Final Report: Spotify Collaboration Network

Drew Schiller, Jesse Gabriel

"Whenever I work with different artists, I expand as a songwriter and as a producer, and I always want to try and find the bridge between my world and their world."

- Steve Aoki

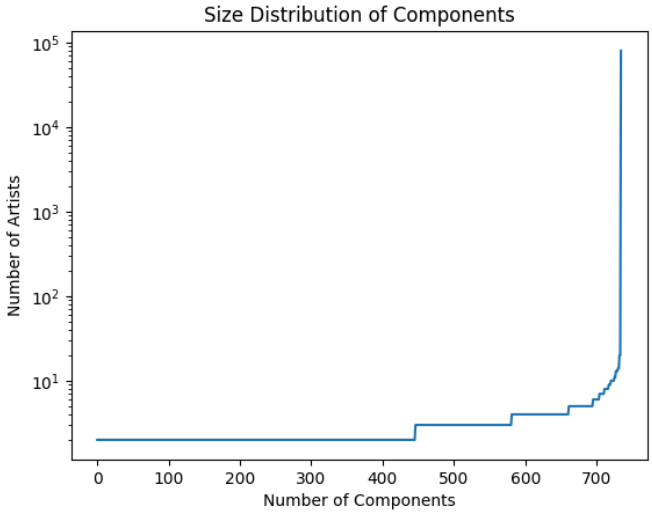
INTRODUCTION

Music is a medium that transcends language and cultural barriers, and musical collaborations commonly act as the bridges across these barriers. These bridges can paint a portrait of how connected the modern music industry is, which in turn can be used to draw conclusions about both collaborative influence as well as the musical bubbles that we as listeners often tend to. A way of acquiring this information is through a collaboration network graph. In such a graph, nodes represent artists and edges represent one or more collaborations between the linked musicians. The emergence of streaming services in the digital age has produced ample data on the numerous collaborations between artists from around the world, with Spotify providing a large dataset and powerful API that allows a collaboration network to be constructed and examined with relative ease. The purpose of this report is to thoroughly mine and analyze

the distribution of the Spotify collaboration network by extracting communities and computing various measures on both localized regions and the overall graph.

1 DATA PROCESSING AND WHOLE GRAPH MEASURES

The data used for reading in the nodes and edges consisted of two CSV files acquired from a Kaggle Spotify dataset: one listing every artist and their metadata, and one listing pairs of artists that collaborated. In reading-in these nodes and edges, we considered various filters and generalizations of the data to simplify the process of converting it into a graph. Firstly, amongst the artist metadata was a metric that Spotify uses called ‘popularity’ which scales from 1 to 100, with higher popularity artists having more listeners and streams. Given that artists with low popularity are significantly more likely to have limited data on genre, songs, and collaborations, we decided to filter out any artist with a popularity metric less than 15. Secondly, we filtered out trivial nodes, more specifically, “isolated” artists that had not had a collaboration with another artist listed in the dataset. The last consideration was made by examining the connected components of the graph of this filtered data; we found that 97.4% of the graph was in one large



connected component that consisted of 79,851 artists and 218,961 collaboration edges. For the sake of simplicity in computing other graph metrics, we treated this large component as the overall graph for the remainder of the measures.

Degree Skew and Connectivity Measures

From this filtered collaboration network, several graph mining metrics were computed to get an overview of the network's basic structure and shape. To get a sense of how typically collaborative any given artist in the network is, we computed the degree distribution of the graph as well as the average degree, which we found to be 5, indicating moderate connectivity. In the skew analysis of the degree distribution, we found both the power law coefficient and the Gini coefficient. The power law coefficient being 2.223 tells us that most artists have very few collaborations, and very few artists have many collaborations, which makes sense given how typically more popularity correlates to more collaborations, and very few artists are popular. The Gini coefficient was found to be 0.688, which also indicates uneven collaboration amongst artists in the graph. In conjunction with the high inequality of the collaboration distribution, the number of leaf nodes, or artists with only one collaboration, was also computed and found to be 39,101. This is profound, as it consists of more than half of the artists present in the filtered collaboration network, showing the distinct range of k-connectivity in the network.

Mining Connectivity of Bridge Edges

In our connectivity analysis, we looked at bridge (or cut) edges and their influence on the structure of the network. We defined the influence of a bridge based on the size of the smaller of two connected components that the bridge divides the graph into. Interestingly, the most

influential bridge we found only impacted 19 vertices out of approximately 80,000. This finding suggests that despite identifying key bridges, the overall network remains highly interconnected, indicating few vulnerabilities in its connectivity. Future analysis might consider refining the measurement of influence, perhaps by examining changes in centrality metrics when a bridge is removed, to better understand the true impact of specific collaborations.

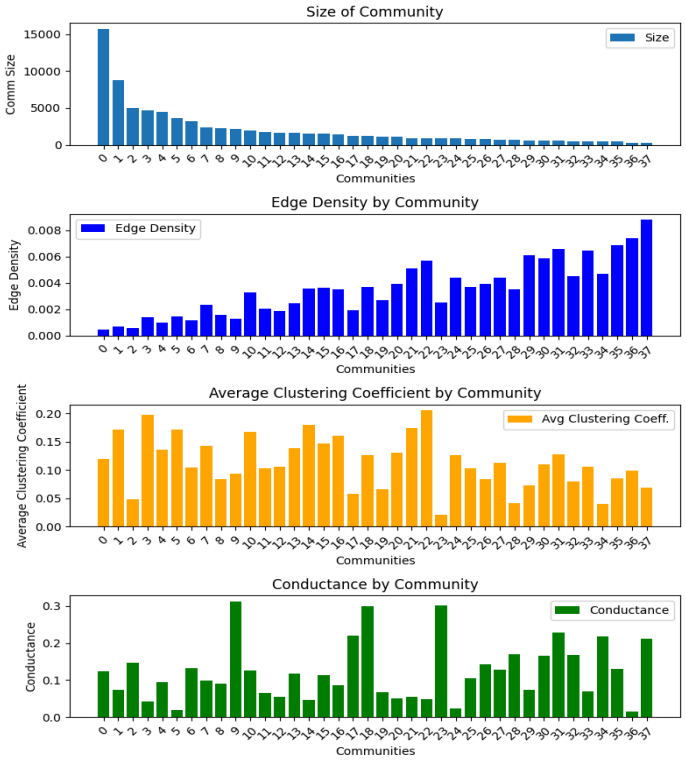
2 COMMUNITY DETECTION AND IDENTIFICATION

Our methodology for finding communities within this collaboration was primarily based upon use of the Louvain community detection algorithm. To determine how strongly structured these communities were, the network modularity with respect to the communities was calculated, and a result of 0.779 was found. This high modularity value indicates that the communities found using Louvain have a well defined community structure.

Evaluating Communities with Density Measurements

In our community evaluation, we focused on subgraphs of communities with at least 200 artists. We computed various density measures such as edge density, average clustering coefficient, and conductance. The average conductance across the communities was measured at approximately 0.115, with the lowest at 0.015 and the highest at 0.297. These values suggest that while most communities are fairly insular and exhibit strong internal cohesion, there are exceptions where communities interact more significantly with others, indicated by the higher maximum conductance. Similarly, the average relative edge density was quite low at 0.0036

across the board. This indicates that, generally, the communities are sparsely connected, and that not nearly all the edges that could exist, actually do. The clustering coefficients ranged widely, from a low of 0.0045 to a high of 0.206, with an average of 0.112. This spread again highlights varied levels of tight-knit relationships within communities—some are highly clustered, suggesting a dense web of mutual collaborations, while others are more loosely connected.



Identification via Genre Label

In our identification analysis of communities, we started by uncovering the top five genres for each community. Interestingly, we observed that community affiliations often align more with regional identifiers rather than purely by genre, contrary to our initial predictions. For example, hip hop as a genre appears extremely frequently across various communities,

commonly labeled in conjunction with a regional identifier such as “region + hip hop” or “region + rap”. Notably there were some exceptions such as “classical” and “sleep” music. The largest community lacked specific regional labels, leading us to surmise that it likely represents a broader category of American/Western popular music. The table provided offers a snapshot of these findings.

Community	Top 1	Top 2	Top 3	Top 4	Top 5
Comm. 1	pop (458 times)	electro house (450 times)	rap (415 times)	dance pop (411 times)	edm (400 times)
Comm. 2	trap latino (299 times)	latin pop (260 times)	reggaeton (254 times)	reggaeton flow (238 times)	latin hip hop (214 times)
Comm. 3	classical performance (235 times)	country (136 times)	adult standards (131 times)	classical (117 times)	mellow gold (116 times)
Comm. 4	funk carioca (342 times)	funk mtg (244 times)	sertanejo pop (170 times)	funk ostentacao (165 times)	brazilian hip hop (145 times)
Comm. 5	german hip hop (424 times)	german underground rap (193 times)	german pop (183 times)	german trap (181 times)	turkish pop (176 times)
Comm. 31	hungarian pop (127 times)	hungarian hip hop (69 times)	magyar trap (63 times)	magyar alternative (28 times)	hungarian underground rap (24 times)
Comm. 32	minimal techno (36 times)	electronica (34 times)	tech house (32 times)	microhouse (30 times)	minimal tech house (14 times)
Comm. 33	mahraganat (53 times)	egyptian hip hop (46 times)	egyptian trap (34 times)	arab pop (34 times)	khaliji (33 times)
Comm. 34	ccm (55 times)	latin christian (45 times)	latin worship (41 times)	rap cristiano (33 times)	worship (33 times)
Comm. 35	sleep (98 times)	environmental (28 times)	lullaby (21 times)	sound (13 times)	water (10 times)
Comm. 36	lithuanian pop (84 times)	lithuanian hip hop (21 times)	lithuanian edm (17 times)	lithuanian trap (12 times)	lithuanian electronic (8 times)

Combination Analysis of Genre and Density Measures

After identifying the top genres in each community, we analyzed the communities by sorting them based on edge density measures to discern patterns among the top and bottom groups. We found that smaller, more specific communities such as "sleep" music and "Lithuanian pop" exhibited higher relative edge densities and clustering coefficients but lower conductance,

indicating tight-knit, insular groups. Conversely, communities characterized by broader genres like "gaming dubstep" and “electro pop” showed lower clustering, lower edge density, and higher conductance, suggesting they are more open and have broader, less regionally confined interactions. This pattern reveals how niche interests tend to form more cohesive clusters within the network, while more mainstream or varied genres display greater connectivity with diverse musical styles.

Hierarchical Analysis of the Largest Community

In our initial analyses, we noticed that the largest community was notably huge, broadly defined, and also correlated with familiar genres in American music. Thus, we decided to conduct a further examination of this largest community. We applied the Louvain method again specifically to this subgraph and identified approximately around 15 significant sub-communities. These showed a similar size distribution skew as the initial analysis, but with lower modularity and higher average conductance scores, approximately around 0.5. Notably, the largest sub-community, which predominantly featured hip hop and rap genres, had the lowest conductance score of around 0.1. This additional genre analysis revealed more distinct genre-based delineation within the subgraph, with rap as the largest. Interestingly, some regional identifiers like “UK drill” were also observed, possibly corroborating the notion of a higher likelihood for collaborations between artists from Western countries.

3 CENTRALITY MEASURES AND ANALYSIS

In measuring the centrality of the collaboration network, we wanted to analyze it on two levels: a global analysis and subgraph analyses. However, due to the sheer computational complexity of computing certain measures on such a large graph, we decided not to consider the global centrality and instead only look at the subgraph centralities. For each subgraph, we computed betweenness centrality, eigenvector centrality, Katz centrality, as well as a composite centrality measure equal to a weighted sum of the three.

Community Subgraph Analyses

For the first community, which largely describes American popular music, the top artists for nearly all the centrality measures were hip hop or electronic artists. This can likely be explained by the tendency for hip hop artists and DJs to collaborate heavily with other musicians. Increased collaborations clearly results in more influence in the collaboration network, which explains why the largest hip hop artists are at the top. More specifically, betweenness centrality is especially dominated by electronic DJs like Steve Aoki and Diplo, whose collaborations allow for a large number of shortest paths to run through, effectively bridging diverse areas in the network. In the third community, which largely consists of classical and orchestral musicians, we noticed that even though the community is third largest in size, all of the centrality measures except for Katz are all mostly the same group of artists, namely Johann Sebastian Bach, Traditional, John Williams, Jean Sibelius, and Il Divo. This is likely due to the fact that these artists typically register classical covers as collaborations, which can cause high centrality in commonly covered composers such as Bach. Katz centrality in this subgraph likely represents the more adult contemporary portion of the community subgraph. Each additional subgraph did

have distinct centrality features, but due to our lack of knowledge in many of the specific foreign communities, we left our analyses to the largest subgraphs.

4 SHORTEST PATHS BETWEEN ARTISTS

The collaboration network could additionally be used to find paths of collaboration between two artists. By running the Dijkstra's shortest paths algorithm between two artist URIs in the graph, we can find the shortest collaboration path between the two. From this, we can run a Spotify API search on an edge by edge basis by searching each edge's two artists. This search takes in a query and response type, which in this case is a track, then responds with the top Spotify searches for the query under the given type. The top result from each search is the most search-worthy song that the two artists worked on together, and all of these songs compounded together will give a list of song collaborations that starts at the first artist and ends at the second artist. The average shortest path length for the network was 6.87 collaborations, but in practicality the pathfinder will likely result in shorter paths. The connectivity of an artist given by a user is typically higher than the average artist connectivity, since the user is typically more likely to know more popular and thus more connective artists, which results in shorter paths. As an example of running the pathfinder, for JAY-Z to Mick Jagger:

FUTURE WORK

Although we have found many interesting results, some of our techniques are still limited in various ways. We have identified several avenues of ongoing study that could potentially enhance our findings.

Comparison with Null Models

To establish the significance of our findings, future analyses will compare our network's structure to null models. This will help determine if the observed network patterns are meaningful or could arise by chance. Such comparisons could prove beneficial for validating our discovered network properties in connectivity, centrality and more. .

Exploring Edge Weights

Currently, our graph model is undirected and unweighted, which limits the depth of insights it can provide. We've brainstormed various potential edge weightings that, when integrated with our analytical techniques, could reveal deeper insights into musical collaborations. For instance, artist popularity scores (whether the minimum, maximum, or their difference) could be used as weights. Additionally, with access to specific songs, as discussed in the temporal multi-graph section, we could implement weightings based on song popularity or the frequency of collaborations. These weightings could significantly influence our path-finding results, potentially allowing us to discover the most influential paths between artists like Kanye West and less well-known artists based on cumulative popularity. This additional layer of information could greatly enhance our understanding of the network dynamics within musical collaborations.

Temporal Multi-Graph Construction

One exciting direction for future research that we've considered is the development of a temporal multi-graph model where edges represent specific tracks over time. This approach has the potential to refine our understanding of community and centrality metrics by providing a more accurate measure of artist collaborations. It could allow for the analysis of temporal effects such as the evolution of collaboration networks and triadic closure, as well as introduce more nuanced edge weightings, such as the number of songs. However, the challenge remains in accessing comprehensive track data due to limitations with Spotify's API, which may require innovative solutions or a possible collaboration with Spotify.

PROJECT ROLES

The following section describes what each member of the group contributed to the project.

Jesse Gabriel focused on analyzing the structural properties of the network by computing the distribution, connectivity, and other whole graph measures. He worked on identifying communities within the graph and classifying them based on their regional and genre affiliations. Jesse also computed measures specific to communities and subgraphs and worked in interpreting many of the results.

Drew Schiller was responsible for the initial data acquisition and processing, preparing the final graph structure used in our analyses. He computed various centrality measures and synthesized them to form a composite centrality index. Utilizing the Spotify API, Drew also developed methods to find the shortest paths between artists.

Both Jesse and Drew collaboratively developed the study's methodology, brainstormed potential research questions, and laid out the graph mining techniques to answer these questions. In addition, they both put together the final paper and the in-class presentations.