DEPARTMENT OF PHYSICS AND ASTRONOMY "A. RIGHI"

SECOND CYCLE DEGREE

PHYSICS

Network Analysis of the Spotify Artist Collaboration Graph

Margherita Pulga

Martina La Rosa

Academic Year 2024/2025

Contents

Chapter 1

Introduction

Music is a universal cultural language that transcends borders, eras and social boundaries. It exists in countless forms and traditions across the world, reflecting the rich diversity of human expression. From ancient melodies to electronic compositions, music resists rigid classification, continually evolving and blending genres. In the digital age, this vast and dynamic soundscape can be explored through data-driven platforms like Spotify, which offer unprecedented access to musical trends, preferences, and patterns worldwide.

Driven by a strong passion for music and by our curiosity for knowledge, we decided to combine them in order to have access to a deeper understanding of the music word today.

In order to do this, we began our analysis by consulting the article "Tobin South, Network Analysis of the Spotify Artist Collaboration Graph" [1], in order to explore the connections, to analyse the correlations and to extrapolate the information for understanding what the music is built on today and how to describe its important features.

Chosen the aforementioned article as our reference and starting point, we decided to analyse and delve into the world of music from a scientific point of view, exploring the music as a cultural but also a social network. Our approach is justified by the fact that, although music itself has been studied extensively, new music streaming and processing technologies allow a unique data-driven analysis of music and listeners.

To this end, Spotify represents, as a dominant player in the music consumption market, the best source of data for analyzing both modern and classical music.

Therefore, we decide to replicate and to further develop the analysis of [1], focusing our study specifically on the examination of the collaboration network between musical artists.
We took the data from the dataset in reference [2], involving artists who have charted on Spotify and their collaborators in features, identifying and exploring small-world properties within the network. These data are, interestingly, a photograph of today's society and an important way to extrapolate information about how people get influenced by a social phenomenon such as music. They are powerful and full of significance, and our aim with this project is to study, analyse, understand them, and to show their properties and the innumerable possibilities of reflection they can offer.

Thus, the project is structured as follows: we begin by analysing fundamental structural properties of the graph, including the vertex degree distribution, artist popularity through the complementary cumulative distribution function (CCDF) and the relationship between follower count and popularity. We also examine how node degree varies with popularity, and explore the friendship paradox within the network.
Subsequently, we apply the PageRank algorithm to identify influential nodes. The analysis then shifts toward a genre-based perspective: we study the popularity of musical genres, considering both the average popularity of artists within each genre and the distribution among the top 10 most popular genres. We further investigate genre connectivity by modeling genres as a hypergraph.
Finally, we conduct an in-depth examination of the network's community structure using the Louvain community detection algorithm [3] [4], a particularly effective method in this context due to its scalability and ability to reveal hierarchical organization in large networks. A dedicated section is devoted to this method and its implications for understanding the underlying organization of musical collaborations.

Results show that the network is scale-free, with a power-law degree distribution and highly popular artists acting as central hubs. Louvain community detection reveals well-defined collaboration clusters, largely aligned with genre and industry-driven patterns. Additionally, genre-based hypergraphs are constructed to explore inter-genre relationships. These findings provide valuable insights into digital music ecosystems, collaborative dynamics, and potential applications in music recommendation systems.

Chapter 2

Collaboration Network

Networks, or -in mathematical terms- **graphs**, are models that depict the relationships and links between objects. Generally speaking, a graph is defined as an ordered pair $G = (V, E)$, where V is the set of vertices (or nodes) and E is the set of edges that connect pairs of these vertices. In this analysis, the edges are undirected, which means that each pair of connected nodes has no specific orientation [5].

In the particular graph this project aims to examine, the musical artists who have released songs on Spotify represent $nodes$, while $edges$ indicate collaborations between artists who have appeared together on the same song or album.
The largest connected component of the artist collaboration network includes approximately 1.250.065 nodes, which makes up a substantial portion of the overall network of just over 2 million artists. However, most artists have very low popularity and only the most popular are found in the core of the network.

Within this network of approximately 156,000 artists, over 300,000 are the edges connecting them, i.e. the collaborative connections. Thus, on average, each artist collaborates with roughly 3.5 other artists. Not all the connections represent the same type of collaboration: the majority of them indicates artists who have worked together on a song through co-performance or co-production; while some edges represent other relationships, like "appears-on" edges, which may also be directed. Furthermore, 'appears-on' edges can also capture cases in which an artist incorporates another artist's work without direct collaboration. This refers in particular to the cases in which an artist samples, remixes, or remakes a song.
Another parameter which has been taken into consideration is the case in which artists choose to not include a song they collaborate to in their profiles. However, often, when two artists collaborate on a song, that song appears in both artists' profiles.
Although these directed edges are meaningful, for the purposes of this graph analysis, all edges are treated as undirected. The directionality data are instead captured in a measure called the artist reciprocation coefficient, which reflects the proportion of an artist's songs that points to other artists who also point back to them.

2.1 Popularity

One of the most important parameter to consider in our analysis is, without a doubt, the popularity of the artist. Popularity influences—and is influenced by—the number of collaborations and the artist's

connectivity in the network.
Artists with higher popularity tend to have more opportunities to collaborate with others, and these collaborations, in turn, often further boost their popularity by expanding their audience and increasing their visibility. This means that popular artists often occupy central positions in the network, forming dense clusters of connections that can significantly impact the overall structure and dynamics of the collaboration graph. Understanding the role of popularity is therefore crucial for interpreting the formation and evolution of the artist collaboration network.

Since Ed Sheeran is the artist on Spotify with the highest number of total streams, Spotify defines an artists popularity as the fraction of an artists total streams compared to Ed Sheeran's, floored to an integer value:

$$\text{Popularity} = \frac{\text{Artist Total Streams}}{\text{Ed Sheeran Total Streams}}$$

The popularity of the artists is then given as a number between 0 and 100, with Ed Sheeran who is the only one with a popularity of 100.

As a consequence of this definition, the vast majority of artists have low popularity, while a relatively small group accounts for most of the music streams. This results in the following pattern:

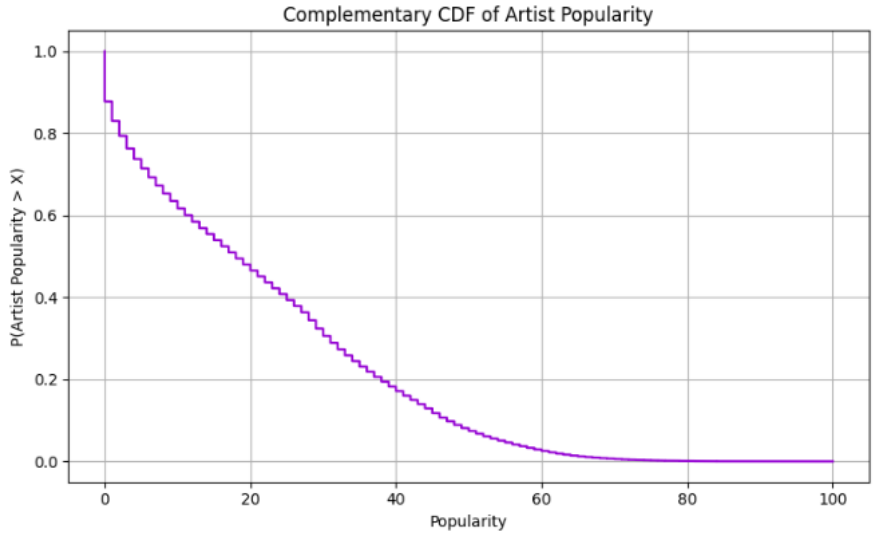

Figure 2.1: Complimentary Cumulative Distribution Function of the Popularity of all artists in the graph.

Which shows as only a small number of singers contribute to the majority of music streams.
Interestingly, it is true that nodes with higher popularity generally tend to have more followers, but this is not consistently the case for all high-popularity nodes. Indeed, the graph in Figure 2.2 shows how a few top artists dominate overall listening.

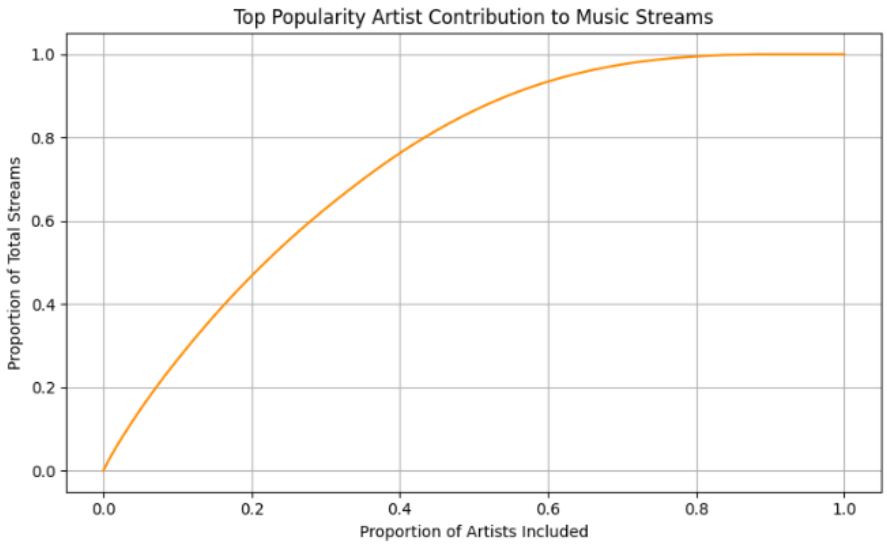

Figure 2.2: Contribution of remaining artists to total popularity as low popularity nodes are removed

This is a cumulative plot that illustrates how much the most popular artists contribute to the total number of streams. The *x-axis* represents the proportion of artists included, ranging from 0 to 1. The *y-axis* shows the percentage of total music streams accounted for by these artists. The curve is upward-sloping and steep at the beginning, indicating that a small fraction of highly popular artists generates a disproportionately large share of total streams. This pattern is characteristic of a power-law distribution or "long tail" effect.

The mechanisms behind this discrepancy are difficult to pinpoint, as they relate to Spotify's varied strategies for encouraging users to follow artists versus simply listening to their music. One possibility is that an artist's follower count reflects the size and engagement of their fan base more than it does general public awareness or listening behaviour.

2.2 Vertex Degree Distribution

Another important parameter for our considerations is the vertex degree.
The **degree of a vertex** represents the number of connections it has, which, in this context, corresponds to the number of collaborations an artist has participated in [5].
In order to find it, we first compute some network statistics, such as the number of nodes and edges, which quantify the graph's scale; the network density, defined as the ratio of actual to possible edges, reflecting how interconnected the graph is; the number of connected components, revealing how fragmented or cohesive the network is; the average clustering coefficient and the degree associativity coefficient, which indicates whether artists with similar numbers of collaborations tend to connect with one another.

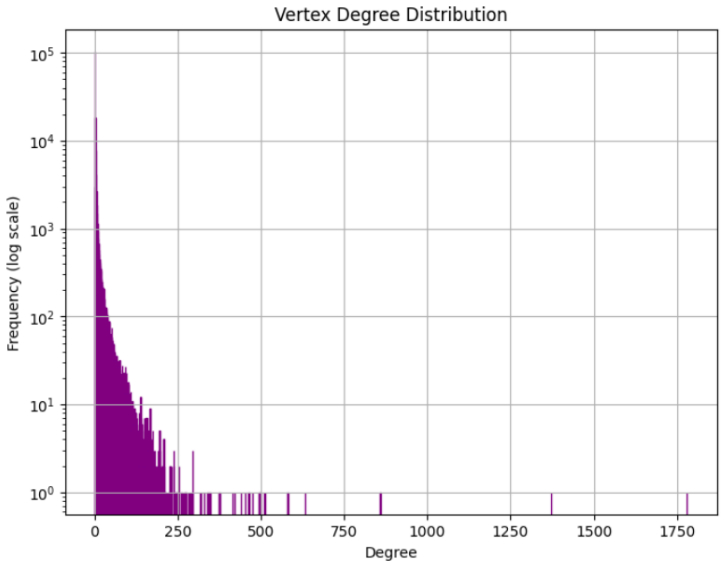

Figure 2.3: Vertex Degree Distribution: The histogram is plotted on a logarithmic scale to account for the long-tailed nature typical of real-world social and collaboration networks.

As shown in Figure 2.3, there are many artists with very few connections and a few that are "hubs" with many connections. This suggests to investigate for **power law**:

$$P(k) \sim k^{-\gamma}$$

The degree distribution follows a power law with $\gamma = 2.0216$. This fits within the lower end of the typical range of $2 < \gamma < 3$ for scale-free networks and suggests the dominance of highly connected hubs in the network, as the tail of the distribution is heavier. Figures 2.4 shows the Logarithmic degree distribution and the power law fit. From the above plot, we can see that the largest connected component of the entire graph follows a strongly power law $(R^2 = 0.93)$, with slope -1.93.

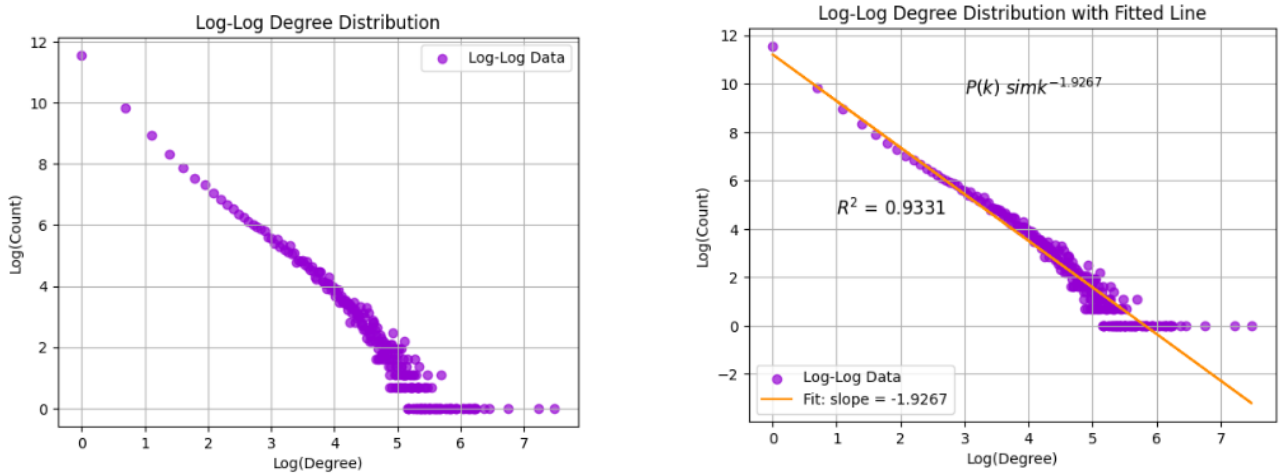

Figure 2.4: Log-log plot of the degree distribution with the fitted power law line (right panel).

This distribution highlights the existence of a few highly connected vertices (hubs) within the network. Specifically, 83 artists have more than 200 collaborations. The most connected artist is Steve Aoki and stands out with 498 direct collaborative connections, including 23 connections with artists who each have over 200 collaborations.

Chapter 3

Graph Analysis

3.1 Friendship Paradox

We talk about **Friendship Paradox** when, on averaged, an artist has less collaborators than the average of collaborations of the artist's collaboration. This means that, in total, there are more "friends of friends" than direct friends, because an individual's friends are likely to share connections with other friends, leading to overlapping and repeated links within the network [6].

We can illustrate the friendship paradox in complex networks using a scatter plot, where each point represents an artist. On the *x-axis*, we plot the artist's number of collaborations (degree), and on the *y-axis*, we plot the average degree of their collaborators:

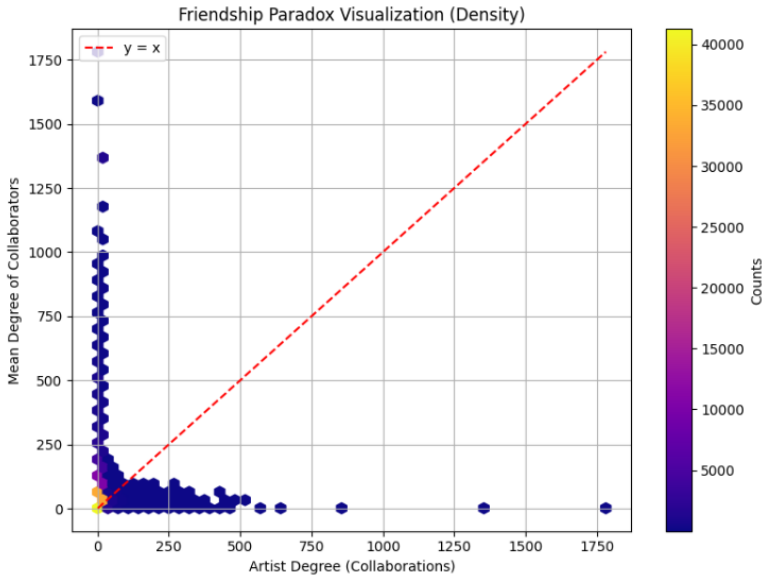

Figure 3.1: Representation of the Friendship Paradox

This plot has to be interpreted in the following way: the plot uses hexagonal bins to represent the density of data points—darker or more saturated areas indicate regions where many artists share similar degree values. A dashed red line is drawn along the line $y = x$, which acts as a reference: if the majority of points lie above the red line $y = x$, it indicates that the average degree of an artist's collaborators

is greater than the artist's own degree. The presence of dense regions above the red line confirms the presence of the friendship paradox in the network.
Thus the friendship paradox refers to the peculiar phenomenon where, on average, your friends (or collaborators, in this case) have more connections than you do. This arises due to the way highly connected nodes (hubs) are overrepresented in the neighbour lists of other and the network we are examining shows a strong visual indication that the friendship paradox holds.

3.2 Eigenvector Centrality

A central motivation behind this project is to explore how humans analyze and assign value to music, particularly in terms of identifying what makes certain music or artists "important." One effective approach for this type of analysis, especially within the context of networked data, is the application of **network centrality measures**. While centrality can be quantified using various methods such as betweenness, closeness, and the PageRank algorithm and eigenvector centrality.

Eigenvector centrality is a measure used in network analysis to assess the influence of a node within a graph, not merely based on the quantity of its connections (as in degree centrality), but also on the quality or importance of the nodes it is connected to. In other words, a node achieves a high eigenvector centrality score not only by having many connections, but by being connected to other nodes that are themselves highly central. This makes the measure particularly useful for identifying key influencers in complex networks, such as social, biological, or collaboration networks.
In the context of the current analysis, eigenvector centrality is computed on a graph$G$, which likely represents a network of artists, where nodes correspond to individual artists and edges represent relationships or collaborations between them. Our aim is to identify the most influential artists in the network, i.e. those who are not only well-connected but are also connected to other prominent figures.

After computing eigenvector centrality for all nodes in the network graph G, the code proceeds to extract and visualize the most influential subset of the network. Specifically, the top 100 nodes with the highest eigenvector centrality scores are selected, and a subgraph is created from this selection. This subgraph represents the eigenvector central core of the network -i.e., the group of nodes that are structurally most central and well-connected to other influential nodes.

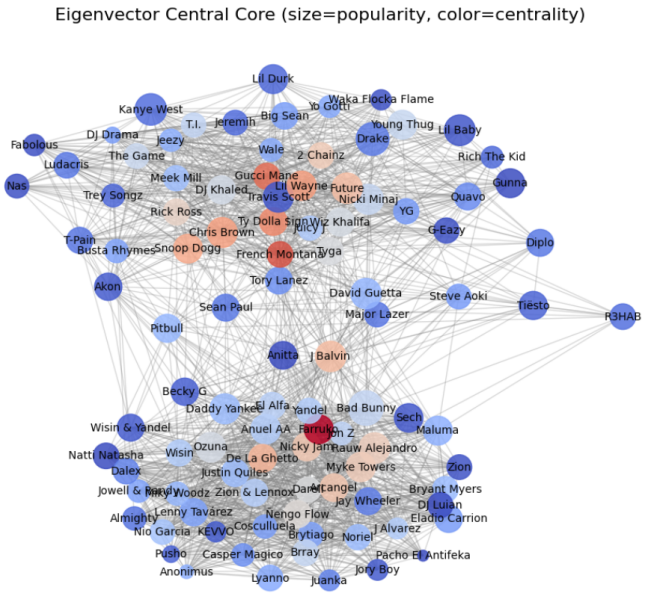

Figure 3.2: Eigenvector Central Core of the Artist Collaboration Network.

As can be seen from the plot above, the most influential nodes—those with the highest eigenvector centrality—are represented in red and are generally positioned towards the center of the graph layout. These nodes are not only well-connected themselves but are also connected to other highly central nodes, highlighting their structural importance within the network.
The most central node in the entire network is J Balvin: he is an artist who not only collaborates widely but does so with other highly connected artists. Besides, a strong Latin music cluster is visible in the lower part of the graph.This reveals that latin artists dominate the central structure, indicating tightly interconnected collaborations across reggaeton, Latin pop, and trap, while U.S. hip-hop artists, although globally popular, are relatively peripheral in terms of network centrality.

3.3 PageRank-Based Analysis of Artist Influence in the Network

To further investigate node importance in the artist network, the PageRank algorithm is applied to the full graph G. Originally developed for ranking web pages, PageRank evaluates the relative importance of nodes based on the concept of recursive endorsement: a node is considered important if it is linked by other important nodes. Unlike eigenvector centrality, PageRank incorporates a random walk model with a damping factor $(\alpha = 0.85)$, simulating the probability that a user randomly jumps to a new node during traversal. Therefore, PageRank is a variant of EigenCentrality, since it also assigns nodes a score based on their connection and their connections' connections. The difference is that PageRank also takes link direction and weight into account, so links can pass influence in one direction, and pass different amounts of influence. This measure uncovers nodes whose influence extends beyond their direct connections into the wider network [7]. Mathematically, PageRank (PR) is defined as:

$$PR(A) = \frac{(1-d)}{N} + d\left(\sum_{k=1}^{n} \frac{PR(P_k)}{C(P_k)}\right), \tag{3.1}$$

where $PR(A)$ is the PageRank value of page A that we are considering (in this case the pages correspond to nodes), N is the total number of pages, n is the number of pages containing at least one link toward

A. P_k represents each page and $PR(P_k)$ are the pagerank values of each page P_k. $C(P_k)$ is the number of links going out of page A, while d, also known as the **damping factor**, is usually set equal to 0.85. This parameter was introduced to take into account the probability, at any step, that the person will continue following links. The probability that they instead jump to any random pages is $(1 - d)$. In our network this corresponds to how people listen to artists and how these artists are connected. From the eq. (3.1) we can notice that by increasing the number of total links that point to A, one increases also the PageRank. The PageRank values are the entries of the dominant right eigenvector of the modified adjacency matrix rescaled so that each column adds up to one [8]. The eigenvector is

$$R = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ : \\ PR(p_n), \end{bmatrix} \tag{3.2}$$

where R is the solution of the equation

$$R = \begin{bmatrix} (1-d)/N \\ (1-d)/N \\ : \\ (1-d)/N \end{bmatrix} + d \begin{bmatrix} l(p_1,p_1) & l(p_1,p_2) & ... & l(p_1,p_n) \\ l(p_2,p_1) & : & : & : \\ : & : & l(p_i,p_j) & : \\ l(p_N,p_1) & ... & & l(p_N,p_N) \end{bmatrix} R \tag{3.3}$$

where the adjancency function $l(p_i, p_j)$ is the ratio between number of links outbound from page j to i to the total number of outbound links of page j. The adjacency function is 0 if the page p_j does not link to p_i. Besides, it is normalized such that, for each j,

$$\sum_{i=1}^{N} l(p_i, p_j) = 1, \tag{3.4}$$

i.e. the elements of each column sum up to 1. Thus, this is a variant of the eigenvector centrality measure.
To explore the distribution of centrality across the network, two plots are produced and represented in Figure 3.3:

- Histogram of PageRank Centrality;
- Sorted PageRank Plot.

Together, these steps provide a quantitative and visual understanding of influence hierarchy within the artist network, offering a complementary perspective to eigenvector centrality by considering both link structure and probabilistic navigation.

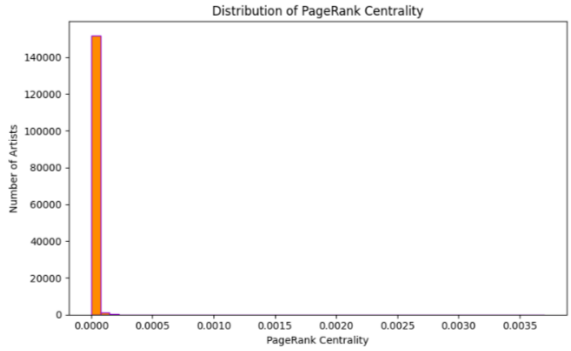

Figure 3.3: Frequency distribution of PageRank values across all artists. The skewed distribution typically reveals that only a small number of nodes receive high PageRank scores, indicating a heavy-tailed structure where influence is concentrated among few key actors.

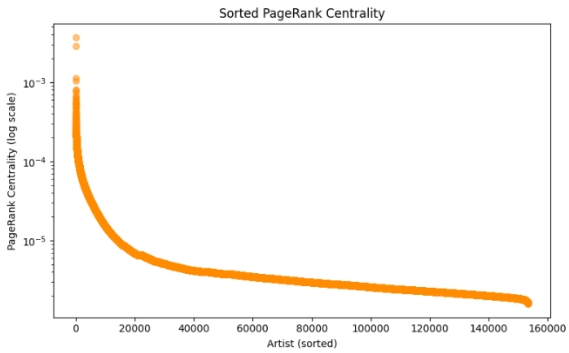

Figure 3.4: The use of log scaling helps visualize the disparity between the top nodes and the rest of the network. This emphasizes the exponential decay of centrality and highlights the sharp drop-off after the most influential artists.

3.3.1 PageRank and Node Degree

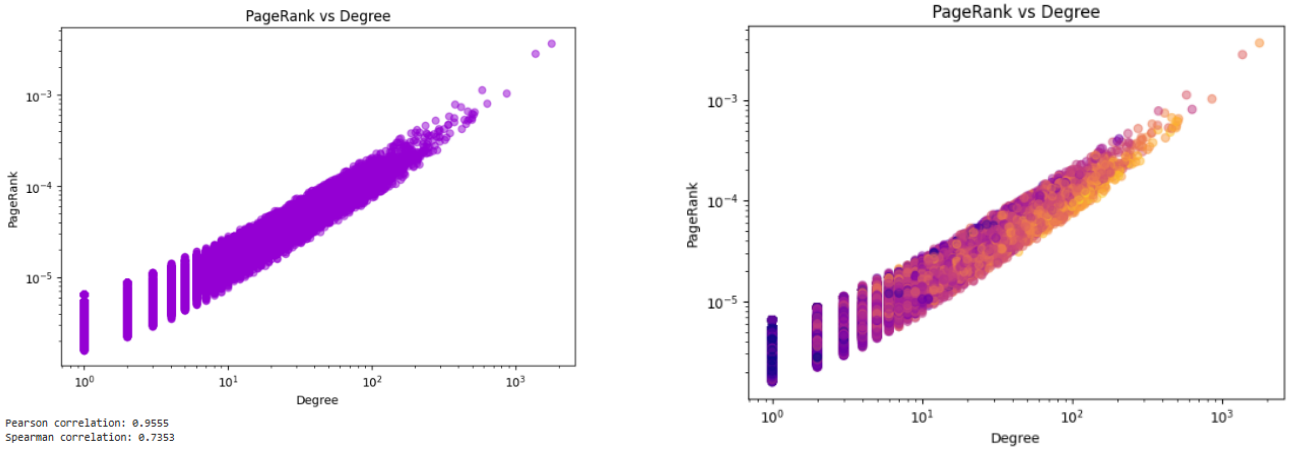

Figure 3.5: **Left panel**:Increasing behaviour in log-log scale shows a strong positive correlation between degree and PageRank: artists with more collaborations have a higher PageRank. **Right panel**: The color is added according to popularity, to see which groups dominate in some region of the graph.

To better understand the structural importance of artists within the network, we analyze the relationship between PageRank and the degree of each node. In this context, the degree represents the number of collaborations an artist has, i.e., the number of edges connected to a given node. PageRank, on the other hand, captures a notion of global importance and quantifies not only the number of connections but also the influence of the nodes to which a node is connected.

The analysis is conducted on the largest connected component of the network to ensure the applicability of global centrality measures.
The first plot (left panel) presents a scatter plot of PageRank versus degree in log-log scale. The results indicate a strong positive correlation between degree and PageRank, as confirmed by both *Pearson's*

correlation coefficient (0.9555)[1] and *Spearman's rank correlation* (0.7353)[2]. This suggests that artists with a higher number of collaborations tend to be ranked higher in terms of PageRank, supporting the idea that the quantity of connections is a primary driver of centrality in this network. Additionally, the dispersion of PageRank values for a given degree is relatively low, implying that degree alone is often a sufficient predictor of an artist's influence, at least in PageRank terms.
The observed distribution is consistent with that of a *scale-free network*, a common feature of real-world systems such as the web or social networks indeed. Most nodes have low degree and PageRank (bottom left), while a few highly connected nodes exhibit much greater centrality (top right), acting as hubs within the graph.

In the right panel, node colors are added based on their popularity values, using a plasma colormap. This allows for an exploration of whether popularity introduces additional structure or clustering within the PageRank-degree space. The colored scatter plot reveals that while popularity does vary across the spectrum, no dominant clusters emerge. This can suggest that, although popularity and PageRank are related, they are not strictly aligned. In particular, high-popularity nodes can be found across a range of degree and PageRank values.

[1] Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

[2] Spearman's rank correlation coefficient is a number ranging from -1 to 1 that indicates how strongly two sets of ranks are correlated. In particular, the Spearman correlation coefficient is defined as the Pearson correlation coefficient between the rank variables[9]

3.4 Genre Analysis

In order to gain insight into the genre composition of the dataset, we analyze the frequency of genres across all artists in the network. Each artist may be associated with multiple genres, and to capture the overall representation, we aggregate all genre labels and compute their total occurrences.

The resulting distribution reveals the top 10 most common genres in the dataset (3.6). This is visualized through a bar chart, where the x-axis represents genre labels and the y-axis shows the number of artists associated with each genre.

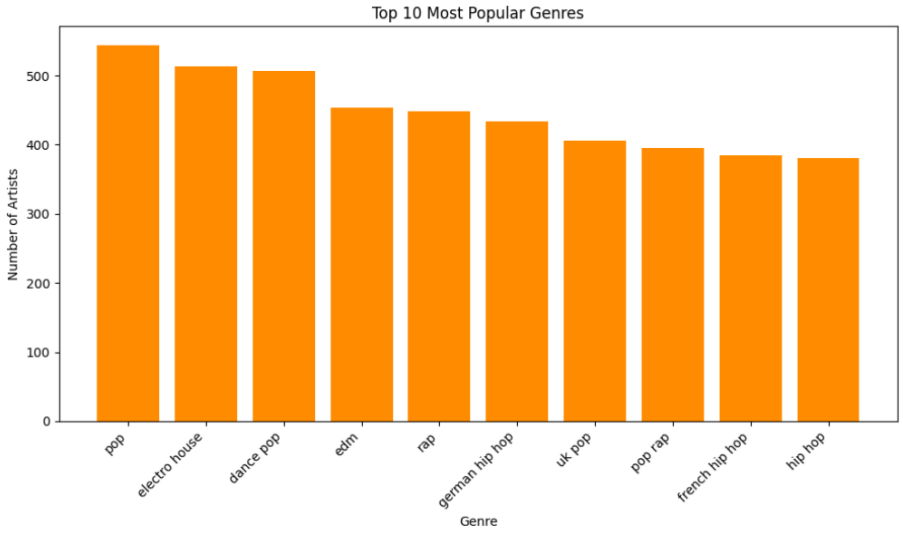

Figure 3.5: Top 10 most frequent music genres among the artists in the dataset. The bar chart displays the number of artists associated with each genre.

In contrast with "The top 10 most frequent genres" analysis, where we counted the total number of artists associated with each genre, a second analysis focused on the overall distribution of genre sizes is plotted in Figure 3.6.

By aggregating the number of artists per genre across the entire dataset, we generated an histogram to visualize how genre popularity is distributed. The x-axis represents the number of artists within a genre, while the y-axis indicates how many genres fall into each size category.

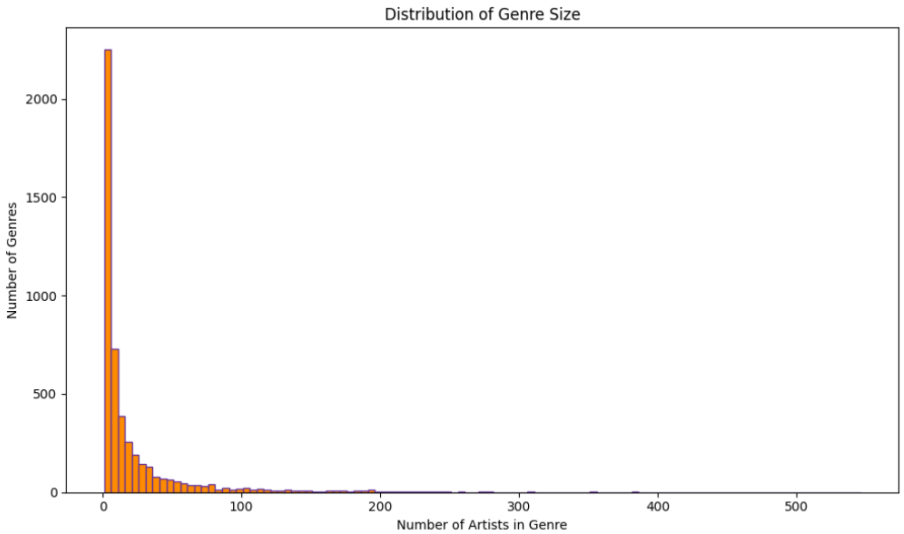

Figure 3.6: Top 10 most frequent music genres among the artists in the dataset. The bar chart displays the number of artists associated with each genre.

While the first graph emphasizes which genres are the most popular on Spotify, the second reveals the underlying asymmetric distribution of genre representation. Indeed, most genres are associated with a small number of artists, whereas only a few genres include a large artist base.

Thus, together, these plots provide a comprehensive view of genre popularity and diversity within the artist network.

3.4.1 Genre Coexistence Hypergraph

To explore the relationship between musical genres, a genre co-occurrence matrix was constructed using the top 10 most frequent genres in the dataset. For each artist, all pairs of genres they are associated with were counted, incrementing the corresponding entries in a symmetric matrix. This approach captures how frequently genres appear together in an artist's profile, offering insight into genre overlap and cross-genre collaboration.

The diagonal entries of the matrix represent the total number of artists in each genre, while off-diagonal elements indicate the number of shared artists between genre pairs -effectively quantifying genre overlap.

The matrix was then normalized by dividing all entries by the global maximum value to facilitate relative comparison across genre pairs. Additionally, the diagonal was explicitly set to 1.0, ensuring that each genre is maximally similar to itself in the normalized space.

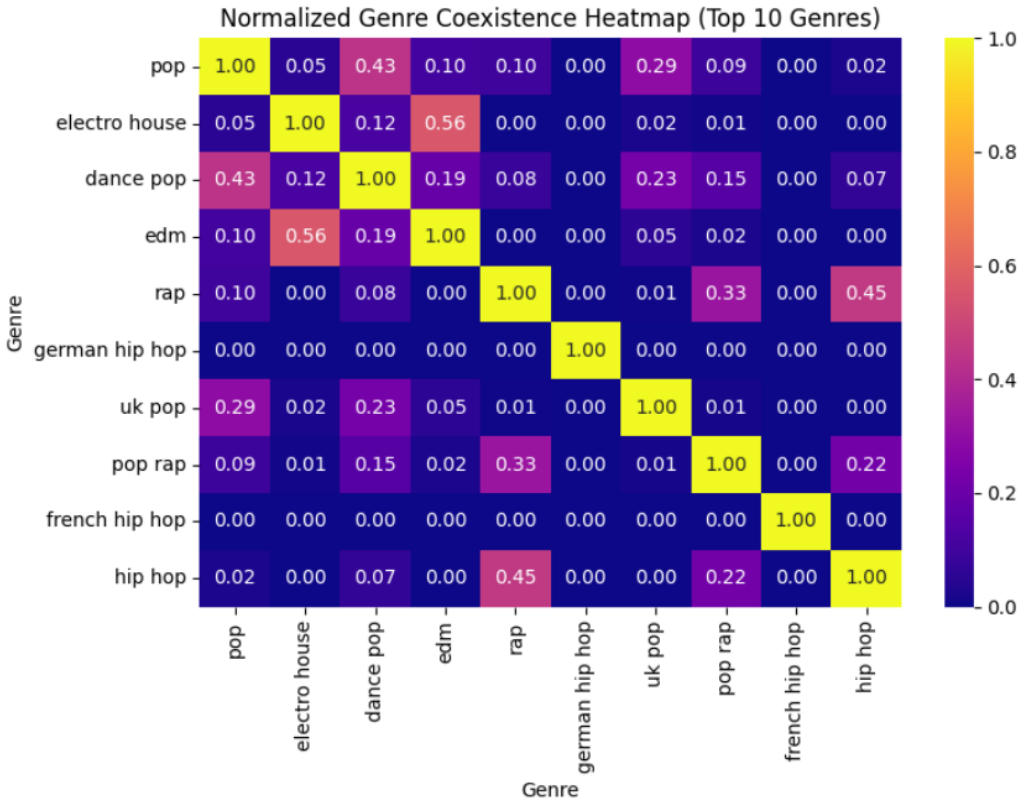

Figure 3.7: Heatmap of Normalized Genre Overlap Among Top 10 Genres

As you can see in Figure 3.7, each cell in the heatmap reflects the relative frequency with which artists from one genre collaborate with artists from another. Higher values (closer to 1.0) indicate stronger cross-genre collaboration, while lower values suggest limited interaction. Notably, genres such as electro house and EDM (0.56), as well as pop and dance pop (0.43), exhibit substantial coexistence, indicating overlapping artist networks.
We considered interesting to analyze this type of graph because it provides an alternative method to examine the closeness of genres and to be aware of the coexistence of multiple genres.

3.4.2 Genre Subgraphs

To further understand the structural properties of different musical genres within the collaboration network, an analysis was conducted to examine the relationship between genre-specific artist popularity and network cohesiveness. For each genre present in the dataset, a subgraph was created consisting exclusively of artists associated with that genre. Genres with fewer than two associated artists were excluded to ensure meaningful computation.

Two key metrics were calculated for each genre subgraph: the average clustering coefficient and the average popularity of the artists. The average clustering coefficient measures the tendency of artists within the same genre to form tightly-knit collaboration groups, reflecting local cohesion in the network. The average popularity represents the mean popularity score of all artists in that genre.
These metrics were then plotted against each other to investigate potential correlations between artist popularity and network clustering within genres:

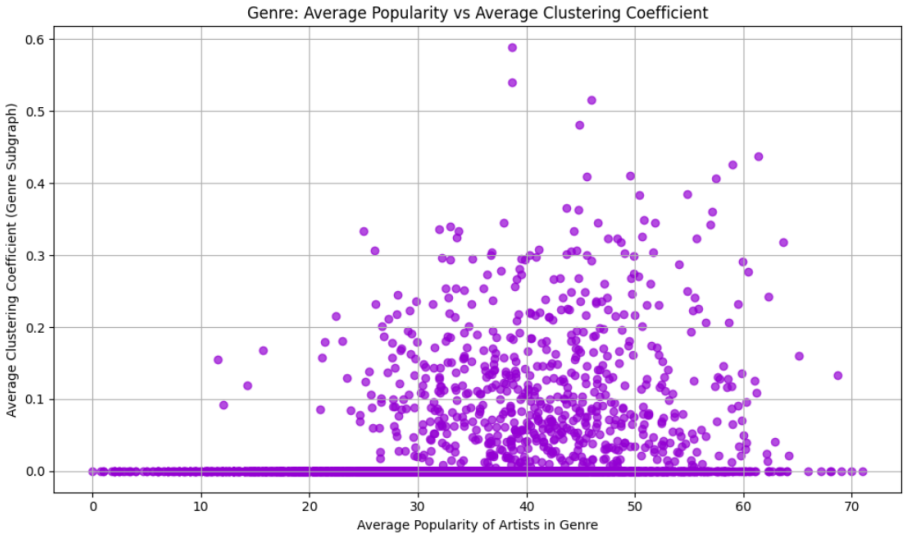

Figure 3.8: Average Clustering Coefficient vs. Average Popularity

Following the computation of average clustering coefficients per genre, an additional analysis was performed to explore the relationship between genre size and network cohesiveness. For each genre, the number of associated artists was calculated to represent the genre size. These values were paired with the previously obtained average clustering coefficients of the corresponding genre subgraphs.
We then generated this cluster to visualize the relationship between the number of artists in a genre and the average clustering coefficient within that genre's collaboration network. This analysis helps to assess whether larger genres tend to have more or less tightly connected artist communities:

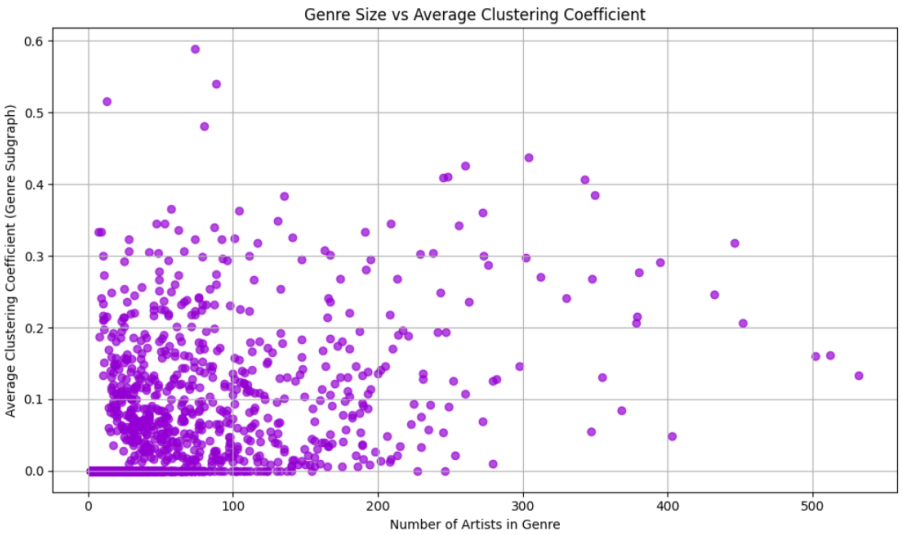

Figure 3.9: Genre Size vs. Average Clustering Coefficient

Thus, Figure 3,8 shows that average popularity in the subgraph is not correlated with the average clustering, while Figure 3.9 illustrates that the size of the genre subgraph does not correlate with a more clustered subgraph. This suggests that clustering and closeness within a genre is not a key element for creating popular music.

Chapter 4

Community Detection and Subgraph Properties

In this section, we investigate the structure of the collaboration network formed by chart-topping artists, defined as those with at least one entry in recognized music charts. After preprocessing the data to extract and convert the genres and $\mathrm{chart\_hits}$ attributes into structured lists, we isolate a subgraph of the full collaboration network containing only these successful artists. The resulting subgraph comprises 19.562 nodes and 72.207 edges.
To analyze the modular organization of this subnetwork, we employ the Louvain algorithm, a widely used method for community detection in large-scale graphs.

The Louvain method for community detection: It consists in a heuristic algorithm that aims to optimize modularity and to quantify the quality of a network's division into communities. It operates in two main iterative phases and is repeated until modularity can no longer be improved. This process results in a hierarchical community structure that captures the organization of the network at multiple scales [4].

In the first phase, each node is assigned to its own community. Then, for each node i, the algorithm evaluates the modularity gain ΔQ that would result from moving i into the community of each neighboring node j. The node i is then reassigned to the community that offers the highest positive gain in modularity and this reassignment process is applied repeatedly to all nodes in the network until no further modularity improvement is possible.
The modularity gain from moving a node i to a community C is given by:

$$\Delta Q = \left[\frac{\sum_{\text{in}} + k_{i,\text{in}}}{2m} - \left(\frac{\sum_{\text{tot}} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_{\text{in}}}{2m} - \left(\frac{\sum_{\text{tot}}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \tag{4.1}$$

where:

- $\sum_{\text{in}}$ is the total weight of links inside community C;
- $\sum_{\text{tot}}$ is the total weight of links incident to nodes in C;
- k_i is the total weight of links incident to node i;
- $k_{i,\text{in}}$ is the sum of weights of links from i to nodes in C;
- m is the total weight of all links in the network.

In the second phase, the resulting communities are collapsed into super-nodes, thereby constructing a smaller, aggregated network. In this network, each node represents a community identified in the previous phase and the weights of the links between these super-nodes are given by the sum of the weights of all edges connecting nodes from the respective communities in the original graph.
These two steps are then repeated iteratively. Each full iteration of these two steps is called a *pass*. By construction, the number of communities (now meta-communities) decreases with each pass, leading to a hierarchical decomposition of the network. The process continues until no further increase in modularity is observed. Typically, only a few passes are needed to reach convergence, and most of the computational time is concentrated in the initial passes. Due to its computational efficiency, the Louvain method is particularly well-suited for large-scale networks.

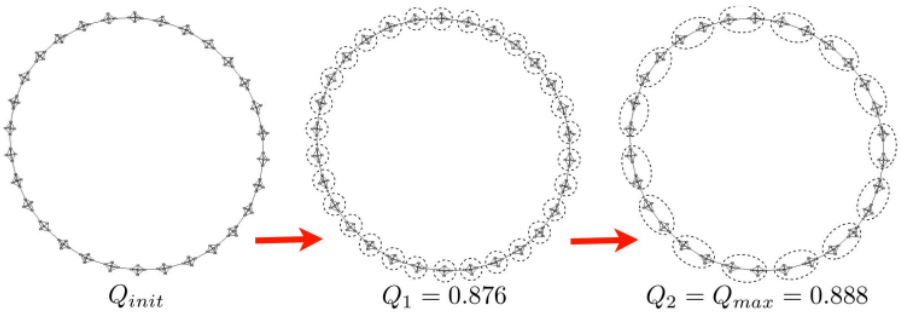

Figure 4.1: From [4]: Application of the algorithm to a benchmark network composed of a ring of 30 cliques, each containing 5 fully connected nodes, linked together by single inter-clique edges. In the initial state (Q_{init}), each node is considered as an individual community. In the first pass ($Q_1 = 0.876$), the algorithm detects each clique as a separate community, reflecting the natural partition of the network. In the second pass ($Q_2 = Q_{\text{max}} = 0.888$), pairs of adjacent cliques are merged into meta-communities, achieving a higher global modularity. This example illustrates how the algorithm builds a hierarchical community structure by iteratively merging communities to optimize modularity.

Applying the Louvain method to the network of artists, the network is divided into 5.910 distinct communities. This indicates a highly modular structure. Besides, The quality of the partition is evaluated using the modularity score, which measures the density of links inside communities compared to links between communities. The computed modularity value is 0.7729, suggesting a strong and significant community structure. Indeed, a high modularity value (greater than 0.7) indicates well-separated communities, suggesting that artists within each detected community frequently collaborate with one another but have relatively fewer collaborations outside their community. This supports the hypothesis that artist collaborations are highly structured.

The analysis of these communities reveals several key insights into community dynamics. The top 10 of these communities by number of artists is seen in Figure 4.2:

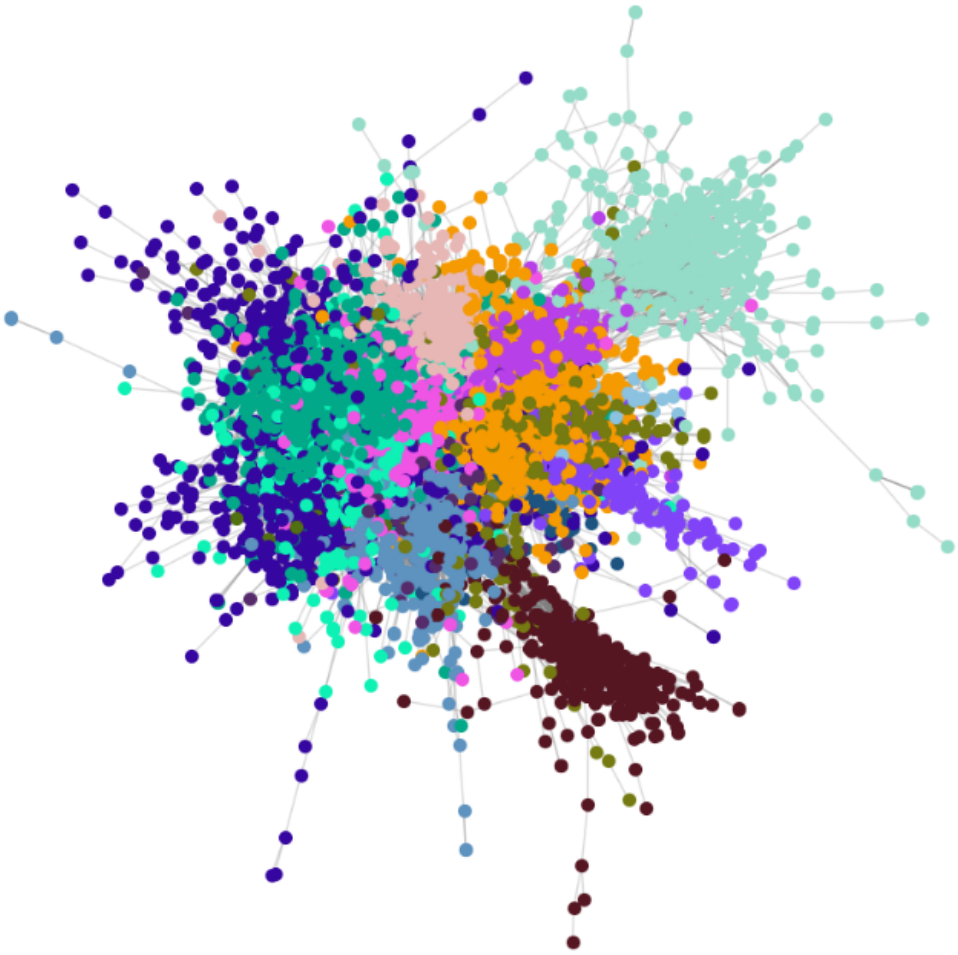

Figure 4.2: Louvain Community Detection for Top 10 communities

Furthermore, we examine the distribution of community sizes. The resulting histogram, plotted on a logarithmic scale, reveals a heavy-tailed distribution: while the majority of communities are relatively small, a few contain a large number of nodes. This distribution reflects a heterogeneous organization in the collaboration network, where a small number of large communities likely correspond to dominant music genres or collaborative circles, whereas the smaller ones may represent niche groups:

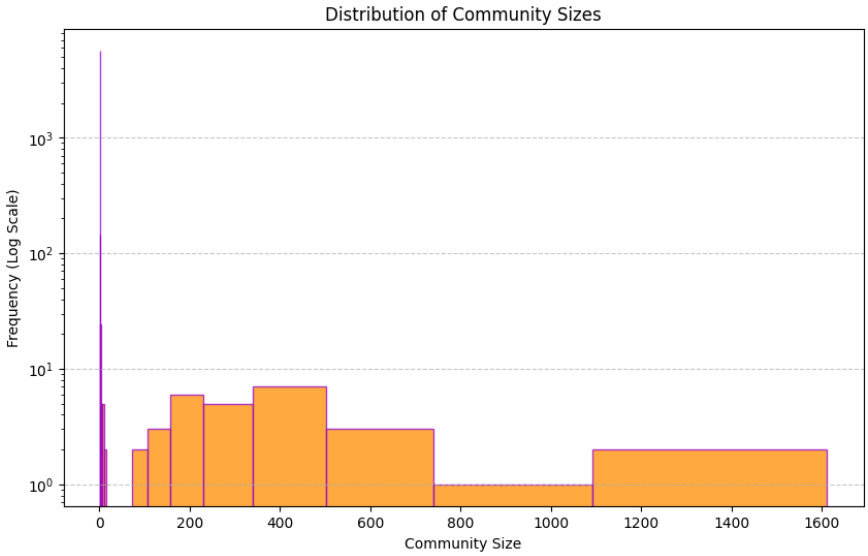

Figure 4.3: Distribution of sizes of all communities identified by the Louvain method for community detection: The distribution indicate that chart-topping artists tend to form tightly connected communities.

4.1 Centrality Analysis of the top 10 largest communities

If we consider, in particular, the top 10 largest communities identified by the Louvain method, we can focus on centrality measures in order to analyze the structural importance of individual nodes and understand the connectivity patterns within each community. Network centrality measures the degree to which a person or organization is central to a network. There are three different ways to measure network centrality [7].

- **Degree centrality**: It is a measure of the number of connections each node has in the network. In this case artists directly connected to others are considered more central. This measure indicates how many direct connections each node has to other nodes in the network.
- **Betweness centrality**: It is a measure of the number of times a node lies on the shortest path between other nodes. This measure shows which nodes are 'bridges' between nodes in a network.
- **Closeness centrality**: It is a measure of closeness or distance to others in the network. In particular, closeness centrality scores each node base on their closeness to all other nodes in the network. Nodes with high closeness centrality scores are less central and have to travel farther along the paths to get to others in the network. This measure calculates the shortes paths between all nodes and then assigns each node a score based on its sum os shortest paths.

In this regard, we computed the three centrality metrics for each node within these communities. These measures allow us to identify the most structurally important nodes (i.e. artists) within each subgraph and to compare the overall cohesion and connectivity patterns across communities. After this, For each community, the code prints the top 3 artists ranked by degree centrality, including their values for all three centrality metrics.

Using the results obtained with the Louvain Method, we can also explore the average degree centrality

Table 4.1: Summary of Centrality Measures for Top Communities

Community ID	Deg. Mean	Deg. Max	Betw. Mean	Betw. Max	Close. Mean	Close. Max
4	0.0346	0.2014	0.0050	0.0869	0.3335	0.4631
31	0.0198	0.1252	0.0037	0.0751	0.3009	0.4127
13	0.0191	0.1556	0.0046	0.0845	0.2863	0.4089
33	0.0163	0.1243	0.0046	0.1037	0.2880	0.4180
38	0.0144	0.1034	0.0062	0.1104	0.2781	0.4043
3	0.0092	0.1206	0.0017	0.0540	0.2930	0.4234
2	0.0088	0.1055	0.0017	0.0415	0.2714	0.3944
41	0.0082	0.0366	0.0091	0.1043	0.1992	0.2787
5	0.0060	0.0528	0.0044	0.1196	0.2082	0.2980
6	0.0056	0.0936	0.0018	0.0914	0.2726	0.4088

per community within the Spotify artist collaboration network. In the following bar chart, each bar represents a distinct community, as identified through the algorithm, with the corresponding $\mathrm{community\_id}$ indicated along the $x - axis$. The $y - axis$ denotes the mean degree centrality for each community. Thanks to this we can quantify the average number of direct collaborations per artist within that group.

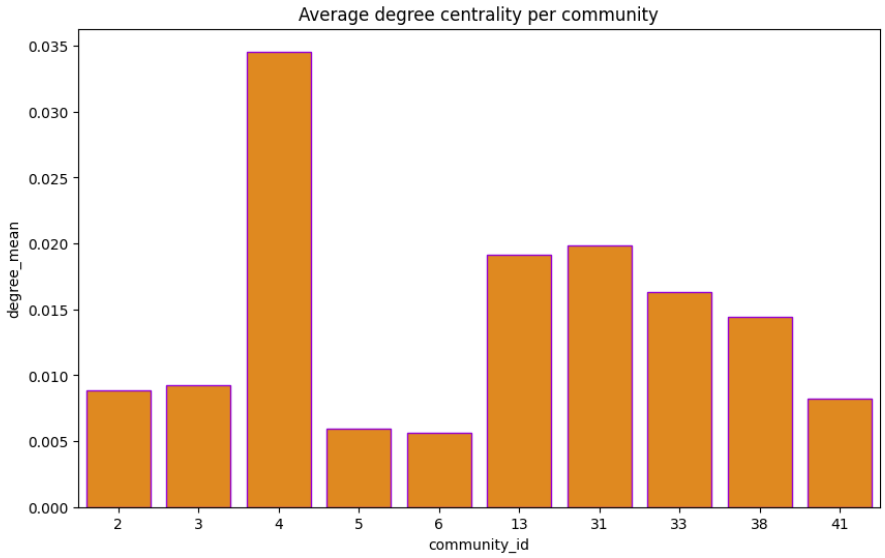

Figure 4.4: Average degree centrality for each detected community in the Spotify collaboration network.

As shown in the figure, there is significant variation in centrality across communities. Notably, Community 4 exhibits the highest average degree centrality, indicating that its members tend to be more extensively connected within the network. This suggests that artists in this group play a particularly central role in terms of direct collaborations, potentially acting as hubs that bridge other communities or facilitate widespread connectivity in the network structure.

We can finally investigate the relationship between degree centrality and popularity for the artists. The outcome shows somethign interesting: there is a positive correlation between degree centrality and popularity—artists with more collaborations tend to have higher popularity. However, the relationship

is not strictly linear.
This shows that collaboration matters. Artists who collaborate more frequently tend to be more popular: visibility, cross-genre exposure and audience-sharing via collaborations can enhance success.
From a practical standpoint, this insight can be useful for music marketing and development: fostering strategic collaborations may help lesser-known artists increase exposure and gain popularity.
Of course, there are also exceptions and outliers in the main trend: some highly popular artists have low degree centrality, indicating they succeeded without extensive collaborations. Conversely, some highly connected artists may not achieve equivalent popularity, suggesting network position is not the only factor, since also talent, branding, timing and musical genre may play significant roles.

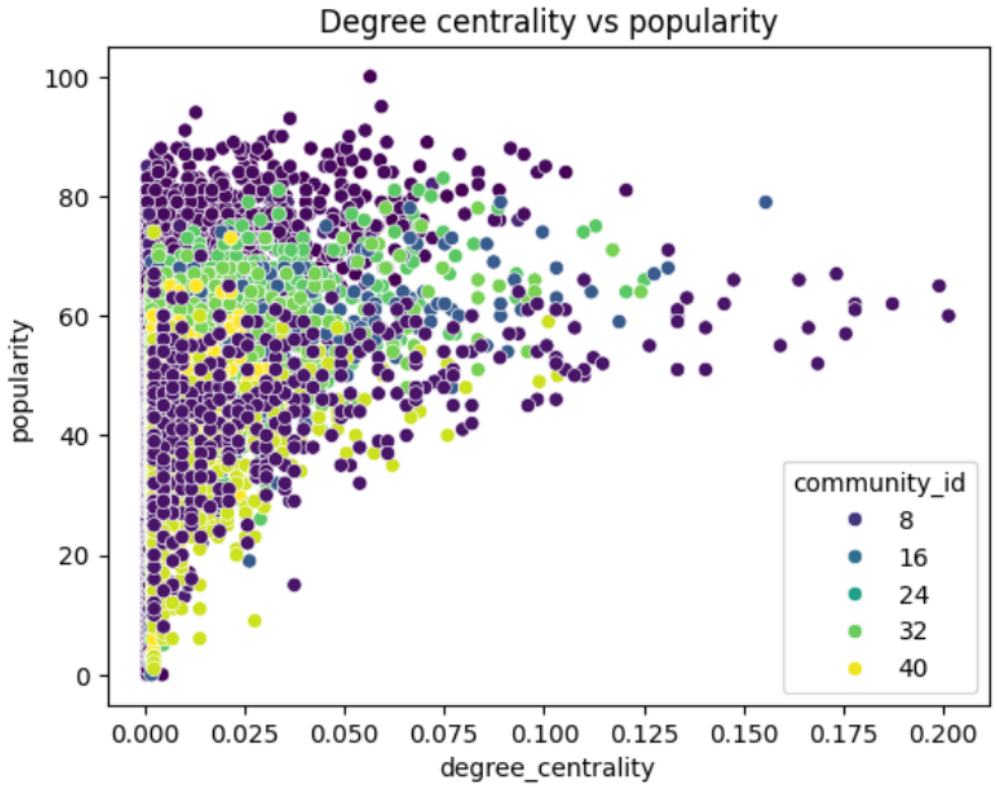

Figure 4.5: Scatter plot showing the relationship between degree centrality and popularity for artists in the Spotify collaboration network.

Chapter 5

Conclusions

The Spotify music artists collaboration graph is a unique and interesting dataset. This study investigated the structure and dynamics of the Spotify artist collaboration network using graph methods. Our results confirm that the network exhibits scale-free properties, with a small number of highly connected artists acting as central hubs.
In particular, the network's power law degree distribution highlights the role of hubs in maintaining connectivity, while centrality measures, including Eigenvector centrality and PageRank, have revealed structurally influential artists. Community detection emphasizes the modularity of collaboration patterns and the application of the Louvain algorithm revealed densely connected communities, underscoring the importance of genre and region in shaping artist network. In fact, this modular structure is largely aligned with genre and industry clusters. Moreover, genre-based analysis showed that popularity and network clustering are not strongly correlated, indicating that network connectivity is just one among several factors contributing to an artist's success.

Importantly, we found that popular artists tend to collaborate more, yet popularity is not strictly determined by network position. Some highly popular artists maintain relatively few collaborations, while some highly connected artists are not necessarily among the most popular. This underlines, once again, the multifaceted nature of success in the music industry.
While this study provides valuable insights into the collaborative patterns in digital music systems, it also has some limitations that should be acknowledged. First of all, the dataset used does not include all artists present on Spotify. It is restricted to those whose songs appeared in the Spotify weekly charts during 2022, which may introduce a bias toward more popular or mainstream artists, thereby excluding emerging or niche performers who may also play significant roles in collaboration networks. Moreover, the network is analyzed as a static snapshot, whereas in reality, collaborations, musical trends, and artist popularity are dynamic and constantly evolving. Temporal aspects such as the emergence of new artists, the dissolution of previous collaborations, or seasonal popularity spikes are not captured in this analysis. Despite these limitations, the study effectively demonstrates the power of network science in uncovering hidden structures, influential nodes, and connectivity patterns within cultural phenomena such as music, offering a valuable framework for further, more temporally sensitive research in this field.

While this study provides valuable insights into the collaborative patterns in digital music systems, it also has some limitations. First of all, this dataset does not contain all artists on Spotify, since the dataset contains artists whose songs made it to the Spotify weekly charts in 2022. Additionally, the network is treated as static, whereas collaborations and trends evolve continuously over time. However, this anal-

ysis demonstrates the power of network science in revealing hidden structures and dynamics in cultural phenomena such as music.

References

[1] Tobin South. *Network Analysis of the Spotify Artist Collaboration Graph*. https://vrs.amsi.org.au/wp-content/uploads/sites/84/2018/04/tobin_south_vrs-report.pdf. 2018.

[2] Julian Freyberg. *Spotify Artist Feature Collaboration Network*. https://www.kaggle.com/datasets/jfreyberg/spotify-artist-feature-collaboration-network. 2022.

[3] Raquel Ana Magalhães Bush. *Analysis of a Spotify Collaboration Network for Small-World Properties*. 2025. arXiv: 2503.09526 [cs.SI]. URL: https://arxiv.org/abs/2503.09526.

[4] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 2008), P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/p10008. URL: http://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

[5] Mark Newman. *Networks: An Introduction*. Oxford University Press, Mar. 2010. ISBN: 9780199206650. DOI: 10.1093/acprof:oso/9780199206650.001.0001. URL: https://doi.org/10.1093/acprof:oso/9780199206650.001.0001.

[6] Scott L. Feld. "Why Your Friends Have More Friends Than You Do". In: *American Journal of Sociology* 96.6 (1991), pp. 1464–1477. ISSN: 00029602, 15375390. URL: http://www.jstor.org/stable/2781907 (visited on 07/09/2025).

[7] Andrew Disney. *Social network analysis 101: centrality measures explained*. https://cambridge-intelligence.com/keylines-faqs-social-network-analysis/. 2020.

[8] Sergey Brin and Lawrence Page. "The anatomy of a large-scale hypertextual Web search engine". In: *Computer Networks and ISDN Systems* 30.1 (1998). Proceedings of the Seventh International World Wide Web Conference, pp. 107–117. ISSN: 0169-7552. DOI: https://doi.org/10.1016/S0169-7552(98)00110-X. URL: https://www.sciencedirect.com/science/article/pii/S016975529800110X.

[9] Jerome Myers and Arnold Well. *Myers, J.,L. Well, A.D. (2003) Research Design and Statistical Analysis, Hillsdale, NJ: Lawrence Erlbaum Associates.* Jan. 2003.