Graph-Based Analysis of Collaborations in Music: Identifying Key Collaboration Factors

Simone Caregnato
2154604
simone.caregnato@studenti.unipd.it

Giacomo D'Ovidio
2142907
giacomo.dovidio@studenti.unipd.it

***Abstract*—This project investigates the structural patterns behind artist collaborations in the contemporary music industry using a graph-based approach. By modeling the Spotify artist network as a graph, we analyze how specific artist features influence the formation of these connections. In particular, we focus on two key attributes: musical genre and record label affiliation. Using standard graph analysis techniques we explore how these features relate to network structure and cluster formation. Our results highlight musical genre as the most significant driver of collaborations, with label affiliation playing a marginal role. The study provides a data-driven perspective on how artist communities are shaped in the modern music ecosystem.**

I. Introduction

A. *Motivation*

In recent years, collaborations between artists have become a defining trend in the music industry. Popular albums and chart-topping tracks frequently feature guest appearances, and certain artists seem to reappear across numerous projects. This recurring pattern raises a key question: Are these collaborations genuinely the result of artistic choice, or are they influenced by external factors such as commercial strategies, genre alignment, or record label decisions?

This project stems from that question. Our aim is to explore the underlying dynamics of artist collaborations by analyzing them through the lens of network theory. A graph structure offers a natural way to represent these relationships, where artists are nodes and collaborations are edges, enabling a systematic analysis of how connections form.

We specifically focus on two artist features that are both relevant and accessible through Spotify's public API:

1) **Musical genre**, which may reflect artistic compatibility or audience overlap.
2) **Record label affiliation**, which can influence collaborations through contractual or promotional decisions.

By investigating how these features correlate with collaboration patterns, we aim to better understand the forces that shape modern musical partnerships.

B. *Graph Structure and Data Collection*

In our analysis, the artist collaboration network is modeled as a graph where nodes represent artists and edges represent collaborations, specifically featuring relationships. A collaboration is defined as the presence of a guest artist (featuring) on one or more tracks within the most recent album released by a given artist. To ensure the relevance and recency of the data, we include only artists who have released an album after 01/01/2023.

The graph construction process is based on data obtained from two main sources: the Last.fm API [[1]] and the Spotify API [[2]].

We begin by using the Last.fm API to retrieve a list of popular artists, as determined by their listener counts. For each artist identified, we then use the Spotify API to fetch their latest album. From this album, we extract the list of collaborators i.e. featured artists appearing on the tracks. Only collaborations with artists already present in the dataset are retained in the graph, no additional artists are added at this stage, unless the analysis follows the 1-hop expansion method described in Section I.D.b.

a) ***Attribute Collection:***

The graph's node attributes are collected in the following ways:

The musical genre of each artist is retrieved directly from the Last.fm API, which provides tags based on user interactions and artist categorization.

The record label is obtained from the metadata associated with the artist's latest album on Spotify. In cases where the album lists multiple labels or distributors, we exclude the artist from the dataset to avoid ambiguity in the affiliation. A notable example of this situation is the latest collaborative album by Future and Metro Boomin, which is released under several different labels and thus is not included in our final graph. These cases are relatively rare, representing approximately 3% of the collected dataset, and therefore do not significantly impact the overall analysis.

b) ***Attribute Normalization:***

To reduce fragmentation and ensure consistent analysis, a remapping process is applied to both genre and label attributes. This involves grouping a wide range of subgenres and sublabels into a smaller set of main categories. The goal is to capture overarching patterns without being distorted by overly specific or inconsistently labeled data.

For musical genres, many subgenres are mapped to the following top-level categories: Pop, Hip hop, Rock, Metal, Punk, Electronic, R&B, Latin, Country, Reggae, Classical, Other.

For record labels, sublabels are consolidated under their parent companies or marked as Other. The main categories are: Warner Music Group, Universal Music Group, Sony Music Entertainment, BMG, Other.

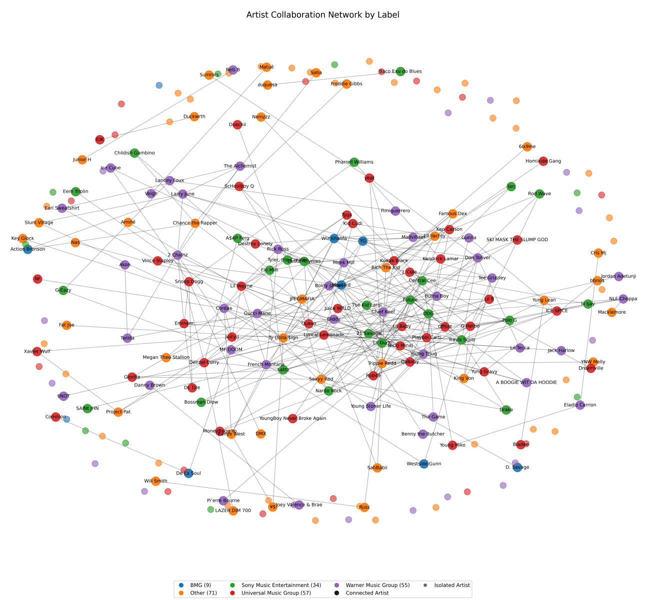

Fig. 1: Example of the resulting artist collaboration graph, with record label attributes normalized, relative to the Hip Hop macro-genre

C. *Methodology and Metrics*

To systematically examine the structural properties and attribute-based patterns across different networks, we employ a consistent framework organized into three main components. This standardized approach ensures statistical rigor and enables meaningful comparisons between networks to extract robust conclusions.

a) ***Basic Network Properties:***

Each network analysis begins with fundamental structural metrics that characterize the overall topology. These measurements include:

- **Network size and connectivity**: Number of nodes and edges, network density, and degree distribution statistics (mean, median, standard deviation, and range)
- **Component structure**: Number of connected components, size of the largest component, and its relative proportion of the total network

b) ***Homophily and Assortativity Analysis:***

The core of our analysis focuses on understanding how nodes with similar attributes tend to connect more frequently than expected by chance. This is accomplished by employing several complementary metrics:

- **Homophily ratio**: The proportion of edges connecting nodes of the same attribute value
- **Blau's Heterogeneity Index**: Measures the diversity of attribute connections for each node
- **E-I Index by attribute**: Quantifies the tendency for each attribute category to form internal versus external connections
- **Attribute assortativity coefficient**: Measures the overall tendency for similar attributes to connect, ranging from −1 (perfect disassortativity) to +1 (perfect assortativity)

Statistical Validation:

To determine whether observed patterns are statistically significant, we compare our results against two null models:

- **Rewiring model**: Preserves degree sequence while randomizing connections
- **Attribute shuffling**: Maintains network structure while randomizing attribute assignments

Statistical significance is assessed through 100 iterations of each null model type, with p-values calculated to determine if observed homophily and assortativity exceed random expectations. The comparative analysis is illustrated through histogram plots showing the distribution of homophily ratios and assortativity coefficients from null models against the observed values, clearly demonstrating statistical significance.

c) ***Community Detection:***

This final component of analysis is performed on the largest connected component of each graph. We employ the Louvain algorithm to identify communities that maximize modularity without prior knowledge of attribute labels.

To analyze the results, we compare the modularity scores between attribute-based and algorithmically detected partitions. Moreover, a more intuitive visualization is presented through percentage-based stacked bar charts, where each bar represents a detected community and colors indicate the proportion of different attribute categories within that community.

D. *Conducted Analyses*

Our analysis is applied multiple times to different graphs. Initially we examine the full network of collaborations, by focusing each on the genre and label attributes.

a) ***Second-Order Effects by Genre:*** Since genre exhibits strong clustering behavior in the global network, we further investigate whether label-based patterns become more evident when we control for genre. To this end, we isolate subgraphs corresponding to the four largest genres —Pop, Rock, Hip hop, and Electronic— and perform community detection and homophily analysis with respect to the label attribute.

This second-order analysis allows us to examine whether, within genre-specific communities, labels act as organizing principles.

b) ***1-hop Neighborhood Network Expansion:*** To better capture the complexity of collaboration patterns in the music industry, we extend the traditional direct-link approach through a 1-hop neighborhood network expansion. While the original graph only considered edges between artists in the top-N list retrieved via the Last.fm API, this expanded representation relaxes that constraint by incorporating both first-degree collaborators of top artists and the connections among those secondary artists themselves. For instance, if a top artist A has collaborated with a non-top artist B, and B has also collaborated with another non-top artist C that collaborated with A, both links A-B and B-C are included in the network.

Due to the substantial increase in node count introduced by second-degree collaborations, we take into account only 1000 top artists based on their Last.fm popularity, and then used the 1-hop expansion to include their collaborators.

II. Obtained Results

While all analyses are conducted according to the methodology outlined in Section I.C, the results are not always presented in the same structured manner. This choice is made to enhance readability and to focus attention on the most relevant findings.

For the full set of results, including all data and figures, please refer to the [**report/results**]) directory in the https://github.com/Dovid308/MusicCollabNetworkAnalysis repo, where the results are divided into subdirectories based on the graphs analyzed.

A. *Attribute: Genre*

a) ***Basic Network Properties:***

The network exhibits a sparse structure with limited connectivity between artists. The fundamental structural characteristics are summarized in Table 1.

TABLE I: Basic network properties.

Metric	Value
Nodes	1,319
Edges	925
Density	0.001064
Average Degree	1.40
Median Degree	0.00
Standard Deviation Degree	2.82
Max Degree	27
Connected Components	817
Largest Component Size	421 (31.92%)

b) ***Genre Distribution and Homophily Analysis:***

TABLE II: Genre distribution in the network

Genre	Count (%)
Pop	298 (22.59%)
Rock	253 (19.18%)
Hip Hop	234 (17.74%)
Electronic	124 (9.40%)
Other	112 (8.49%)
R&B	108 (8.19%)
Punk	49 (3.71%)
Metal	40 (3.03%)
Unknown	39 (2.96%)
Country	21 (1.59%)
Latin	19 (1.44%)
Classical	15 (1.14%)
Reggae	7 (0.53%)

TABLE III: Homophily measurements

Homophily Metric	Value
Homophily Ratio	0.5362
Attribute Assortativity Coefficient	0.3653
Average Blau's Heterogeneity Index	0.2337

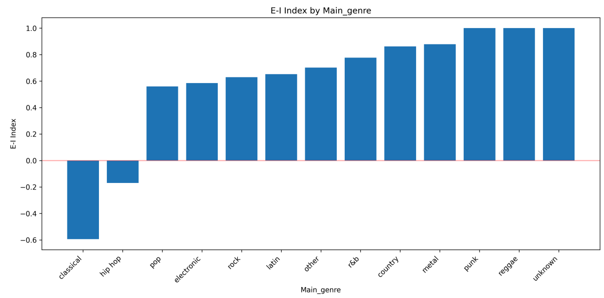

Fig. 2: E-I Index by genre

The analysis shows that genres with a strongly negative E-I Index — indicating a higher tendency to form internal connections — are Classical (-0.5942) and Hip Hop (-0.1693). In contrast, genres with higher positive values exhibit a clear prevalence of external links over within-group ties: Pop (0.5593), Electronic (0.5856), Rock (0.6296), Latin (0.6522), R&B (0.7765), Country (0.8621), Metal (0.8788), and Punk, and Reggae, each reaching the maximum value of 1.0000.

c) ***Statistical Validation:***

TABLE IV: Comparison with null models

Model	Avg Homophily ratio	Avg Assortativity
Observed	0.5362	0.3653
Rewiring Model	0.2656	−0.0050
Attribute Shuffling	0.1441	−0.0049
P-value	0.000000	0.000000

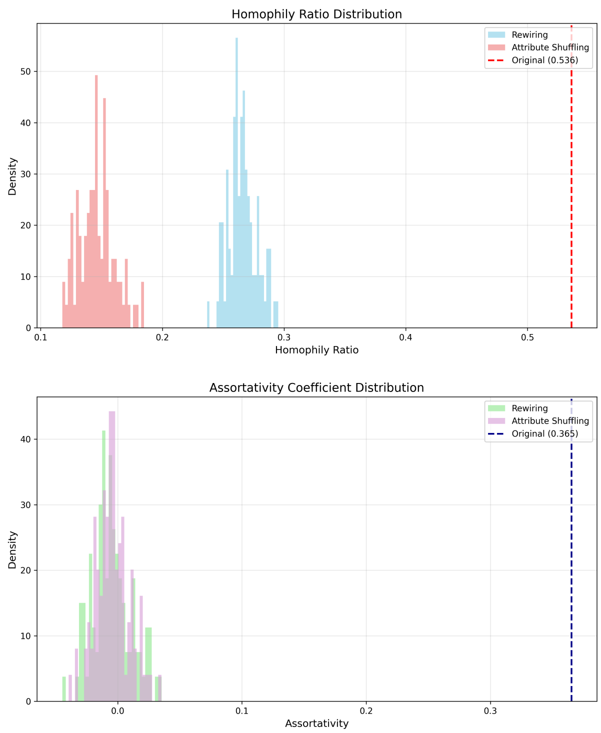

Fig. 3: Comparison of observed homophily ratio and assortativity values against null model distributions

The analysis confirms the significance of the observed homophily ratio (0.5362) and assortativity coefficient (0.3653) compared to null models. This confirms that the observed genre-based clustering is not a result of random chance.

d) ***Community Detection Results:***

TABLE V: Community detection comparison

Metric	Value
Nodes Analyzed	421
Louvain Communities	13
Genre-based Communities	12
Louvain Modularity	0.6140
Genre-based Modularity	0.1845

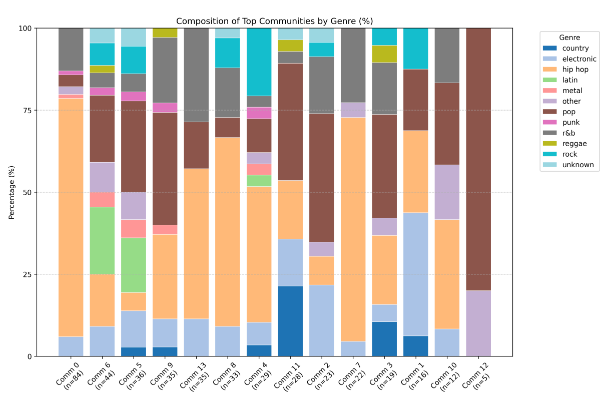

Fig. 4: Genre composition of algorithmically detected communities

The results show a substantial difference between the modularity scores of the Louvain-detected communities (0.6140) and those grouped by genre (0.1845), indicating that genre alone does not fully explain the observed community structure. While some Louvain communities are relatively homogeneous (e.g., dominated by hip hop or rock), many contain a mix of genres, suggesting that other factors may play a more prominent role in shaping artist communities.

B. *Attribute: Label*

a) ***Basic Network Properties:*** Network properties are nearly the same as for the genre network, only with a smaller number of nodes (1289) and edges (832), as we did not include artists whose last album was distributed by more than one labels.

b) ***Label Distribution and Homophily Analysis:***

TABLE VI: Label distribution in the network

Label	Count (%)
Other	456 (35.38%)
Warner Music Group	319 (24.75%)
Universal Music Group	316 (24.52%)
Sony Music Entertainment	150 (11.64%)
BMG	48 (3.72%)

TABLE VII: Homophily measurements

Homophily Metric	Value
Homophily Ratio	0.3425
Attribute Assortativity Coefficient	0.1239
Average Blau's Heterogeneity Index	0.2745

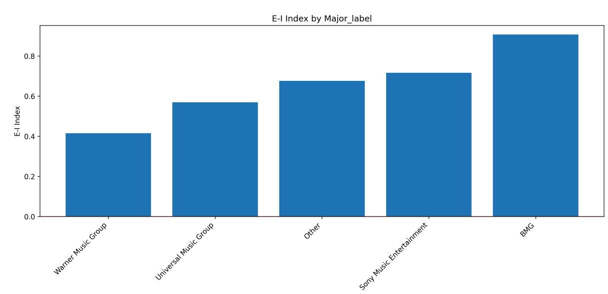

Fig. 5: E-I Index by label

The label distribution shows a significant proportion of independent artists. However, homophily metrics suggest a weak tendency for artists to collaborate within their own label. The low assortativity coefficient (0.1239) and the moderate homophily ratio (0.3425) indicate limited label-based clustering. This is further supported by the E-I Index, which remains relatively high across all label groups.

c) ***Statistical Validation:***

TABLE VIII: Comparison with null models

Model	Avg Homophily ratio	Avg Assortativity
Observed	0.3425	0.1239
Rewiring Model	0.2462	−0.0046
Attribute Shuffling	0.2607	−0.0074
P-value	0.000000	0.000000

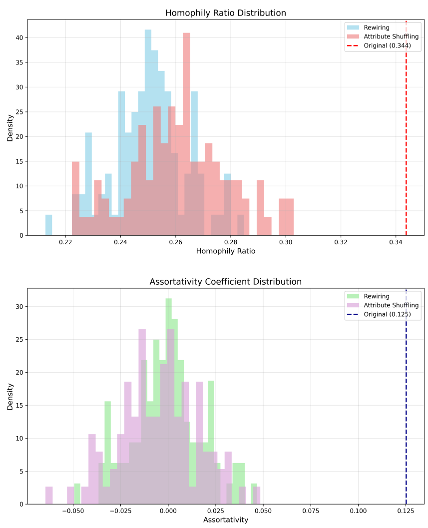

Fig. 6: Comparison of observed homophily ratio and assortativity values against null model distributions

Despite the relatively low absolute values of homophily and assortativity, statistical validation confirms that these values are significantly higher than what would be expected by chance. Both null models - degree-preserving rewiring and attribute shuffling - yield substantially lower averages, indicating that label affiliation does exert a measurable, though limited, influence on collaboration patterns.

d) ***Community Detection Results:***

TABLE IX: Community detection comparison

Metric	Value
Nodes Analyzed	383
Louvain Communities	15
Label-based Communities	5
Louvain Modularity	0.6269
Label-based Modularity	0.0457

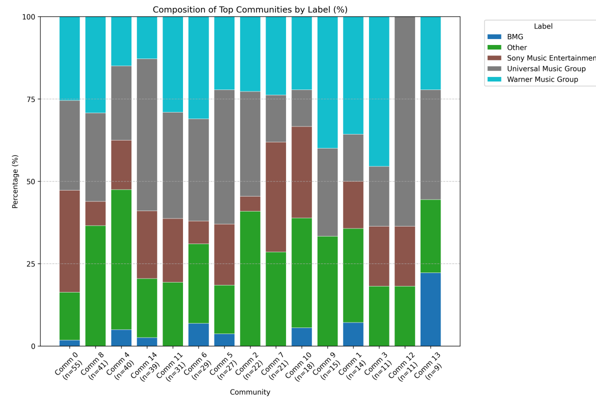

Fig. 7: Label composition of algorithmically detected communities

Community detection further supports this interpretation. While the Louvain algorithm identifies 15 communities with a high modularity score (0.6269), grouping artists based solely on their major label affiliation (as defined in Section I.B.b, where labels were mapped to 5 major categories: Other, Warner, Universal, Sony, and BMG) results in a much lower modularity (0.0457). The barplot reveals that algorithmic communities tend to mix artists from different labels, reinforcing the idea that while labels matter, they are not the dominant factor shaping collaborative structure.

C. ***Second-Order Effects: Labels Conditioned on Genre***

To assess the influence of record labels on collaboration patterns, we examined subnetworks for four major genres: Pop, Rock, Hip Hop, and Electronic. The analysis reveals that when controlling for genre, label affiliation exhibits minimal predictive power regarding collaboration structures, with most metrics failing to surpass random expectations.

The Pop subnetwork (294 nodes, 37 edges) demonstrates weak label-based clustering. The homophily ratio (0.3243) and assortativity coefficient (0.0237 versus 0.1239 in the full network) both approach random values. While Blau's Heterogeneity Index (0.0900) suggests marginally more uniform connections than the full network (0.2745), statistical tests confirm this difference is insignificant. These results indicate that label affiliation does not meaningfully structure collaborations even within this homogeneous genre.

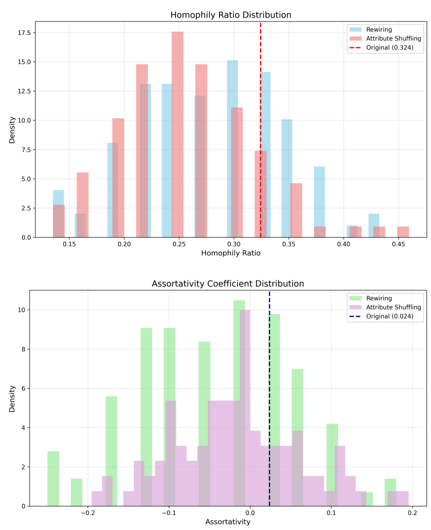

Fig. 8: Comparison of observed homophily ratio and assortativity values against null model distributions, specifically for the Pop subgraph

The Rock subnetwork (250 nodes, 9 edges) suffers from extreme sparsity, rendering its metrics unreliable. Though the homophily ratio (0.4444) appears elevated and Blau's Index (0.0667) suggests homogeneity, the near-zero assortativity (0.0110) and limited edge count prevent definitive conclusions about label effects.

In contrast, the dense Hip Hop subnetwork (226 nodes, 301 edges) provides robust evidence against label-driven collaboration patterns. Both homophily (0.2724) and assortativity (0.0455) fall below the full network averages (0.3425 and 0.1239 respectively), while Blau's Index (0.3778) indicates substantial cross-label collaboration. This contradicts industry assumptions about Hip Hop's label-centric nature.

The Electronic subnetwork (122 nodes, 20 edges) shows negative assortativity (-0.0850), suggesting a slight tendency toward cross-label collaborations. However, the limited number of edges makes this finding unreliable, and statistical validation aligns these patterns with random distributions. As with the other genres, we find no evidence of meaningful label-based homophily within this genre-specific context.

Methodological Considerations:

The selected genres (Pop, Rock, Hip Hop) represent broad musical categories encompassing diverse artistic styles. This heterogeneity may dilute potential label effects. Conversely, niche genres like Classical (represented by only 15 artists in our dataset) might exhibit stronger label homophily due to their specialized markets and tighter artistic communities. However, our sample size precludes definitive analysis of such genre-specific dynamics

This second-order analysis systematically demonstrates the absence of significant label-based homophily when examining collaborations within individual genres. While genre membership strongly predicts collaboration patterns, our findings show that record label affiliations do not meaningfully influence artist partnerships when controlling for musical genre.

a) ***1-hop Neighborhood - Genre:***

Genre Distribution and Homophily Analysis: The genre distribution remains broadly stable with 1-hop expansion, but some notable shifts emerge, e.g. hip hop increases from 17.74% to 23.47%, and the genre classical grows from 1.14% to 9.16%. This rise anticipates the high homophily we will observe for classical, where artists tend to collaborate predominantly within their own genre.

All homophily metrics increase in the 1-hop network, with the assortativity coefficient rising from 0.3653 to 0.6243.

The E-I Index pattern also becomes more pronounced, reinforcing the observation that some genres exhibit stronger internal collaboration than others.

D. 1-hop Neighborhood- Label

a) ***Basic Network Properties:***

TABLE X: Basic network properties of the one-hop label network.

Metric	Value
Nodes	1707
Edges	6756
Density	0.004640
Average Degree	7.92
Median Degree	2.00
Standard Deviation Degree	20.04
Max Degree	121
Connected Components	464
Largest Component Size	1172 (68.66%)

b) ***Label Distribution and Homophily Analysis:*** The label distribution remains broadly consistent when including 1-hop neighbors. The major labels show similar proportions, with Warner Music Group increasing from 24.75% to 27.77% and Sony Music Entertainment from 11.64% to 13.06%. Minor labels such as BMG slightly decrease, but these variations do not significantly affect the overall distribution.

TABLE XI: Homophily measurements in the one-hop label network

Homophily Metric	Value
Homophily Ratio	0.4904
Attribute Assortativity Coefficient	0.1608
Average Blau's Heterogeneity Index	0.3102

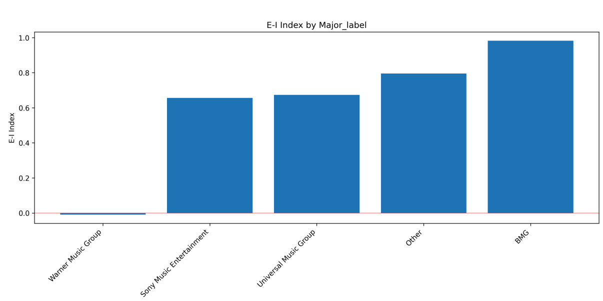

Fig. 9: E-I Index by label in the one-hop label network

Expanding the network to include one-hop connections leads to a notable increase in all homophily-related metrics: the homophily ratio rises from 0.3425 to 0.4904, assortativity increases from 0.1239 to 0.1608, and Blau's index also grows. Interestingly, in the 1-hop setting, the E-I Index for Warner Music Group turns negative, indicating that Warner-affiliated artists collaborate more within their own label than with others. This is a sign of internal cohesion and contrasts with the previous labels network, where all labels showed positive E-I values.

Statistical Validation and Community Detection: Again, the observed homophily and assortativity are significantly higher than in the null models ($p < 0.000001$). Compared to the original network, the increase in both metrics is substantial when including 1-hop neighbors: homophily rises from 0.3425 to 0.4904 and assortativity from 0.1239 to 0.1608.

However, community detection tells us something different - Louvain and label-based modularity both decrease (0.6140 -> 0.5050 and 0.1845 -> 0.0944), indicating that extending the network with 1-hop results in more mixed and less label-aligned communities.

TABLE XII: Comparison of homophily and assortativity metrics with null models, and community detection results in the one-hop label network

Model	Avg Homophily ratio	Avg Assortativity
Observed	0.4904	0.1607
Rewiring Model	0.3684	−0.0402
Attribute Shuffling	0.2605	−0.0044
P-value	0.000000	0.000000

Metric	Value
Nodes Analyzed	1172
Louvain Communities	12
Label-based Communities	6
Louvain Modularity	0.0500
Label-based Modularity	0.0943

III. Conclusion

A. Main Findings

Across all analyses, genre consistently emerged as the stronger predictor of collaboration. The genre network displayed moderate homophily (homophily ratio: 0.5362; assortativity: 0.3653). In contrast, label-based networks exhibited weaker homophily (homophily ratio: 0.3425; assortativity: 0.1239), and communities detected algorithmically were not properly aligned with actual labels.

Second-order analyses, where label effects were examined after filtering the collaboration graph based on genre, confirmed that label affiliation has minimal influence on collaboration patterns. Even in dense genre graphs such as the Hip Hop one, label-driven clustering appears to be weak.

Expanding the network by 1-hop neighborhood increased all homophily-related metrics, especially for genre, where assortativity hit 0.6243. On the other hand, the 1-hop label network showed increased homophily, but a decrease in label-based modularity.

Collectively, these analyses reveal that genre is the primary organizing principle behind collaborations. The role of record labels in defining these collaborations is much weaker, directly challenging our initial expectation that label affiliation would be a significant driver.

B. Limitations and Future Directions

We recognize several limitations in our methodology that could have influenced the results. In this section we outline these limitations and hypotethesize how they could be addressed.

Firstly, our approach to constructing the collaboration graph treated all collaborations equally, not weighting the number of times artists collaborated. A single collaboration held the same weight as joint projects. Future work could benefit from incorporating weighted edges based on collaboration frequency.

Our decision to analyze the global music market presents a second limitation. This market is in fact a composite of geographically localized markets. This inherent diversity might obscure the true influence of specific factors like label affiliation. Future research could benefit from focusing on more homogeneous markets, such as the European, US, or individual national markets.

Finally, the categorization of record labels into five macrogroups was, to some extent, arbitrary. While this simplification made our analysis manageable, it's quite possible that a more detailed classification of labels would reveal more subtle collaboration trends. Future analyses could explore this direction by refining the label categorization process.

References

[1] Last.fm, "Last.fm API Documentation." 2024.
[2] Spotify, "Spotify for Developers." 2024.