# Melodic Links: A Network-Based Study of Artist Collaborations

**Submitted By:-**
Nikhil Kumar - 2022322
Aditya Kumar Sinha - 2022034
Pandillapelly Harshvardhini - 2022345

## Abstract

*With the era of digital streaming and international music exchange, artist collaborations have become a hallmark of modern music production. This paper builds and examines a large-scale network of artist collaborations based on data web-scraped from Spotify and augmented through MusicBrainz and a bespoke local language model pipeline. Of more than 156,000 artists, a fine dataset of 51,002 confirmed musicians was constructed, encompassing rich metadata like genre, origin, language, number of followers, and Spotify popularity. The collaboration network, constructed from more than 158,000 edges, reflects the hidden structural features of the global music scene. By using exploratory network analysis, measures of assortativity, and cross-border collaboration research, we examine how popularity, language, and geography affect artistic collaborations and commercial success. Our research provides new insights into the topological, cultural, and economic aspects of contemporary music collaborations.*

# Table of Contents

# Introduction

Artist pairings have emerged as a unifying force that reshapes the sound of modern music, fueled by technological innovations and platform-driven streaming environments like Spotify. Artist collaborations tend to cross genres, geography, and language, crafting a complex international network of collaboration. The understanding of such collaborations' structure and implications necessitates a data-centered methodology that recognizes both popularity and diversity among participants.

This study provides a robust network analysis of artist collaborations based on a carefully vetted dataset that comes from Spotify. Raw data was initially made up of approximately 156,000 artists that were filtered to 51,002 confirmed artists by eliminating noise and confirming identities through the MusicBrainz API. The collaboration information, as an artist pair edge list, was also augmented with inferred metadata—genre, origin country, and language—using a blend of local large language models and publicly verified databases.

We intend to investigate the ways in which artists are interconnected and grouped in the network, the influence of popularity and location on these relationships, and whether such collaborations might affect or anticipate commercial success. By conceptualizing the music scene as a sparse, complex, and highly interlinked social graph, this work sets the stage for investigating macro- and micro-level processes by which artistic collaboration drives the behavior of the global music scene.

# Deliverable 1: Novel Dataset

## Raw Data:-

The basis of our music collaboration network project started by gathering raw data from Spotify, involving around 156,000 artists featured on the platform up to January 2025. The initial dataset included the following primary attributes for every artist:

- Spotify ID
- Artist Name
- Follower Count (as of January 2025)
- Spotify Popularity Score

Spotify Popularity Score, which varies from 0 to 100, is a private measure indicating the popularity of an artist or a track on Spotify. Though the formula isn't revealed by Spotify, it's based on various factors like:

- The number of plays (with recent streams weighted more heavily),
- The rate at which a track is gaining traction,
- User engagement metrics (e.g., likes, shares, playlist additions),
- Presence in curated playlists,
- Listener behavior, such as skip rate.

But this dataset was missing important contextual factors like artist genre, country of origin, and dominant song language, which are needed for productive network analysis. Additionally, Spotify's API does not offer country of origin or song language by default, restricting the analysis depth of the raw data.
To account for all of the above limitations, we decided to enrich this dataset further.

### Initial Data Cleaning and Validation:-

A preliminary check showed that the raw dataset contained numerous non-legitimate artists since Spotify permits any user to self-identify as an artist. To maintain dataset integrity, we removed such noise through the MusicBrainz API:

- For each artist, we queried MusicBrainz using their name.
- If no valid response was received, the artist was flagged as inauthentic and removed.
- This filtering reduced our dataset to 51,002 verified artists, highlighting the significant amount of noise in the original dataset.

Additional Data:  Edge List-Music Collaboration Data:-

Besides the artist metadata, we also obtained a CSV file of music collaborations. The file had two columns:

- **Artist A**
- **Artist B**

Each row depicts a collaboration edge between the two named artists. This is the skeleton of the music collaboration network, with artists as nodes and collaborations (e.g., tracks together, albums together, or features) as edges.

Having filtered the artists, we used the same checking on the edge list:

- Edges were kept only if both artists in the pair were present in the cleaned list of verified artists.
- This guaranteed that the ultimate network was made up of only authentic collaborations between confirmed, actual artists.

Feature Enrichment (Novel Data Creation):-

In order to enhance the dataset with genre, origin country, and main language, we researched a number of APIs. Most of them were limited by rate limits or required paying. To get around this, we used a hybrid solution:

- We utilized a local LLM-based model (through the Ollama framework) to make inferences of missing attributes by querying the name of each artist and obtaining structured results for:
    1. Primary genre
    2. Country of origin
    3. Primary language of music

- These results were subsequently cross-checked with the MusicBrainz API, using multithreaded parallel querying and cycling API credentials to deal with rate limits.

This approach enabled us to generate a rich metadata profile for each artist in a scalable and reliable way.

Final Dataset Overview:-

Following the data cleaning and enrichment process, our last dataset consisted of:

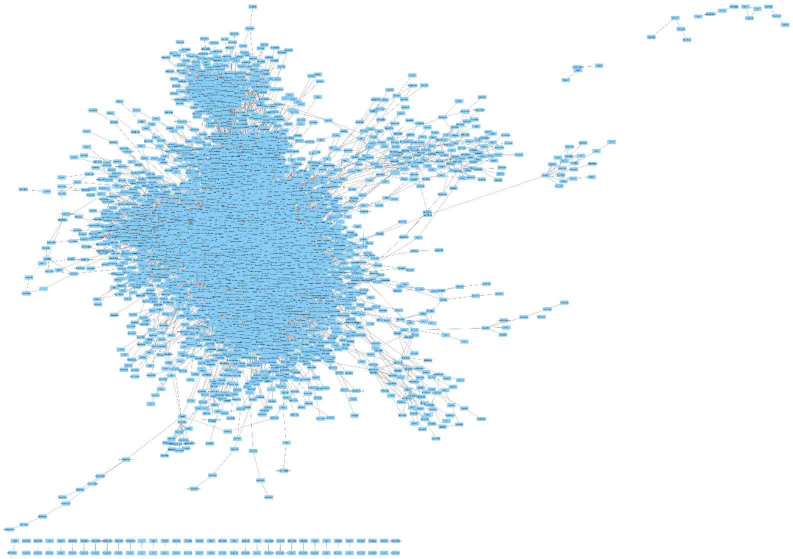
- 51,002 verified artists
- Cleaned edge list of real collaborations
- The following attributes for each artist:
    - Artist Name
    - Spotify ID
    - Follower Count
    - Spotify Popularity Score
    - Primary Genre
    - Country of Origin
    - Primary Language

This final dataset serves as a robust foundation for meaningful network-based exploration of artist collaborations and enables several analytical insights covered in subsequent sections of the report.

## Deliverable 2: EDA On Network

Here, we examine the structural features of our collaboration network of music artists, built from the cleaned and augmented 51,002 verified artist dataset. Each artist is one node, and an undirected edge connects two artists if they have ever collaborated on some song or track. The graph captures the collaborative behavior topology of the global music industry.

Network Visualization:-



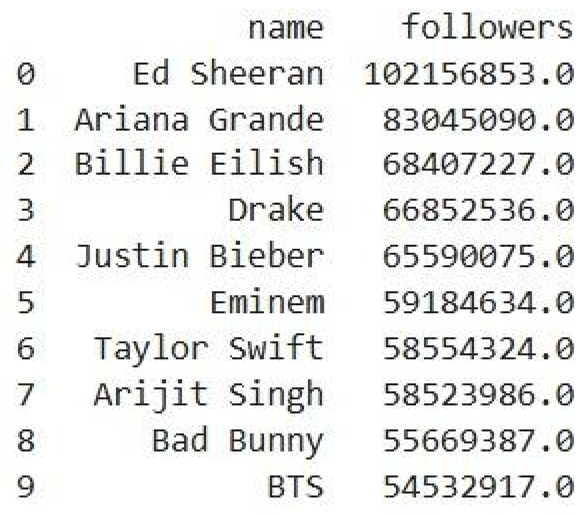
This is a subset of the original network. This represents the top 1% (~ 5000 nodes) of the overall network nodes on the basis of follower count. Every node is an artist

Basic Network Statistics:-

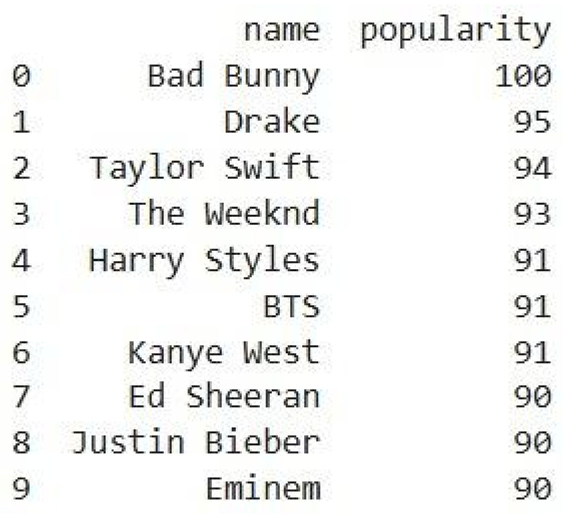
| Metric | Value |
|---|---|
| Number of Nodes | 51,002 |
| Number of Edges | 158,232 |
| Type of Network | Undirected |
| Network Density | 0.0001 |
| Is the Fully Network Connected? | No |
| Number of Connected Components | 478 |
| Size of the Largest Connected Component | 49,698 |
| Maximum Degree | 780 |
| Minimum Degree | 1 |
| Average Degree | 6.2 |

- The network density of 0.0001 represents a very sparse network, as is common in scale level real-world social or collaboration graphs.
- The graph is not connected, with 478 connected components, but the giant connected component contains 49,698 nodes, more than 97% of all artists. This giant component enables us to meaningfully investigate global collaboration structures in the music world.
- High Degree Variability: With a maximum of 780 and a minimum of 1, there is a great variation in artist collaborations. This indicates the existence of super-connectors (potentially prominent global artists) with numerous niche or local artists who have just one collaboration.
- Average Degree Insight: The mean degree of 6.2 further supports the sparsity of the network- the majority of artists work with just a few others. Coupled with the large max degree, this suggests a very skewed degree distribution.

## Top Artists by Follower Count & Spotify Popularity Score:-



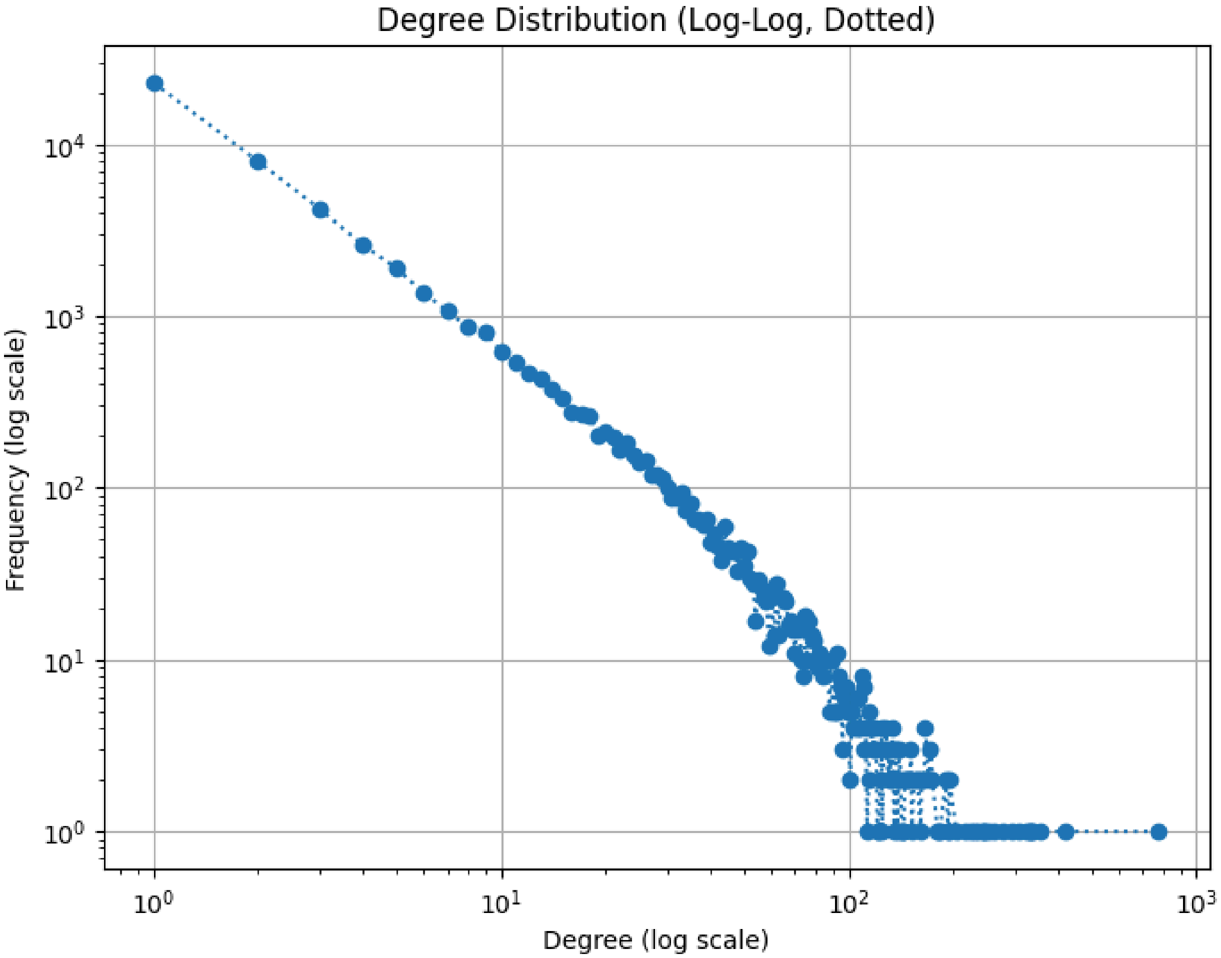
```
             name     followers
0      Ed Sheeran  102156853.0
1   Ariana Grande   83045090.0
2   Billie Eilish   68407227.0
3           Drake   66852536.0
4   Justin Bieber   65590075.0
5          Eminem   59184634.0
6    Taylor Swift   58554324.0
7    Arijit Singh   58523986.0
8       Bad Bunny   55669387.0
9             BTS   54532917.0
```



```
            name  popularity
0      Bad Bunny         100
1          Drake          95
2   Taylor Swift          94
3     The Weeknd          93
4   Harry Styles          91
5            BTS          91
6     Kanye West          91
7     Ed Sheeran          90
8  Justin Bieber          90
9         Eminem          90
```

## Degree Distribution:-

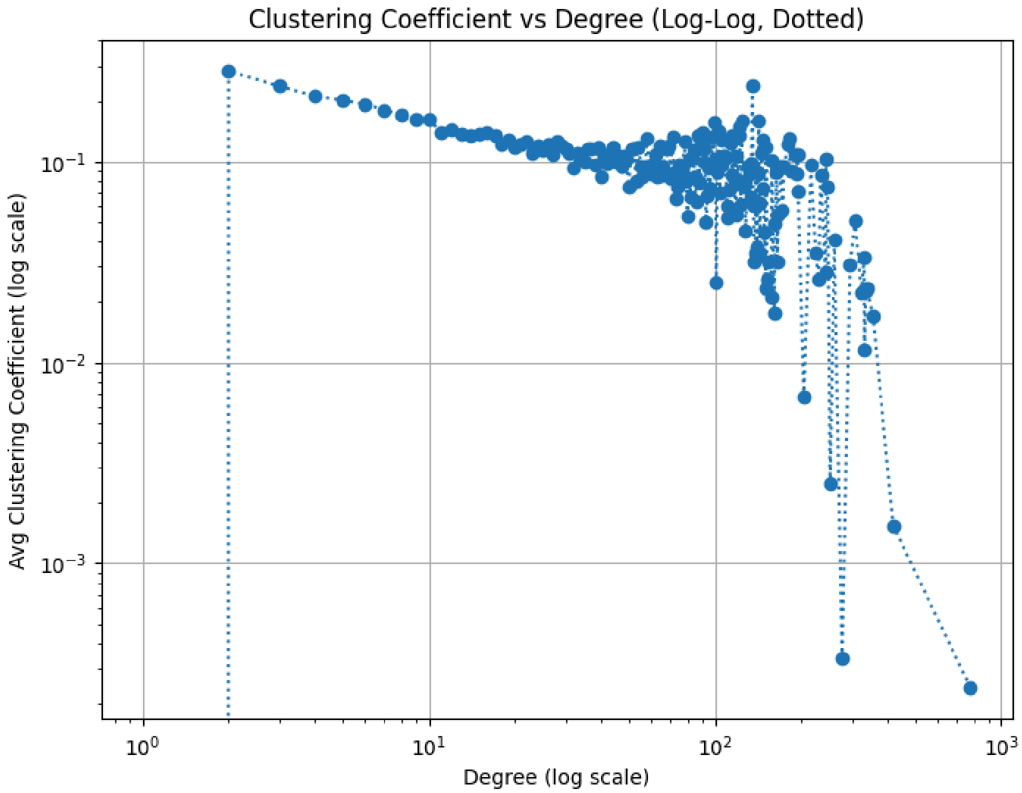
The node degree indicates the number of other artists it has worked with. The degree distribution analysis tells us something about how collaboration is spread out in the population of artists.



Observations:-

- Power Law Behavior: The plot has a roughly linear trend on a log-log scale, which indicates a power-law distribution — one of the signature of scale-free networks. What this implies is that few artists work with many other artists, and most work with only a few.
- Long Tail: There's the visible long tail to the right, indicating a few nodes (artists) having very high degrees (super-collaborators). These would be global celebrities who work frequently across genres and geographies.
- Noise at Higher Degrees: At extremely high levels (above ~100), the information is sparse and noisy with irregular gaps. This is as it should be, given the sparsity of very high-degree nodes, and marks the boundaries of the data size.

## Clustering Coefficient vs Degree:-

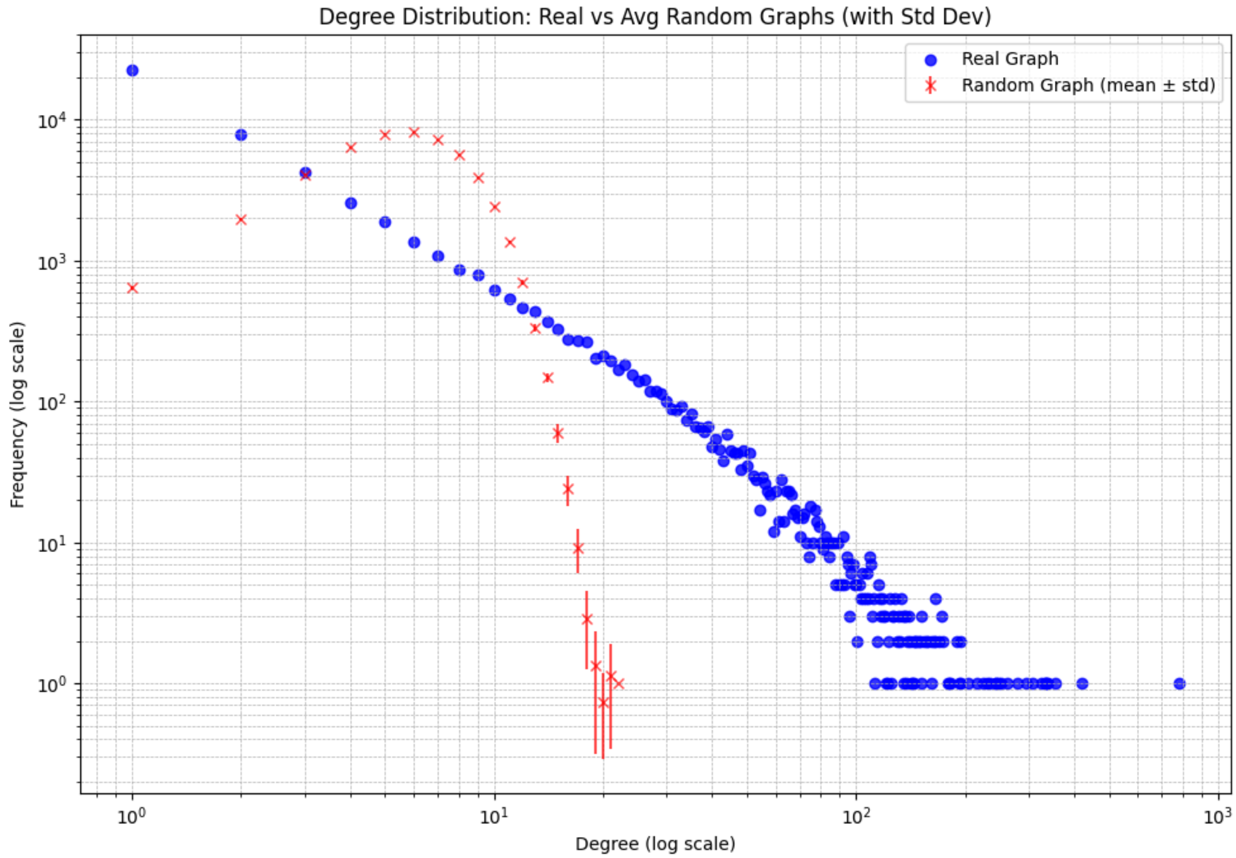


## Observations:-

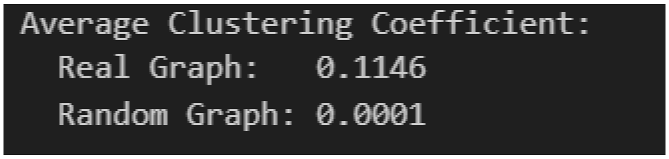
- Negative Correlation Trend: The typical clustering coefficient will drop as the degree increases, a common occurrence in many real-world networks. High-degree nodes are often bridges between clusters, which results in lower clustering.
- Dense Local Communities at Low Degree: Low-degree nodes have comparatively high clustering coefficients, which means small-scale or local cooperation tends to coalesce into tight clusters — say, artists from the same genre, label, or region.
- More Fluctuations at Higher Degrees: As the degree gets higher (particularly above 100), the clustering coefficient values are more volatile, with greater variance. This is an expression of the variability in how high-degree artists collaborate — some collaborate with close groups (still high clustering), others bridge communities (low clustering).

Comparisons with Random Network:-

For this, we created 100 Random Network of comparable nodes and with same edge probability. And then we compared the degree distributions of our music collaboration network with those random networks averaged over 100 instances.



Next, we compared the average clustering coefficient across the two:-



```
Average Clustering Coefficient:
  Real Graph:   0.1146
  Random Graph: 0.0001
```

- The actual graph (blue) displays a long-tailed, roughly power-law degree distribution—a hallmark of scale-free networks commonly encountered in real systems (e.g., social, biological, or technological networks).

- In contrast, the random graph (red) displays a bell-shaped distribution, characteristic of Erdős–Rényi models, where nodes tend to have degrees near the average.
- The average clustering coefficient is ~0.1146 in the actual graph, much larger than ~0.0001 in the random graph.
- This suggests tight local connectivity or community structure in the actual graph-nodes prefer to cluster together tightly, as opposed to the almost tree-like, sparse clustering of random graphs.

# Deliverable 3: Multi-Level Assortativity in Music Collaboration Networks: Connectivity and Popularity

## Assortativity in the Music Collaboration Network:-

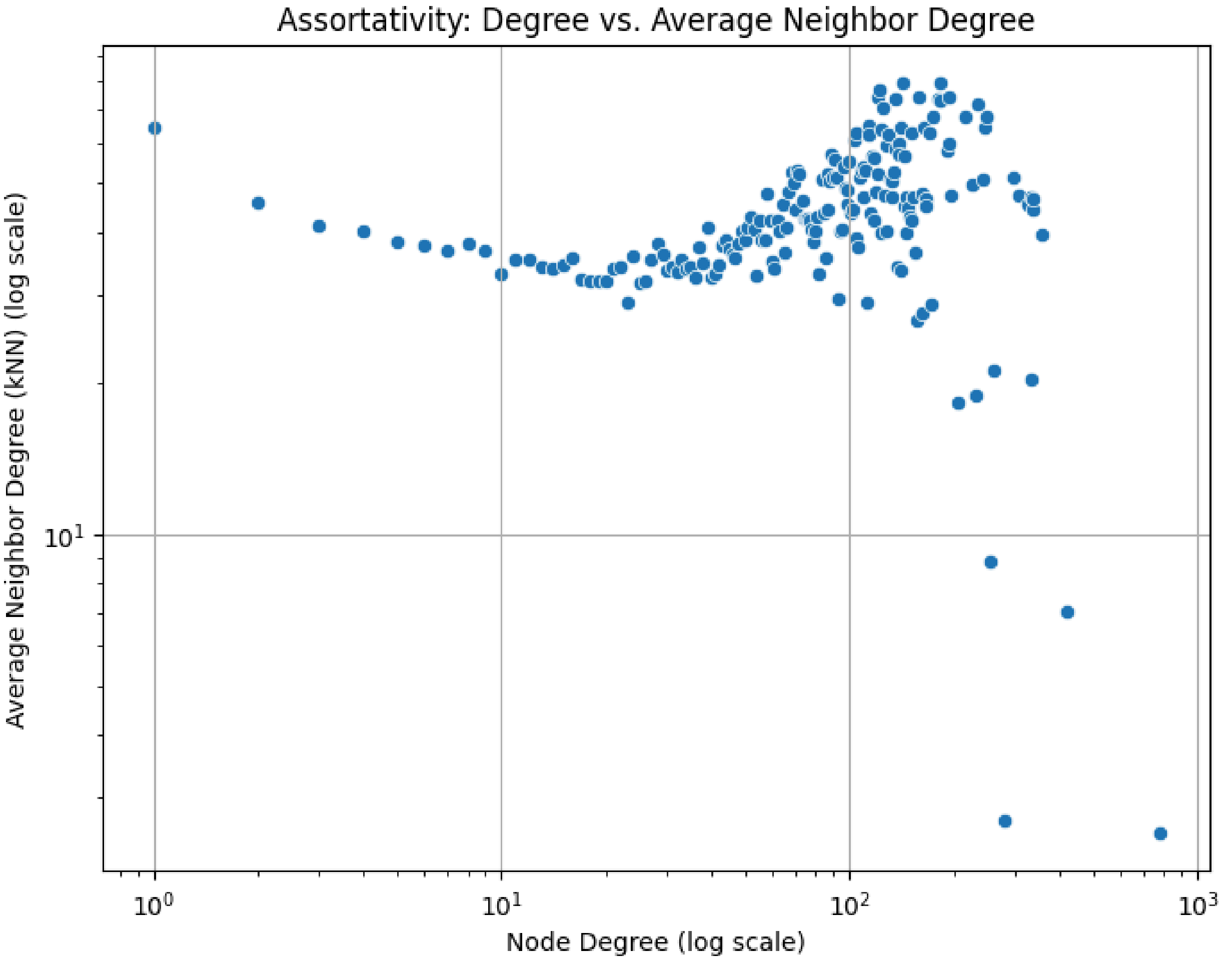
### Background:-

In actual social networks like Facebook, actor networks, or research collaborations, we tend to see a phenomenon called assortative mixing by degree, where highly connected nodes (hubs) tend to connect to other highly connected nodes. This is a sort of "rich-club" effect, in which popularity or centrality dictates who talks to whom.

We speculated that a similar trend could be seen in the music industry, specifically in artist collaboration networks.

### Assortativity by Degree (Topology-Based):-

We computed the average degree of each node's neighbors, also known as knn, and plotted this against the node's own degree. An increasing trend in this relationship would indicate assortativity.



Plot: Degree vs. Average Neighbor Degree (log-log scale)

Result:

The plot showed a clear upward trend for mid-to-high-degree nodes. This suggests that highly connected artists (collaboration hubs) are more likely to collaborate with other highly connected artists. This mirrors the hub-to-hub preference seen in other social systems.

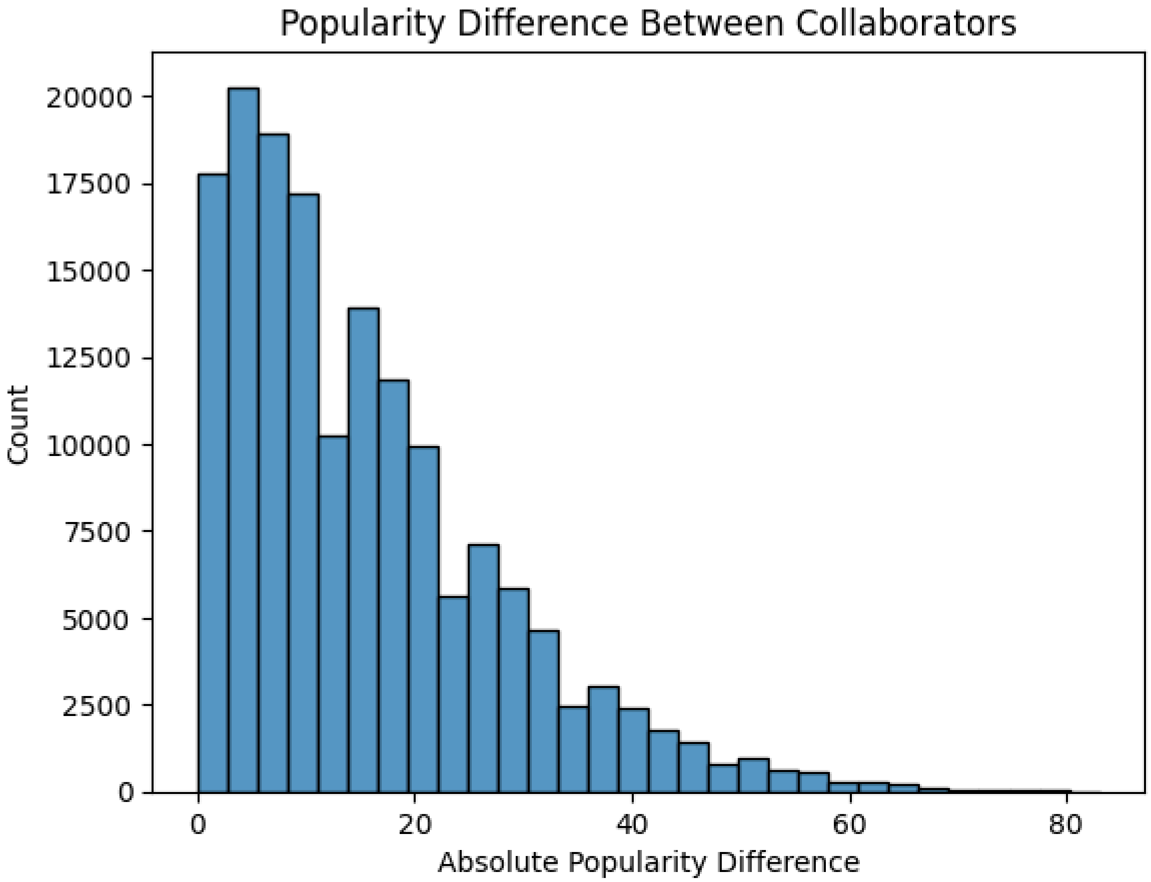
We further supported this result with the degree assortativity coefficient from NetworkX, which yielded a positive value, confirming assortative mixing by degree.

## Assortativity by Popularity Score:-

Although degree does not necessarily translate into perceived popularity or public fame, Spotify offers an objective popularity score per artist based on streamed data, activity from listeners, and interaction.

We thus investigated whether artists prefer to work with others of similar popularity by:

- Computing the absolute difference in popularity for each collaboration pair.
- Creating a histogram illustrating the distribution of these differences.



Plot: Histogram of Popularity Differences Between Collaborators

Result:
The histogram was left-skewed, meaning most collaborations occurred between artists with similar popularity levels. This suggests an additional form of assortativity, not captured by degree alone, where social capital or market standing influences collaboration patterns.

Interpretation:-

Our results lend support to the theory that assortment in the music world works on more than one level:

- Topological Assortativity: Hubs (popular) artists work with other hubs.
- Popularity Assortativity: Artists sharing similar Spotify popularity scores work with each other.

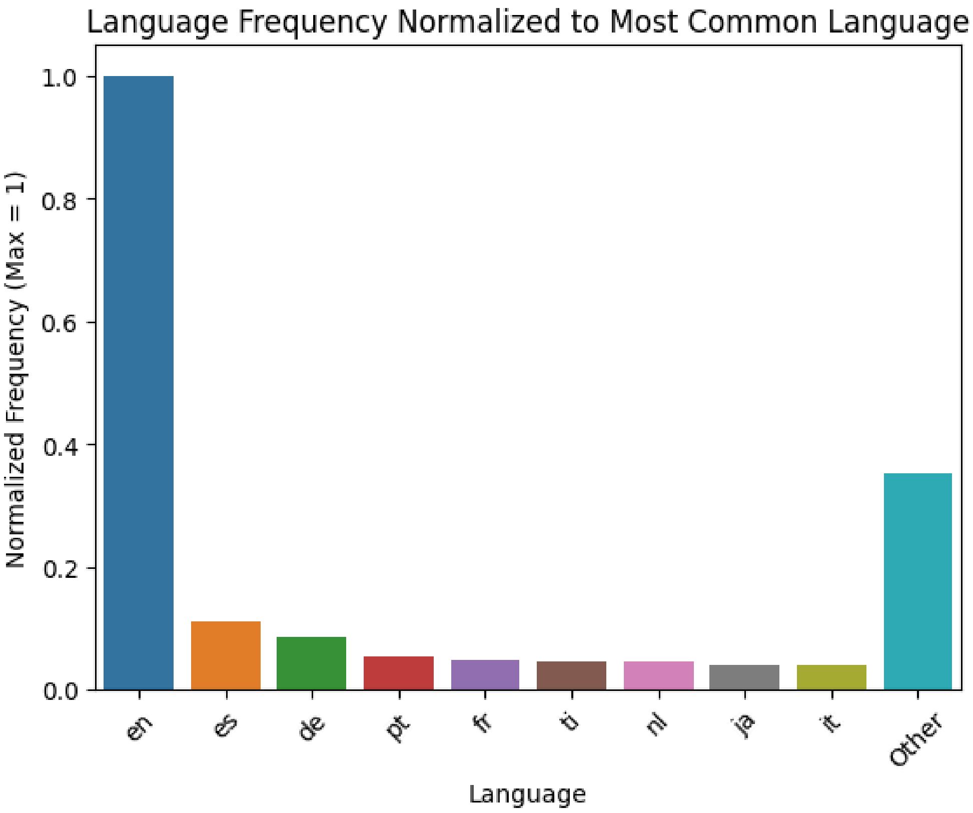
This implies that structural position (degree) and visibility in the market (popularity score) both affect collaboration, creating a stratified system of assortment.

## Deliverable 4: Do Global Language Hubs Predict Music Fame? An Empirical Counterpoint

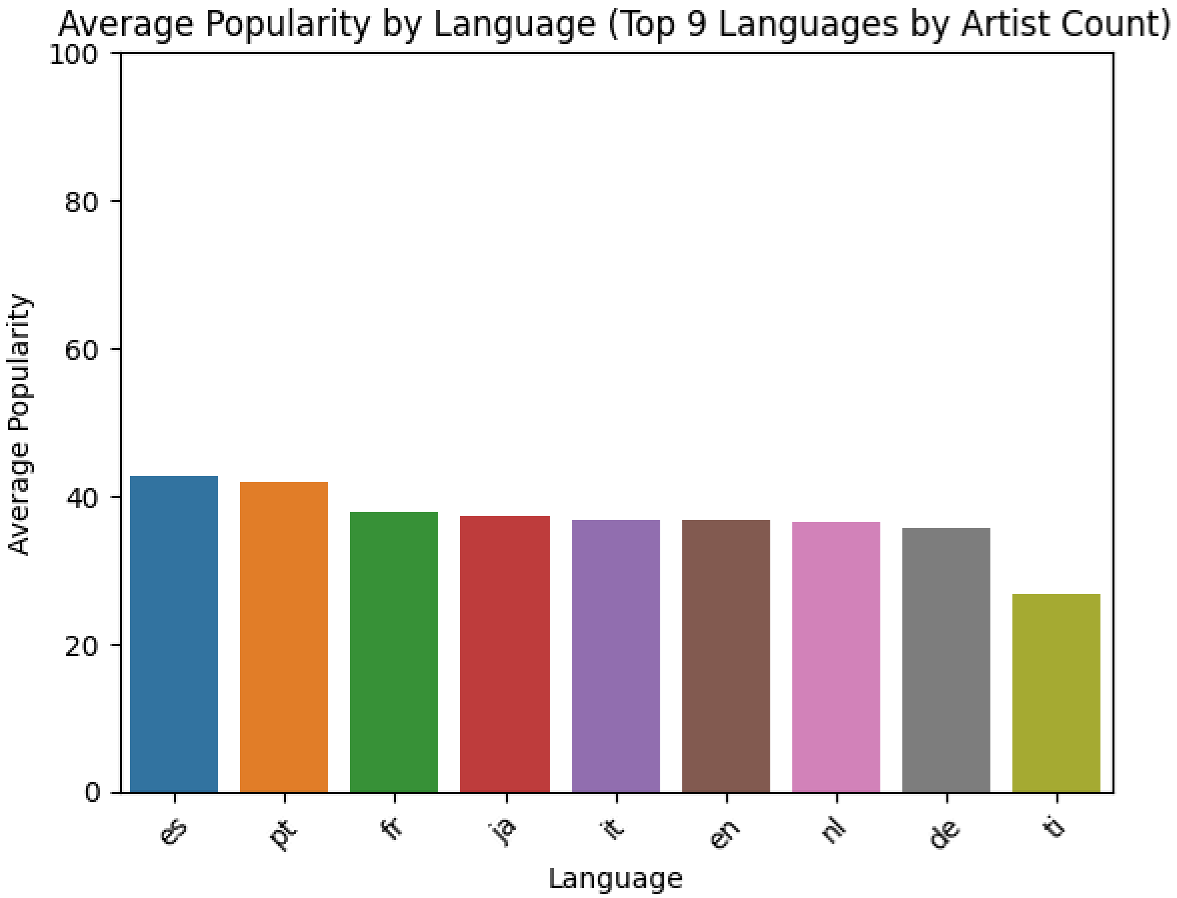
Background:-

In the study "Links that Speak: The Global Language Network and Its Association with Global Fame" by Ronen et al. in 2014, the authors compared multilingual networks (book translations, Wikipedia editions, and Twitter) to derive a quantitative model of global language influence. They found that English is the most central node in global language networks, highly correlated with the number of globally famous individuals. Spanish, French, German, and Chinese are other intermediate nodes.

Critical Observation in Music Collaboration Network:-



Plot: Showcasing the Normalized Frequency of the primary languages of songs by various artists



Plot: Average Popularity of Primary Song Languages

However, when we analyzed our network, consisting of approximately 50,000 nodes, with around 27,000 artists identified as English-language singers, a different trend emerges. Contrary to expectations from the global fame model, the average popularity of artists in the English language group is not significantly higher than that of artists in other language groups. When average popularity was calculated across leading languages by artist number, several languages had comparable or similar average popularity values, indicating that:

- Language centrality in fame networks does not necessarily correspond to popularity in music networks.
- English dominance in artist number does not necessarily mean greater per-artist popularity.

Interpretation:-

This disparity can be explained by the platform-specific dynamics of music listening, including:

- Algorithmic promotion on streaming platforms (e.g., Spotify, YouTube),
- Regional preferences
- Genre-specific niches with international reach (e.g., K-pop in Korean, Reggaeton in Spanish),
- Cross-lingual collaboration and remixing of popular music.

These considerations indicate that the popularity of music might be more democratized across languages than conventional fame or literary visibility, which are more closely connected to language centrality.

# Additional Deliverables

## 1. Cross-Border Collaboration Patterns in the Global Music Network

Background:-

Globalization and online platforms such as Spotify have turned the music industry into a more globally connected sphere. One of the interesting questions is whether artists are more likely to work together in their home countries (intra-country) or different countries (inter-country).

This analysis examines the geographic aspect of artist collaboration, providing insight into how nationality might (or might not) limit creative collaborations.

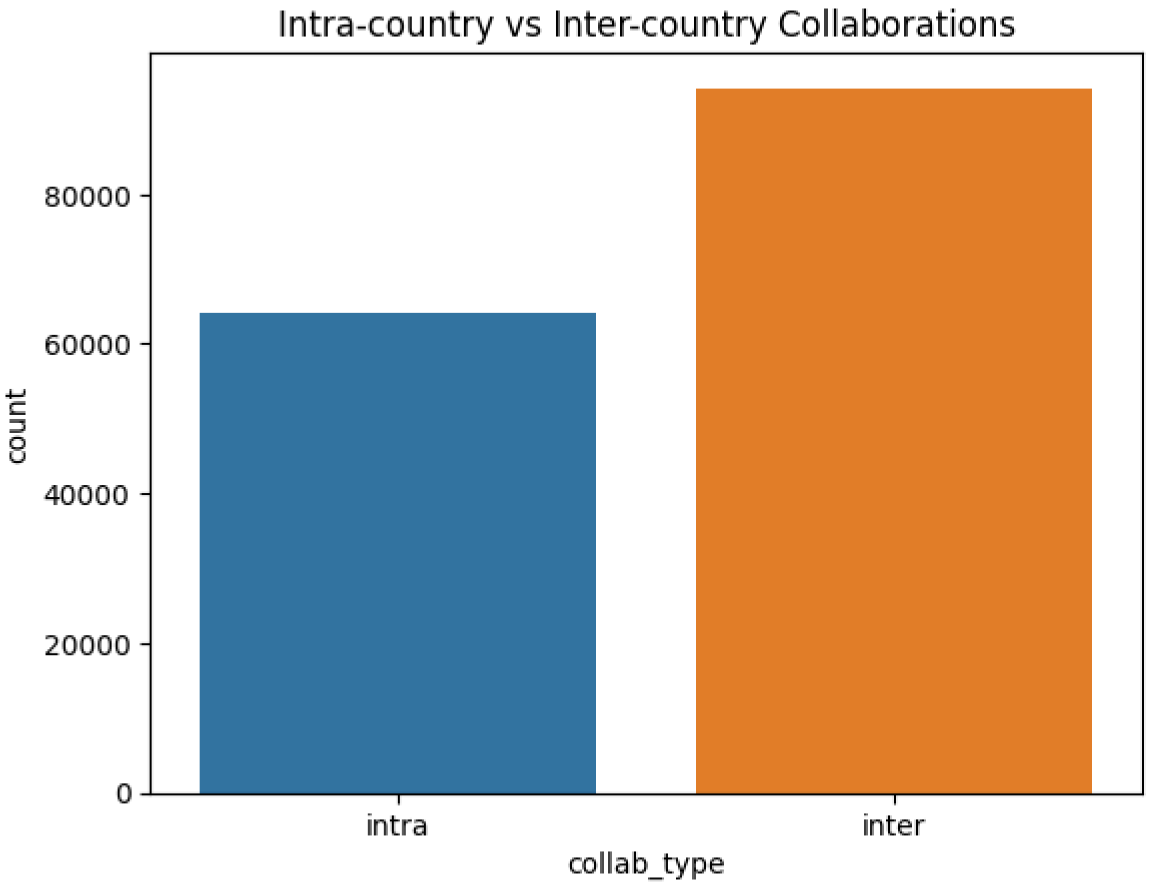
Methodology:-

To determine this, we used the country field from the artist metadata and inspected the collaboration edges in our music graph. For every collaboration (an edge), we used the Spotify IDs of the collaborating artists to retrieve their countries of origin. We considered each collaboration as:

- Intra-country if both artists belong to the same country.
- Inter-country if the artists belong to different countries.

This was done by a simple lookup dictionary and row-by-row comparison in the collaboration DataFrame.

Results:-



Plot: Inter vs Intra Country Colloborations by music artists

The classification results reveal more inter-country collaborations than intra-country ones. The bar plot clearly depicts a bias towards cross-border collaborations. This implies that:

- The music collaboration network is integrated internationally, with extensive cross-cultural collaborations.
- Geographic borders are weak barriers in art partnerships, probably because of digital production software, international labels, and global social networks.

## **2. Collaborations and Commercial Success in the Music Network**

Background:-

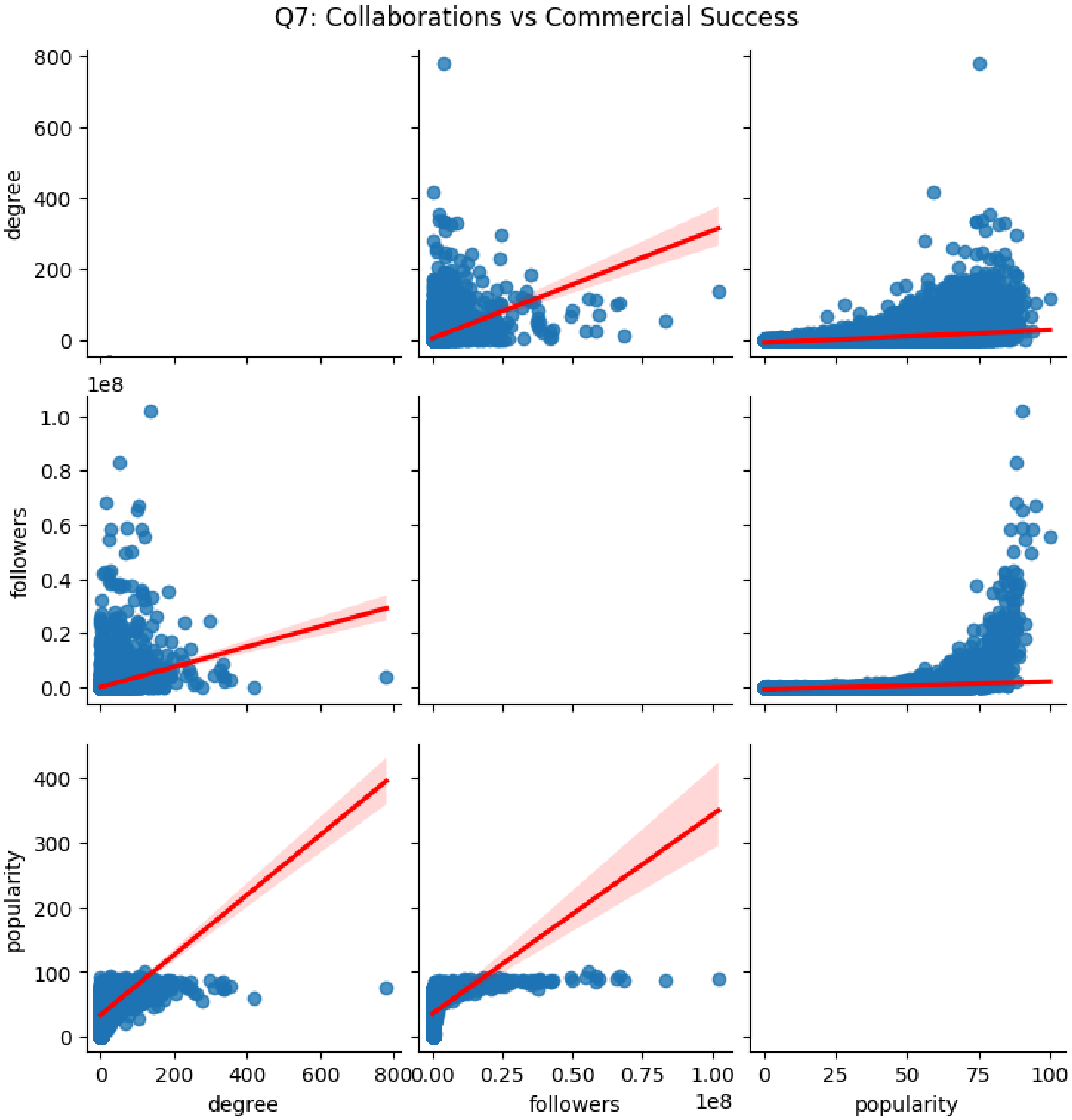
In collaboration and social networks, the size of a node — i.e., the number of direct links — frequently approximates influence, reach, or visibility. In the music industry, artist collaborations are analogous to links that potentially contribute to exposure for each other, access to each other's audiences, and eventually, commercial success. This study investigates whether more collaborative artists (i.e., with a greater degree in the collaboration network) are also more commercially successful, as measured by Spotify's popularity score and number of followers.

Conceptual Framework:-

This investigation is based on well-established network science concepts:

- The Preferential Attachment Model (Barabási & Albert, 1999) explains how nodes with greater connectivity (i.e., higher degree) tend to attract new connections over time - a "rich-get-richer" effect. In the music network, this implies that popular artists who are well-connected might continue to receive more collaborations and attention, increasing their popularity even further.
- Degree Centrality is proportional to visibility within a network. A greater degree tends to be associated with wider spread of content or influence (Newman, 2003), particularly in co-authorship or collaboration networks.
- Other empirical work in other creative networks (scientific collaborations or actor networks) reveals that high degree nodes tend to have high impact indicators (citations, box office performance, etc.) - indicating a pattern which can be transferred to the music context

Our Findings:-



Plot: Correlation Between Popularity, Follower count & Spotify Popularity Score

By examining the collaboration network:-

- Higher degree (more collaborators) artists always had higher popularity scores.
- The same positive trend between degree and follower count was witnessed.
- A positive relationship between followers and popularity testifies to the consistency of Spotify's commercial indicators of success.

This layered correlation suggests a feedback loop: collaborations increase exposure, leading to higher popularity, which in turn may attract further collaborations, reinforcing the preferential attachment mechanism.

# References

1. Bush, R. A. M. (2025). Analysis of a Spotify Collaboration Network for Small-World Properties. arXiv. https://doi.org/10.48550/arXiv.2503.09526
2. Ronen, S., Gonçalves, B., Hu, K. Z., Vespignani, A., Pinker, S., Hidalgo, C. A. (2014). Links that speak: The global language network and its association with global fame. Proceedings of the National Academy of Sciences, 111(52), E5616–E5622. https://doi.org/10.1073/pnas.1410931111
3. Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509–512. https://doi.org/10.1126/science.286.5439.509
4. Newman, M. E. J. (2003). The structure and function of complex networks. SIAM Review, 45(2), 167–256. https://doi.org/10.1137/S003614450342480
5. Hagberg, A., Swart, P., & S Chult, D. (2008). Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th Python in Science Conference (SciPy2008), 11–15. URL: https://networkx.org
6. McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, 51–56. URL: https://pandas.pydata.org
7. Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). Array programming with NumPy. Nature, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2 URL: https://numpy.org