Artist Genre Diversification: A Spotify Analysis

Quinn Wai Wong, Joseph Merkadeau

Washington University in St. Louis, St. Louis, 63130

May 6, 2022

Abstract

In an age where streaming services like Spotify dominate current music consumption, the exchange of music is easier than ever. How has a specific music genre like Chicago rap become more diverse in its collaborations over time? Using datasets primarily created from the Spotify API, a dynamic network of artist collaborations from 2004-2022 can be analyzed. To assess the changes over time, we will both conduct qualitative visual assessment and calculate network measures like average clustering coefficient and genre assortativity. Expecting decreased clustering from expanding artist collaborations and yet greater disassortativity within genres will help explain if artists tend to work on music outside of their genre or not. Overall, a visual network analysis and average clustering coefficient indicate that there may be increasing integration of across-genre collaborations over time, but assortativity moderates this conclusion by displaying that dense, within-genre collaborations will only serve to become much more dense. More evidence must be collected to replicate such conclusions by implementing analyses with wider-ranging datasets and methods.

***Keywords*— Dynamic Network**, **Assortativity**, **Bipartite Graph**, **Weighted Projection**

1 Introduction

In an age where streaming services like Spotify dominate current music consumption, the exchange of music has become easier than ever [1]. Many new artists are reinventing the way we define genres. Popular Korean artists are using complex jazz harmonies and artists like Lil Nas X are combining contrasting genres in country and rap. Is the rise of streaming related to this growing diversification of genres? This paper is focused on measuring the growth of collaboration across musical genres over the years. In particular, how has a specific genre's top artists diversified in their collaborations over time?

In order to best assess these trends, it is important to understand a brief history of streaming services. Spotify and SoundCloud—two of the most popular music streaming services—were founded in 2006 and 2007 respectively, but they did not become the immensely popular platforms they are today until the mid 2010s [2]. While they had been steadily growing a user base, in 2014, Spotify opened their "Student plan" and "Family plan" subscriptions [3] and the number of active users skyrocketed. From May 2014 to June 2015, Spotify doubled their paid subscribers to 20 million and almost doubled their overall user base from 40 million to 75 million [4]. In order to capture the effect that this boom in music streaming had on artist collaborations, we decided to start our analysis in 2004 to serve as a baseline year just before these streaming services started and finish with the most current data in 2022 (up to April 2022). In our analysis, we have included the initial and current years as benchmark years, as well as 2014 to analyze the network during that streaming boom.

The Spotify API will primarily be used to form a dataset based on Chicago rap's top of artists and their collaborators. This dataset will be used to conduct a dynamic network analysis of artist collaborations from 2004 to 2022. As a qualitative first pass, the network can be analyzed according to visual inspection of the networks, highlighting how cross-genre collaborations grows and becomes more dense. This growth in cross-genre collaboration can then be justified quantitatively using metrics like average clustering coefficient and binary characteristic assortativity. Overall, modeling music collaborations as a network provides both quantitative and qualitative network measures from which to draw analyses. Doing so might underpin the correlation between music streaming proliferation and cross-genre collaboration over time.

2 Methods

2.1 Data Wrangling

To construct the artist collaboration network, a dataset on artists and a dataset of their associated tracks must be acquired. Since such datasets were not readily available, both web scraping [5] and Spotify API [6] pulls were leveraged to create them. The dataset creation process begins with web scraping the top 10 artists in Chicago rap from a Spotify-affiliated website known as Every Noise at Once [7]. Though artists' genres and their popularity scores are available on Spotify, this method required the fastest data retrieval, as Every Noise has a very clean HTML structure to scrape artists names from. Following web scraping, SpotiPy [8] was used to pull JSON files for each of the top 10 artist's tracks from the Spotify API. To do so, each artist's unique identifier (URI) was found by name, then all albums were pulled for each artist. Then, each track was pulled for each album and labeled according to the associated album's release date. This track pulling process was repeated once more for the set of featured artists, resulting in a dataset of tracks as well as a datasets of artists with genres labels. Keeping these two datasets separate was important in facilitating the network construction process.

2.2 Network Construction

Given the artists and tracks datasets, a set of weighted networks were created representing artist collaborations in the even years from 2004 to 2022 using NetworkX [9]. To create the first network representing 2004 artist collaborations, all tracks created during or before 2004 were used. Using these selected tracks, a bipartite graph was formed, where nodes were either a track or an artist and an edge represented if an artist was on a particular track. Then, a weighted projection was applied to the bipartite graph so that nodes were only artists and edge weights were the number of common tracks between two artists. This process was repeated for every other year from 2006 to 2022, forming 10 graphs to use as comparison points over time.

2.3 Key Network Statistics

Clustering Coefficient For any node in a graph, the clustering coefficient quantifies how close its neighbours are to being a clique by finding the fraction of a node's neighbours that are themselves connected. The clustering coefficient of node v_i (cc_i) and the average clustering coefficient of an entire network (cc) is

$$cc_i = \frac{2e_{N_i}}{(k_i)(k_i - 1)}$$

$$cc = \frac{1}{n}\sum_{i=1}^{n} cc_i$$

respectively, where N_i is the set of neighbours of node v_i, $k_i = |N_i|$ is the degree of v_i, e_{N_i} is the number of edges among any $a, b \in N_i$, and n is the number of nodes in the graph.

Binary Assortativity For a binary set of characteristics, let p be the fraction of nodes with characteristic 1 and q be the fraction of nodes with characteristic 2. Then the expected fraction of cross-characteristic edges in the graph is $2pq$. Let k be the actual fraction of cross-characteristic edges in the graph. If $k << 2pq$ then the graph is assortative.

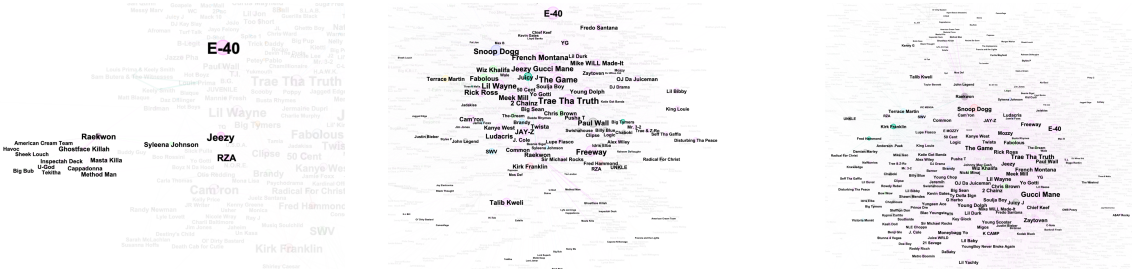
3 Analysis

3.1 Visual Inspection

An initial understanding of the patterns in the artist network can be obtained through visual inspection of the key networks in time. To capture the emergence of streaming in 2014, consider the 2004, 2014, and 2022 networks. Overall, all three networks followed a power law distribution, which was expected considering all three were collaboration networks created from an initial artist node set. In 2004, the network had 347 nodes and 882 edges, resulting in an average of 2.54 collaborations per artist. There is a small group of large nodes (node size being a measure of degree) in the center of the graph—all of which are rappers including E-40, Jay-Z, and Snoop Dogg [Figure 4]. This seems to confirm expectations. Jay-Z was a well connected figure both as an artist and manager in the music industry [10]. Snoop Dogg was also well connected at the time, being featured on Dr. Dre and 2Pac albums along with releasing his own albums featuring a variety of artists [11]. By 2014, the network increased to 1109 nodes and 6816 edges, averaging 6.14 collaborations per artist. There is now a group of around 8 large nodes, all of which are rappers [Figure 5]. By 2022, the network has 1983 nodes and 19402 edges, averaging 9.78 collaborations per artist. Now, less nodes stand out in size as most nodes towards the center of the graph appear to be very well connected [Figure 6]. Again, most of the well-connected nodes are still rappers,

where Gucci Mane leads the way sharing a track with 426 different artists. This also confirms expectations given Gucci Mane's prolific discography of over 70 mixtapes to date [12].

In terms of genres in 2004, many genres and subgenres are fairly self-contained. For example, looking at Raekwon's collaborations, we see that he is almost exclusively connected to members within the rap group he is a part of: the Wu-Tang Clan [Figure 1]. Similarly, American saxophonist Kenny G collaborates almost always with jazz and choir artists. By 2014, Raekwon and Kenny G have become more connected within the network as a whole rather than just their contained group or genre. This trend continues through to 2022, where Raekwon is now connected to a high number of nodes across the graph—still mostly rappers—and Kenny G is now connected to a wide variety of genres including rap, R&B, and K-pop. This trend is also associated with a wider variety of genres within the network, where a well-connected cluster of artists from Bollywood and related genres are also a part of the collaboration network.



2004 2014 2022

Figure 1: Wu-Tang Clan - Raekwon's connections

In summary, from an initial visual inspection of the artist networks over time, artists on average collaborated more with their peers both within and out of their own genre, isolated clusters became more connected to the entire graph, and the network reached further away genres. These conclusions served as a useful starting point for further network analyses.

3.2 Average Clustering Coefficient

To test the earlier hypothesis that artists mostly collaborated within-genre rather than across-genre, isolated clusters became more connected to the entire graph, and the network reached further away genres, the average clustering coefficient was calculated for the entire network as well as the subgraph of rap artists.

Clustering Coefficient Unique Genre Counts

Figure 2: Clustering Coefficient and Genre Counts

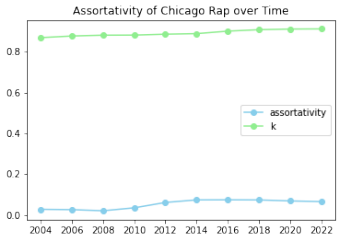
From the clustering coefficient figure, there seems to be an increase in clustering coefficient both within the genre of rap and in the network as a whole. Overall, the increase in clustering coefficient and larger clustering within rap seems to indicate that artists within rap tend to collaborate more densely compared to the entire network. This makes sense considering that Chicago rap artists were the starting point, allowing the rap cluster to become more dense compared to other parts of the network. To start, these statistics seem to agree with the visual inspection.

More specifically, there seem to be a large increase in 2006 and in 2014, as well as a large decrease in the rap subgraph in 2010. Then, from 2014 onwards, the average clustering coefficient seems to roughly even out. As for specifics, the 2006 increase can likely be attributed to the initial set of artists used. Because Kanye West was the only artist of the top 10 initial artists to have released an album in 2004, the sharp increase in 2006 may be because now more of the top 10 artists were used in the construction of the graph. As for decreased clustering of rap in 2010, this might indicate that the set of artists within the genre increased, but their collaborations decreased. Although a significant result, this remained an unanswered question in the exploratory analysis.

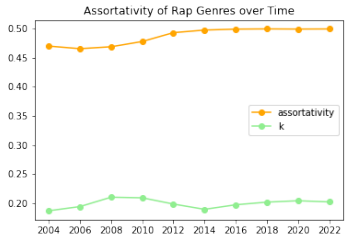
The main result of the average clustering was the sharp increase in 2014 and its fairly constant value afterwards. We predict this is likely because of streaming services' takeover as the main source of listening and sharing music. As services like Spotify gained popularity among artists and listeners leading up to 2014, artists were able to collaborate with each other more easily, leading to a sharp increase in well connected nodes. Once streaming services were established as the main source for music, collaborations still increased but not as drastically. In those post-2014 years, it is possible that the increase in unique genres countered the increased collaborations resulting in a constant clustering coefficient. So far, the visual inspection seems to align with the results of the average clustering coefficient over time.

3.3 Assortativity

From the clustering coefficient analysis, it seems as if both the rap subgraph and the network as a whole are becoming more dense in collaboration. But how much of this increased density is occurring within-genre versus across-genre? Looking at the networks we noticed that even though many new genres entered the graphs in the later years, these genres appeared to be fairly self-contained and only weakly connected to the main section of rap artists. To assess this claim, the assortativity of both Chicago rap and rap in general was calculated and plotted over time.



Chicago Rap



All Rap

Figure 3: Assortativity of rap genres over time

Comparing Chicago rap to all other genres, Chicago rap artists seem to be highly disassortative, where the actual percentage of cross-characteristic edges (k, 80%) is much higher than the expected number of cross-characteristic edges (assortativity, <10%) by nearly an order of magnitude through every measured year. Given that Chicago rap is a relatively small genre compared to genres like pop or alternative R&B, this makes sense, as Chicago rap has a much larger pool of artists to collaborate with outside of their genre than inside.

Comparing rap to all other genres, the opposite trend can be seen, where rap is slightly assortative in genre. Opposite to the Chicago rap assortativity, a large number of artists in the network are likely rap artists, so rappers likely have a larger pool of rappers than pool of non-rappers in the network to collaborate with. In turn, this means that there should be a larger proportion of rap-to-rap edges in the network, thus decreasing assortativity. Furthermore, after 2014, our benchmark year for the takeover of streaming services, the actual fraction of cross-characteristic edges stays constant near 38%, while the expected number stays constant near 50%. Even though these graphs are well connected, over 60% of the edges stay within genre. All in all, while our network does expand greatly and reach many more genres, the collaboration of the majority group of rap artists tend to most frequently collaborate with themselves, causing there to be.

4 Discussion

4.1 Conclusion

The main goal of the network analysis was to determine how genre diversity is reflected in artist collaborations over time, especially with the rise of streaming in the mid-2010s. This was done by first creating an artist dataset and tracks dataset from mostly the Spotify API, which enabled the construction of artist-to-track bipartite networks and then weighted artist collaboration networks from 2004 to 2022. Qualitatively, the artist network began with the existence of notable hubs as well as mostly self-contained genres. As the network reached a greater number of genres over time, these self-contained clusters of genres seemed to become better connected with the rest of the graph, possibly indicating stronger across-genre collaborations. Quantitatively, this was reinforced by overall increase in average clustering coefficient with a notable increase in 2014. The assortativity provided counterevidence to such a strong conclusion, suggesting that there was larger growth within-genre compared to across-genre collaborations. The lack of significant change in assortativity in 2014 helps portray the possibly nuanced effects that streaming services may have had on genre diversification. Overall, a visual network analysis and average clustering coefficient indicate that there may be increasing integration of across-genre collaborations

over time, but assortativity moderates this conclusion by showing that more measures must be used to relate the changes in popular music consumption to artist collaborations.

4.2 Limitations

Given the strong qualitative claims made from visual inspection, one main limitation was the connection of the qualitative claims to strong quantitative evidence. Although the clustering coefficient determined that both within and across-genre collaborations increased over time, assortativity itself was not enough. Two attempts to deepen the analysis was thresholding edge weights and using a Holme-Kim model.

Edge Weight Thresholding In terms of methods, one main drawback to the analysis was the minimal analysis of the weighted edges. Given that edge weights indicate the number of tracks in common between two artists, edge weights could be an important consideration in determining strength of collaborations between artists. Since artists collaborate most frequently with other artists in their own genre, removing any edges with a weight below a certain threshold from the graph should result in separate connected components, ideally representing unique genres. Unfortunately, the results were not as simple as this. When doing edge thresholds on the 2022 graph with a weight of 1, the fraction of cross-characteristic edges for rap vs non-rap genres slightly increased from 0.3836 to 0.3949. For a threshold between 2 to 6, this fraction decreased towards 0.29. For any threshold larger than 6, the network was too sparse to gain meaningful insight on assortativity. Unfortunately, there was no data to support the idea that such thresholds edge weights could be used as a form of community detection, As a result, edge weights are less important to the network than expected, as such thresholds were not too useful as a form of community detection. More could be investigated to make use of edge weights as a analysis for increasing strength in cross-genre collaborations.

Holme-Kim Model Furthermore, we wondered if we would be able to model the real-world network growth using a preferential attachment model. Instead of the traditional Barabási-Albert model (BA model) [13], the Holme-Kim model (HK model) [14] was investigated so that two artists that already exist in the graph can collaborate and create an edge. In other words, the HK model is the BA model with triadic closure. Even adding a number of random edges equal to the average degree for the following year's graph and adding a high 0.99 probability of triadic closure, each average clustering coefficient was much lower than the real-world network. Despite accounting for triadic closures, this still served as a poor model for the real network. This is likely because nodes (artists) are added at random one at a time in the HK network. On the contrary, due to our proposed network construction, each artist in the real network both adds themself as a node and their featured artists. As such, a different baseline model or a different method network construction may be important to better model the real-world network.

Data Constraints The other main limitation was data constraints. In terms of data constraints, despite having such information in their app, the Spotify API only provides information on artists as collaborators. This means that prolific producers or songwriters like Rick Rubin are not. Rick Rubin is an especially good example of this, as his collaborators span the likes of punk rock, rap, pop, and country, making him an especially influential character in musical genre diversification [15]. The data was also limited, as a large proportion of artists lacked any genre labels and only artists' albums were used to generate tracks. Leaving specific artists and specific tracks out may underestimate the across-genre collaborations, as unlabeled artists likely have unique genres and non-album tracks like singles could likely be one-off, genre-diverse collaborations. Tying in data constraints to the methods, only Chicago rap's top 10 artists were used to create a dynamic network model, making the analysis fairly rap-heavy. Being able to create various genre-specific networks would help ensure that this analysis generalizes regardless of the initial genre. Overall, a comprehensive analysis of the review indicates potential effects of the rise of streaming on artist genre diversification, but a greater variety of data and alternative methods must be used to corroborate this finding.

GitHub repository accesible at `https://github.com/cse416a-sp22/final-project-qwong-jmerkadeau`

References

[1] M. C. Götting, "Music streaming statistics & facts." Nov. 2021.

[2] D. Yassin, "A brief history of streaming services." Dec. 2019.

[3] B. Peterson, "Spotify's growth chart should be the primary user manual for startups building businesses on the web." Apr. 2018.

[4] E. Kim, "Why spotify is worth more than 8.5 billion, explained in one chart." June 2015.

[5] L. Richardson, *Beautiful soup documentation*, 2007.

[6] "Web API reference Spotify." 2022.

[7] G. McDonald, “Every Noise at Once Chicago rap.” Apr. 2022.

[8] P. Lamere, *SpotiPy*. Paul Lamere Revision, 2014.

[9] D. A. S. Aric A. Hagberg and P. J. Swart, “Exploring network structure, dynamics, and function using NetworkX,” in *Proceedings of the 7th Python in Science Conference (SciPy2008)* (T. V. Gäel Varoquaux and J. Millman, eds.), (Pasadena, CA USA), pp. 11–15, Aug. 2008.

[10] A. Light and G. Tate, “Hip-hop in the 21st century.” Feb. 2021.

[11] “Snoop Dogg.” May 2022.

[12] “Gucci Mane.” Apr. 2022.

[13] A.-L. Barábasi and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, Oct. 1999 [Online].

[14] P. Holme and B. J. Kim, “Growing scale-free networks with tunable clustering,” *Physical Review E*, vol. 65, Jan. 2002 [Online]. doi: arXiv:cond-mat/0110452.

[15] M. Ray, “Rick rubin.” Mar. 2022.

5 Figures

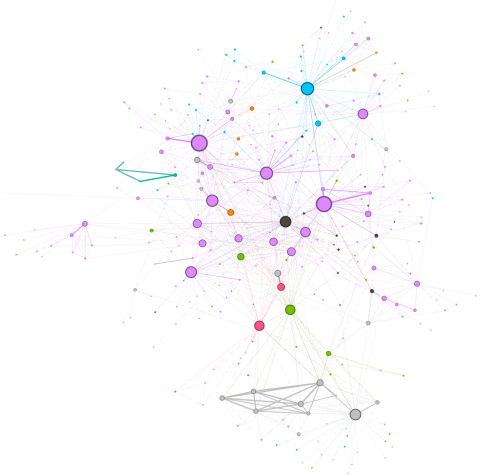


Figure 4: Full artist graph, 2004.

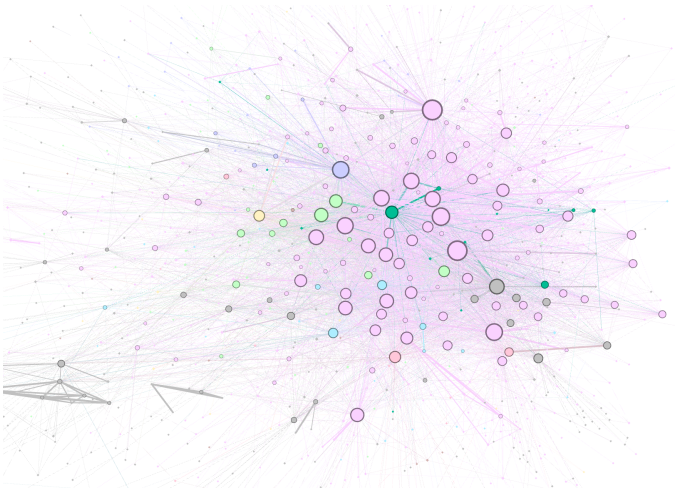


Figure 5: Full artist graph, 2014.

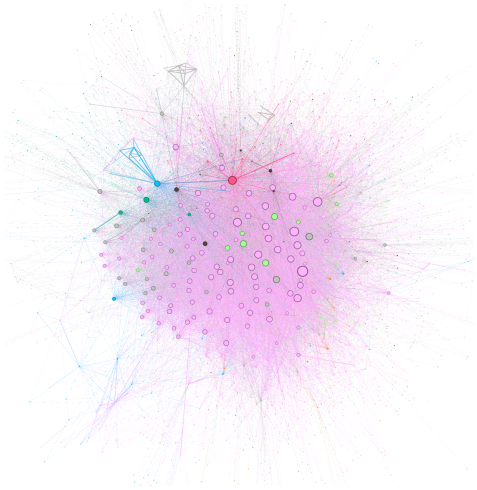


Figure 6: Full artist graph, 2022.

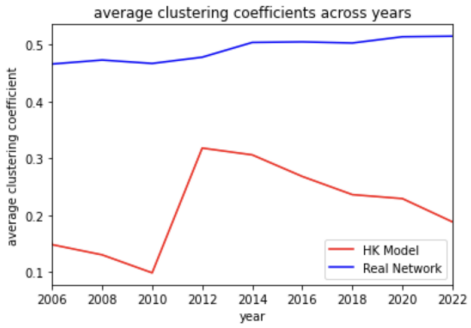


Figure 7: Average clustering coefficient of HK model vs real network.

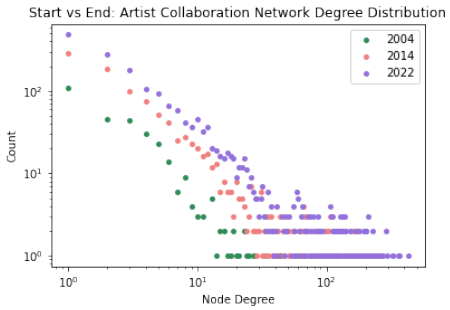


Figure 8: Power law degree distributions by year