Mapping Genre Preferences Onto A Social Network

By Corbin Diaz

Introduction

Social networks play an important role in determining the views and preferences of the people inside them. One question that arises is how large of an impact a social network has on the listening habits of the people within. A reasonable hypothesis is that genre preference would largely group around social ties, as this is certainly true if one gets new music mainly from others within their social circle. In the Karate club study (Zachary, 1977), it was found that strong connections only strengthen the bonds within subgroups. In this context, this would mean people are more and more likely to take their strong ties recommendations seriously and weaker ties less seriously, causing subgroups to polarize in music taste. However, Salganik and Watts (2009) explain how popularity can have a strong influence on whether or not you listen to it, which indicates that considering only local social networks for genre preference may have its limitations. Additionally, this rise of listening apps, music recommendation algorithms, and increasingly complex and niche genres have only made the question “What music do you listen to?” and “Where do you get your music?” more and more complicated.

This paper attempts to answer whether music preference is related to local social networks, through various computational methods and analysis of graphed networks.

The Social Network

The “Extended Chapin” Sample Set

19 participants were asked to answer questions regarding their listening habits, sources, and connections to other participants in the study. An explanation of the questionnaire they were given is outlined in Appendix A. These participants come from my main social circle, which I call the “Extended Chapin Network”. A majority of the participants come from Northwestern students who are currently or have once lived in the Chapin Humanities Residential College. All other participants are close friends of those from Chapin, causing most of them to visit Chapin frequently and become honorary residents of Chapin. As such, this friend group is known to have distinct subgroups within them, connected within by strong ties and to each other by weaker ties, with some lying more on the outskirts of one or more of the subgroups. If it is because of this anticipated knowledge of the social network (and the availability of the participants to me to ask for responses) this social circle was chosen.

Visualization

From the directed graph (Graph C.1), one can immediately see the emergence of two subgroups: The top one, with nodes labeled 0--8, which I refer to as the “Inner Chapin” subgroup, and the bottom one, with nodes labeled 9-18, which I refer to as the “Outer Chapin” subgroup. This lines up with my knowledge, as the Inner Chapin subgroup contains mostly students still currently living in Chapin, while the Outer Chapin subgroup contains students who have either moved out of Chapin or are friends with people from Chapin. The two graphs are

mostly connected by nodes 5 and 8, which also lines up with my knowledge. These participants are frequently in and out of Chapin and frequently talk with those outside of Chapin. The directed graph also gives an insight into how "tight" each subgroup is within and between each other. The Inner Chapin subgroup has many strong ties going in and out of each node, suggesting this to be a very closely knit group. The Outer Chapin subgroup, however, has fewer strong ties leaving each node, despite receiving many strong ties from both groups (such as nodes 9 and 17). This suggests not only that this subgroup may be less close, but also that they view themselves as being less close with Inner Chapin, despite Inner Chapin viewing themselves as being close with them. One theory for this asymmetry is simply that Inner Chapin could be more friendly and more likely to call a connection a close tie. Another theory is less memorable contact. Everyone from Outer Chapin has visited those in Inner Chapin, and due to the small size of the Chapin dorm, this makes them more memorable in Inner Chapin's eyes. The Outer Chapin subgroup, however, may view themselves as having closer ties outside of the Chapin dorm.

This graph, however, is only structured around strong ties. This makes weak ties very hard to see. To remedy this, an undirected graph (Graph C.2) was created by only including connections that went both ways. This graph makes it easier to see the two subgroups. Additionally, the only strong connection remaining between the groups is through node 8, further showcasing how the Inner Chapin group has more connections towards the Outer Chapin group than vice versa. This graph also showcases the weak connections between the groups more clearly, indicating nodes such as 0, 6, 7, 9, and 17 all hold crucial connections between the two subgroups.

The Music Network

The Trouble With Genre

Trying to categorize someone's music taste by genre is, by itself, an impossible task Rentfrow and Gosling (2003) attempted to categorize music based on the structure of the music itself, while Greenberg et al. (2016) decided to categorize music based on emotional and psychological experiences reduced down to just three categories. But both of these approaches fail to fully capture people's taste profile by having too limited and abstractly created categories that only account for a fraction of someone's music experience. Spotify genres have the reverse problem of having too many or and too niche genres making it virtually impossible to understand or compare ones music taste. Genres are also assigned with underlying bias, as some artists like Frank Ocean who have complicated genre profiles are assigned only a simple subset of genres while other rock or country artists are assigned a plethora of different categories.

Overall, the main problem lies in trying to categorize listening habits into either too general or too abstract genre categories. Instead, it would be more accurate to define genre groups based listening habits themselves. For example, "pop" would not be a useful category if everyone listened to mostly pop. One may then try to define abstract subgenres of po, like "energetic pop" or "complex pop", but these would either be too niche or too arbitrary to make concrete divisions with. Instead, one could observe the genre listening habits of people (using some basic genre information) and find that some share specific genre "mixtures", such as some clustering around "Pop-Rock-Indie" and some clustering around "Pop with Hip-Hop and some Electronic". These clusters would then act as "genre categories", and people would be assigned accordingly. This is similar to how Spotify compares songs with one another in order to

determine similarity, and in the same vein we can compare the overall listening habits in order to define “genre” itself. Despite only starting with some basic genre categories, the resulting mixtures are more accurate in capturing shared complex music tastes and creating sortable groups.

Alternatively, by quantifying participants' music tastes in this way, one could measure how similar each participant's genre profiles were to each other. Based on how strong this similarity is, one could assign weak or strong links between participants and create a “Music Network” that measures similarity in genre preferences. An outline of how this hypothesis was computationally implemented and how the music network and genre clusters were assigned can be found in Appendix B.

Visualization

The music network (Graph C.3) shows connections based on similar genre profiles, with thick lines indicating strong similarity and thin lines indicating weak similarity. The nodes are then colored based on which subgroup they lie in. Here, we can see that there seems to be almost a linear progression in music similarity, as the network forms almost an arrow shape. However, the subgroups do not map nicely to the music network. Rather than one subgroup clustering around one part of the music network, it seems evenly dispersed. This seems to suggest that social subgroups do not cluster around a similar genre profile. One thing of note is that the Inner Chapin subgroup seems to be on the “periphery” of the music tastes of the Outer Chapin subgroup. This may suggest that the Inner Chapin genre tastes are influenced by the Other Chapin ones, and genre tastes propagate away from Outer Chapin to Inner Chapin. However, it could also mean that Outer Chapin genre tastes are highly clustered and the Inner Chapin genre

tastes are less so. The arrow-like structure also suggests some sort of spectrum of genre tastes, with clear clusters in this music network. This network in turn was used to verify the genre clusters formed.

The Social-Music Network

Visualization

Here (Graph C.4), genre tastes were mapped onto the original undirected social network, using the genre clusters calculated. The original hypothesis that genre clusters would correspond to social subgroups seems to further be disproved. Some genre clusters can be found throughout the network, such as “High Chill, High Pop, .. ” and “High Pop, Moderate Indie, …”. However, there is still some clustering of music tastes in the network. In general, more Hip-Hop genres seem to lie on the outskirts of the network, while more Indie genres lie more near the center of the network. Pop genres lie scattered throughout. This seems to indicate that there is some correlation between the indie genre and lying in the center of the social network. In particular, note that the three High Indie nodes 6, 9, and 17 have some of the highest number of weak ties between the two subgroups. This seems to indicate that there is some relation between listening to indie music and having many connections that mediate between two subgroups.On the other hand, Hip-Hop nodes tend to have less nodes overall in the network. Whether genre taste causes social ties or the other way around is still hard to determine from the network alone.

How Music Flows In The Network

Participants were also asked where they get their music from, ranked 1 to 4, among people in the social network, people outside the social network, algorithms, and other. The percentage of participants that selected each of these sources as either their primary or secondary source (both were included to increase variation) for each genre cluster is included in Table D.1. Overally, many participants indicated algorithms being their first or second choice, and of note those with High Indie heavily favor algorithms despite being so well connected in the graph. Additionally, those with High Hip-Hop that lie on the outskirts of the graph seem to favor those within the network. This may indicate that overall people find Hip-Hop music more through others, and people find indie music more through algorithms. However, no one genre seems to correlate with any one source, which indicates that the favoring of one source over another in relation to music taste is much more complex.

Additionally, a markov-chain analysis was conducted to better understand how genre tastes flow throughout the model. This analysis assumes that music taste flows probabilistically, with music transferring with a higher probability to nodes with stronger ties (with the highest probability of transfer being back to the node itself). This process is then viewed over 1000 iterations to get an idea of where genre tastes eventually end up. In the end, it was found that no matter where a music taste originates, it ends up at a High Pop or High Indie node about 72% of the time, with just High and Moderate Indie nodes being 55%. In other words, if music taste flowed through the network probabilistically, it ends up at High Indie or High Pop nodes a majority of the time. Of course, music taste likely does not flow this way, and whether or not someone shares their music taste (such as a new song they found) varies from person to person and may be more strict. However, this still shows that there is some correlation between

high-connectedness and the indie and pop genres, and suggests that people with this genre taste are more likely to be taking information in from all over the network.

Conclusion

In the end, it is found that although there is largely a lot of variation in genre preference within the social network, those with more weak ties between distinct social circles are more likely to gravitate towards indie genres, while those on the outskirts tend more towards hip-hop. High connectedness within the social network seems to be linked with those who prefer High Indie or High Pop genres. However, indie genres are roughly more likely to indicate getting their music from algorithms, and hip-hop genres are more likely to indicate getting it from people within the network. Overall, there seems to be a correlation between indie and connectedness. Yet, it hard to determine whether this is a result of high connectedness leading to people having many sources of genre and forming an “indie” taste, or if the indie taste itself is likely to lead to higher connectedness socially. Additionally, it is hard to quantify where music tastes originate from, as whether participants indicate they get music from the social network or algorithms does not seem to correlate largely with any particular genre and seems to go against how genres are situated in the social network.

There are also notable limitations in these computational methods. Many of the directed connections in the social network were ignored to their complexity which may affect how the social network is structured. Categorizing complex music tastes into a set of basic genres, no matter the size, introduces bias in both how the genres are categorized and what the categories are, and this is only further complicated by Spotifies bias in generating genres for each artist. Different results may be found for slightly different genre categorization. Focusing on these

computational methods also tends to treat music taste as a quantifiable noun rather than an active verb that changes over time, and these results could change as participants' music tastes change with time, place, and emotion.

Appendix A: Data Collection

Each participant was given a questionnaire that asked for two main pieces of data: their listening habits, and their social ties. There are many ways to quantify listening habits, but this report considers genre due to its ability to get a simple yet robust overview of someone's music taste. The best way to ask for someone's listening habits was to use Spotify Wrapped, Apple Music Replay, or some other music app's data, since it was not only accurate but for many people readily available. Unfortunately, Spotify Wrapped 2024 does not have genre data, so instead the participants were asked for their top 5 artists. Genre data was then calculated based on these artists in a process described later on.

The second piece of data was their social ties. Participants were asked to rank other participants either 0 (no tie), 1 (weak tie), or a 2 (strong tie). However, I worried that since all of these participants knew me, and knew that I would have access to their responses, there would be some "guilt bias" in my results. To remedy this, I (1) excluded myself from the study, and (2) participants were not asked directly to differentiate between weak and strong ties. Instead, as suggested by my close roommate Benjamin Auby, participants were asked "would you want to go to a concert one-on-one with this person?". The idea is that this would correlate closely with "friends" and "close friends", and while the different phrasing may cause some response bias (some people are more willing to go to concerts with strangers than others, etc.), this bias is still helpful for the study. This is because whether or not you are comfortable going to a concert with someone is thought to closely line up with whether you would consider their listening habits or not. This allowed us to quantify strong ties more easily. Weak ties were then defined as people you have interacted with but wouldn't go to a concert with.

Participants were also asked several “dummy questions” (name a real life person, or name an artist you want to see in concert) in order to prevent them from figuring out what the questionnaire was trying to measure, as this may have influenced results. Among people they had to rank connections with was also “Alex Taylor”, who is non-existent, and was used to ensure participants filled out the survey accurately by filling in a “0”. All this data was discarded.

Participants were also asked to rank where they got their music from, either from people in the study, people outside the study, algorithms, or other. This data was used to better understand how music preferences flowed throughout the resulting network.

Appendix B: Genre Computations

Genre Vectors

To use the clustering of mixtures hypothesis mentioned and create the genre vectors, one would first need genre data to understand listening habits in the first place. As previously mentioned, genre data could not be retrieved directly from participants themselves, and instead was calculated through their top 5 artists. To achieve this, artists were put through the website everynoise.com through a web-scraping algorithm in order to retrieve spotify genre names for each artist. The resulting genres varied widely, so to make comparisons between them they had to be manually searched and binned into a reasonable number of groups that could be measured (results given in Table D.3). I found that all genres could be sorted into one of 10 groups: pop, hip-hop, latin, chill, electronic, instrumental, rock, indie, folk, and country. Each genre was then given a genre score. For example, listening to Taylor Swift for your #1 would add +5 to the "pop" genre, while having Charli XCX as your #3 would add +3 to both "pop" and "electronic". These scores were then arranged in a 10 dimensional vector and then normalized, creating a "genre vector" that allowed music taste to be mapped as a point in space.

This method is not without its limitations. For example, some genres are less clear than others: particular, "chill" acted more as a "catch-all" for genres like r&b, jazz, and piano, while genres like "latin" or "indie" have less concrete definitions of what defines them. These groups were made mostly based on my personal experience and understanding of the artists and genres themselves, while also trying to prevent creating too many categories or putting genres into categories it only weakly fit into. For example, trying to fit latin into other genres runs the risk of losing styles like norteño or banda that are deeply rooted in latin culture. Additionally, Spotify

often gives too many or too few genres to each singer. However, this was mostly remedied by the recategorization, as most singers had only 1-3 genres listed. Still, this method is still prone to overgeneralizing artists to one genre. This was attempted to be fixed by putting some genres into multiple categories, such as “modern indie pop” fitting into both indie and pop.

Genre Adjacency Matrix

Genre similarities were computed using each genre vector. By taking the dot product of each genre vector (analogous to a cross-correlation between two genre profiles), one could obtain a “similarity score” between two participants’ genre tastes. A higher product indicates two vectors more similar to each other, while a product closer to 0 indicates two genre vectors that share zero similarity. This was done for each pair of two participants. These scores were then graphed in order to determine thresholds for weak and strong similarity scores. This showed three main peaks: one from 0-0.55, one in 0.55-0.75, and a final uptick after 0.75. By this, I determined a mapping of no similarity, weak similarity, and strong similarity respectively. This created the genre adjacency matrix, which was used to create a music network mapping strong and weak ties in similarity in music preference.

K-Means-Clustering

The genre vectors were then clustered together using a k-means clustering algorithm. In the process, node 1 was removed from the data for being a high outlier, having a higher “latin” component than anyone else, indicating that putting latin separate from other genres could run into some issues. Additionally, the 10 dimensions were shrunk down to only 2 principal components in order to prevent a mathematical issue known as the curse of dimensionality from skewing the algorithm. These components generally corresponded to either “Hip-Hop” or

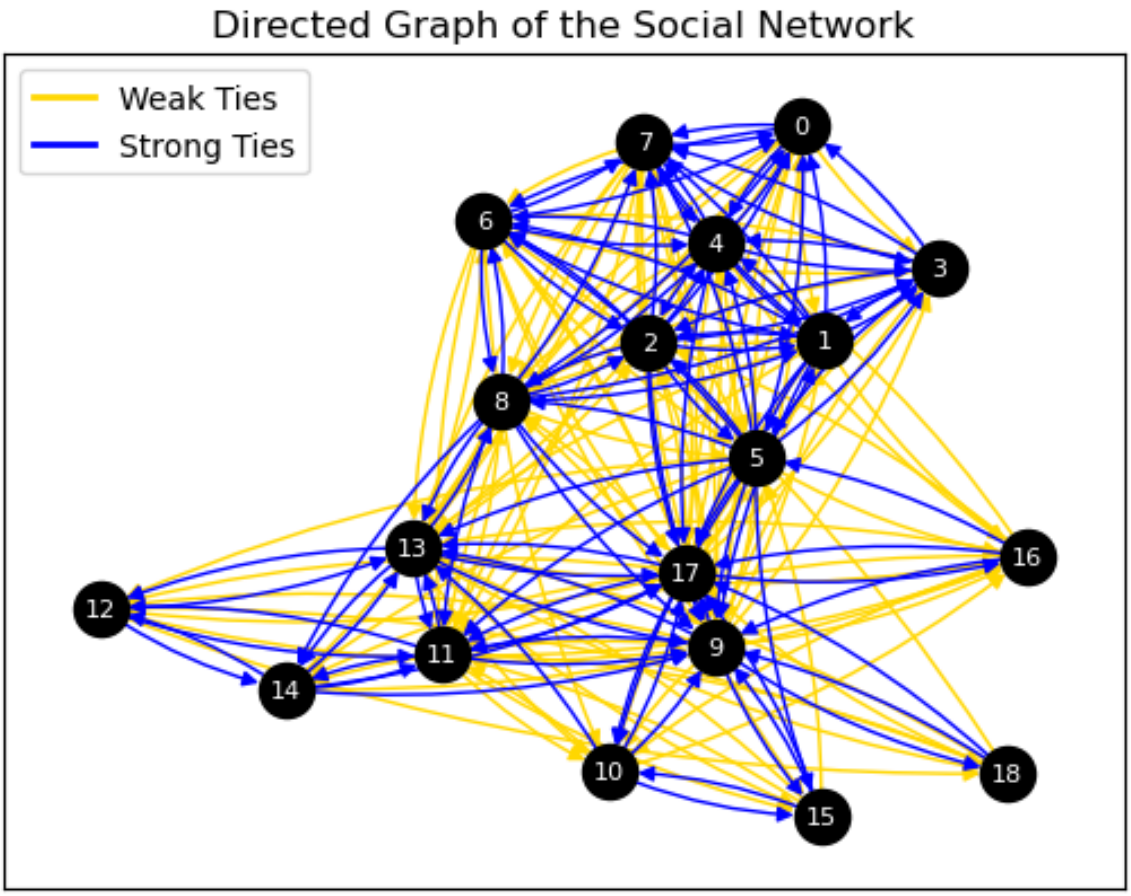
“Pop-Rock”, and all music tastes could be mapped out onto this space (shown in Graph D.5). Despite reducing a lot of information, this process retained nearly 75% of all variation in the data (shown in Graph D.2), and produced results that were verified visually by graphing the groupings onto the music network and confirming that they corresponded to placements on the graph (Graph D.3). This reduction also indicates that the listening habits of the group studied most varied in either Hip-Hop or Pop-Rock; other samples may have listening habits that produce different results. After these steps, an optimal clustering amount was confirmed by simple error analysis (Graph D.4).

The result is 5 clusters, plus the outlier node 1, creating 6 genre groups. In order to better understand these genre groups, the average of all the genre vectors of the nodes in each group were studied and analyzed to create a simple description. For example, a description like “High Pop, Moderate Indie, Moderate Rock, Some Folk” indicates that on average, pop scored higher than 0.4, indie and rock scored higher than 0.2, and folk scored higher than 0.1 (these thresholds were determined by roughly studying each vector for what seemed important).

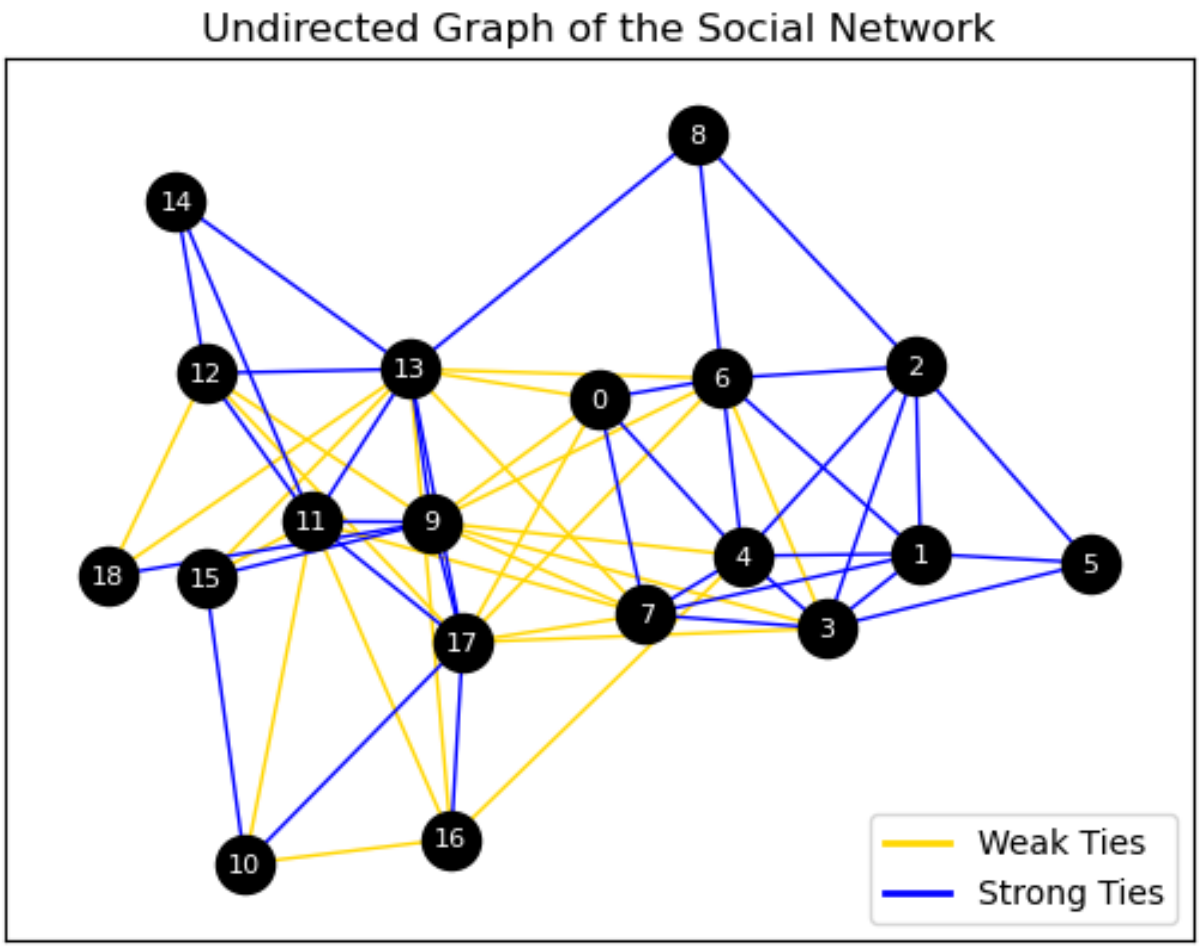
Appendix C: Networks

All but one of the networks in this paper were drawn by the Fruchterman-Reingold force-directed algorithm, which pulls and pushes nodes based on connectedness. Nodes that are close to each other in the graph represent nodes that are close to each other through strong and weak ties, with strong ties weighted more. However, due to the size of the directed graph, the force-directed algorithm caused nodes and edges to bunch up on each other, yielding visually unappealing results. Instead, the Kamada-Kawai path-length cost-function was used, which similarly tries to equate distance between nodes based on how connected they are. In particular, this cost-function only takes into account strong ties to better showcase the subgroup structure, and the weak ties were then superimposed onto this drawing.

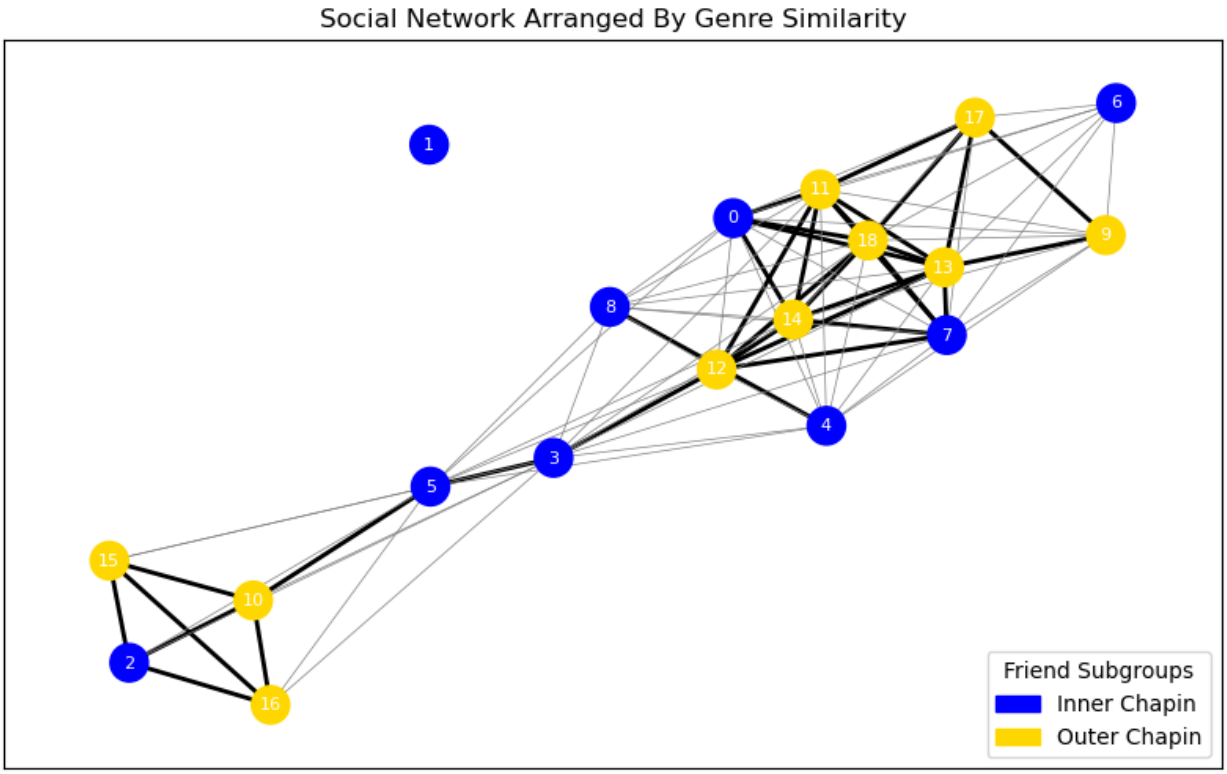
Graph C.1


Directed Graph of the Social Network
Weak Ties
Strong Ties
0
7
6
4
3
2
1
8
5
13
16
17
12
11
9
14
10
18
15

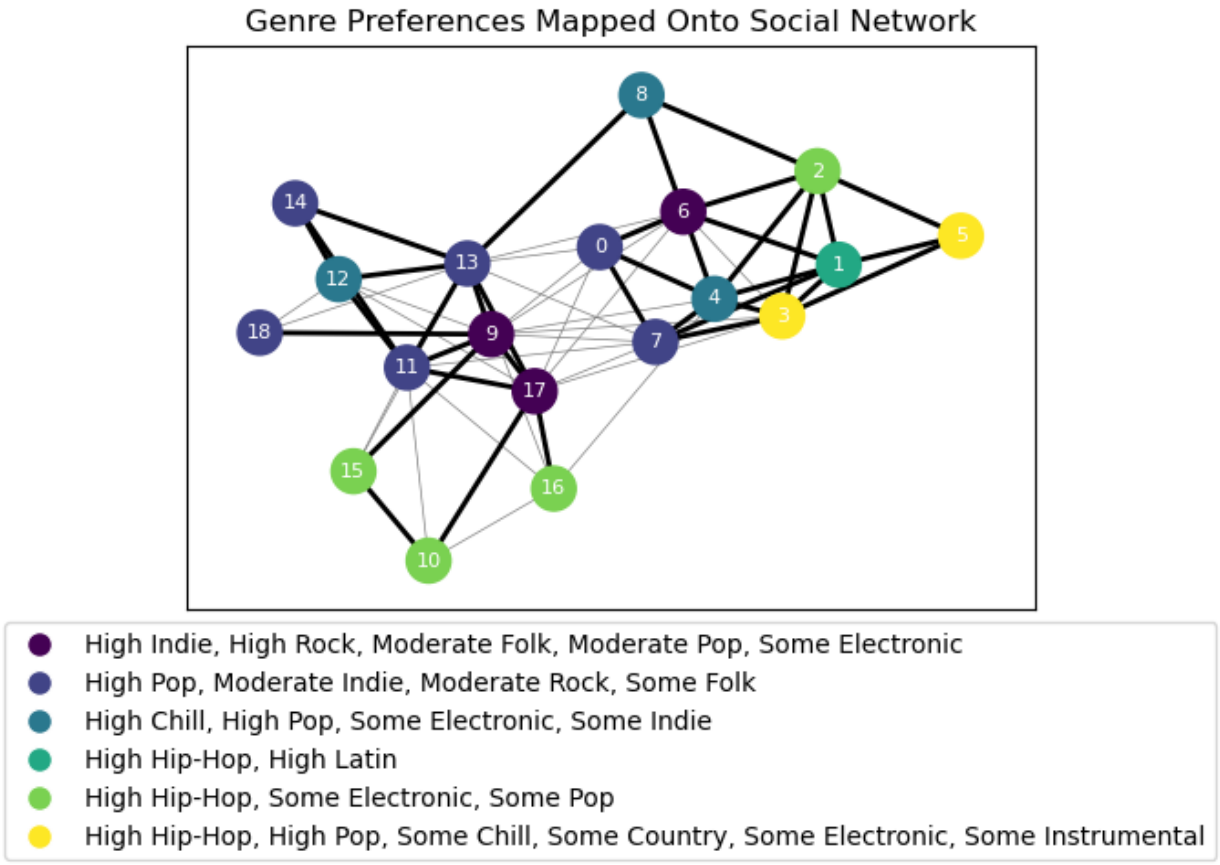
Graph C.2


Undirected Graph of the Social Network
8
14
12
13
0
6
2
11
9
4
1
5
18
15
7
3
17
16
10
Weak Ties
Strong Ties

Graph C.3


Social Network Arranged By Genre Similarity
1
6
17
11
0
18
9
13
8
14
7
12
4
3
5
15
10
2
16
Friend Subgroups
Inner Chapin
Outer Chapin

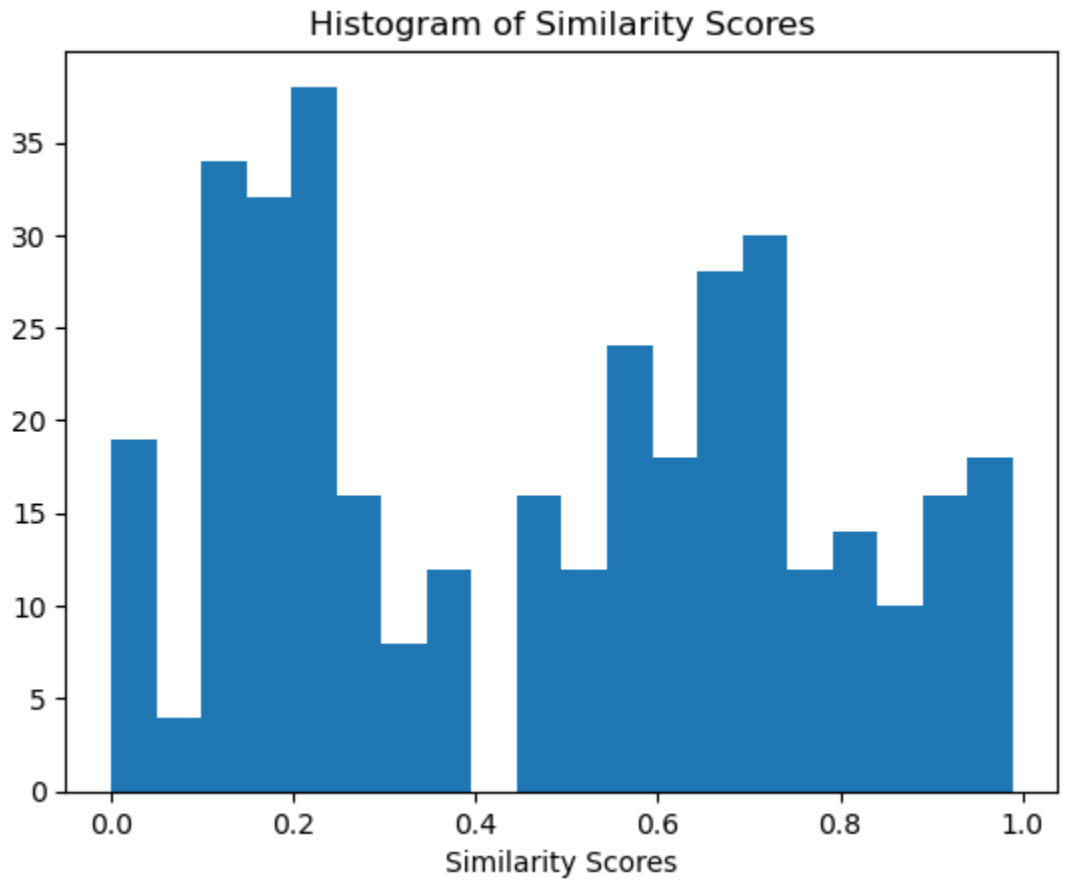
Graph C.4



Graph C.4 Caption: *Note that genre clusters are arranged in descending order of their indie score in their genre vector, from "most indie" to "least indie".*

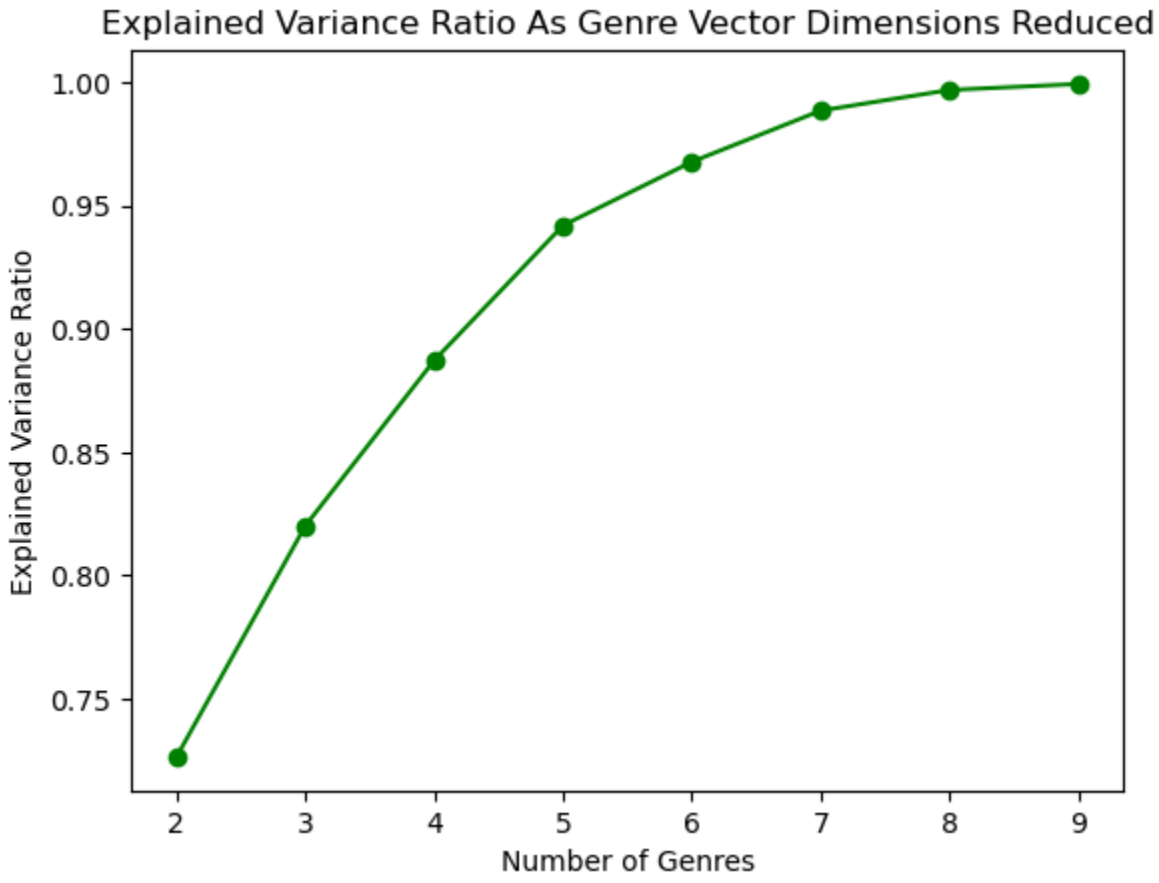
Appendix D: Other Graphs and Important Data

Graph D.1



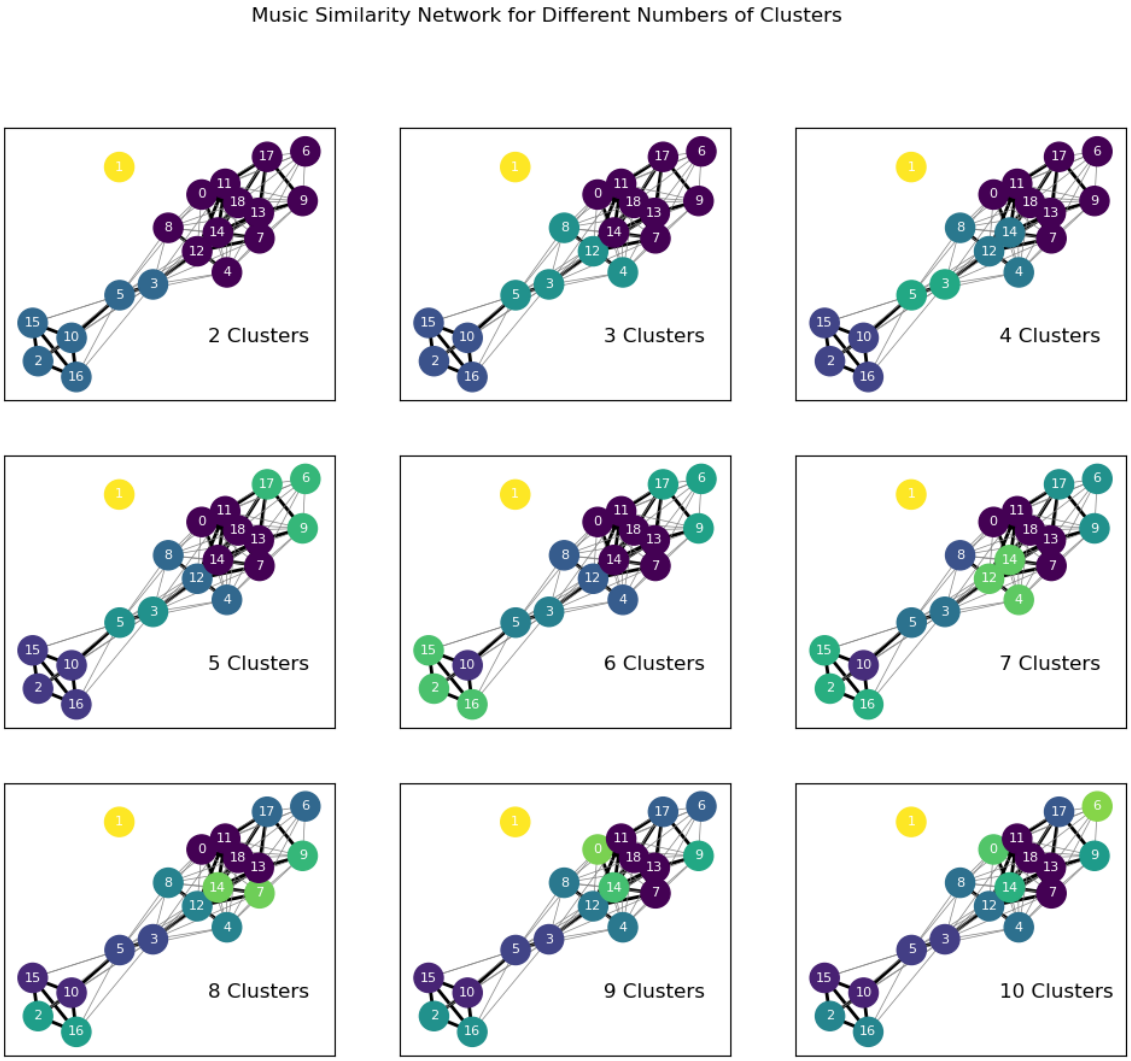
Graph D.1 Caption: *The trimodal distribution indicates that similarity scores are clustering around no similarity, weak similarly, and strong similarities. The cut off for these groups were determined to be 0.55 and 0.75 by the distribution.*

Graph D.2



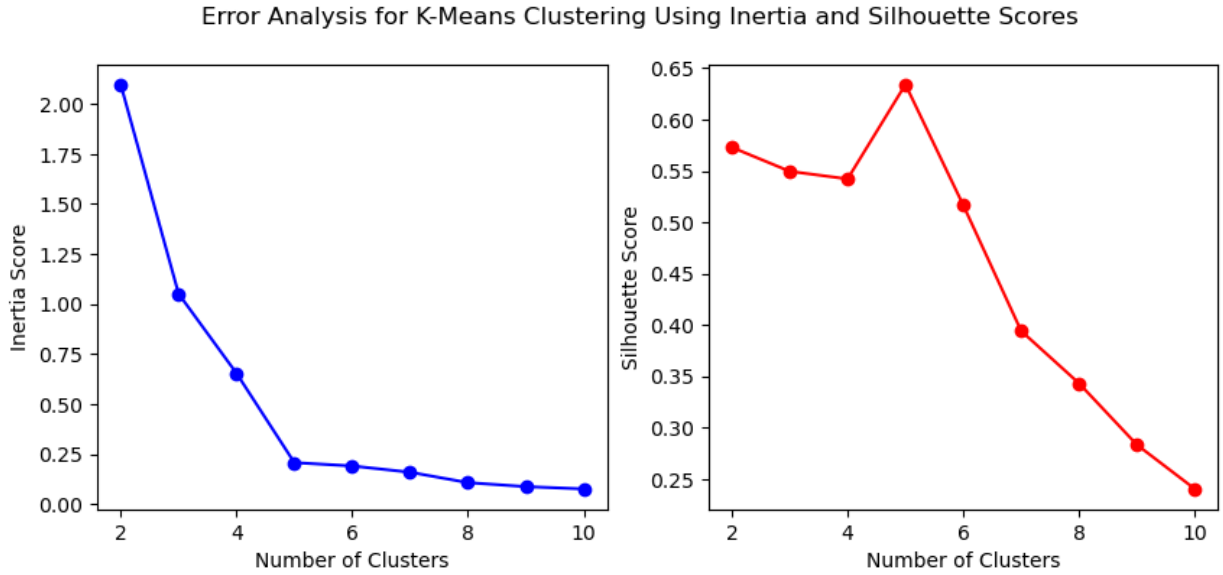
Graph D.3 Caption: Results from PCA feature selection. The explained variance ratio is near 75% when only 2 dimensions of the original genre vector are used. Although this is low, less dimensions means the resulting algorithm will perform better. The resulting 2 dimensions, or principal components, are given as axes in Graph D.5

Graph D.3



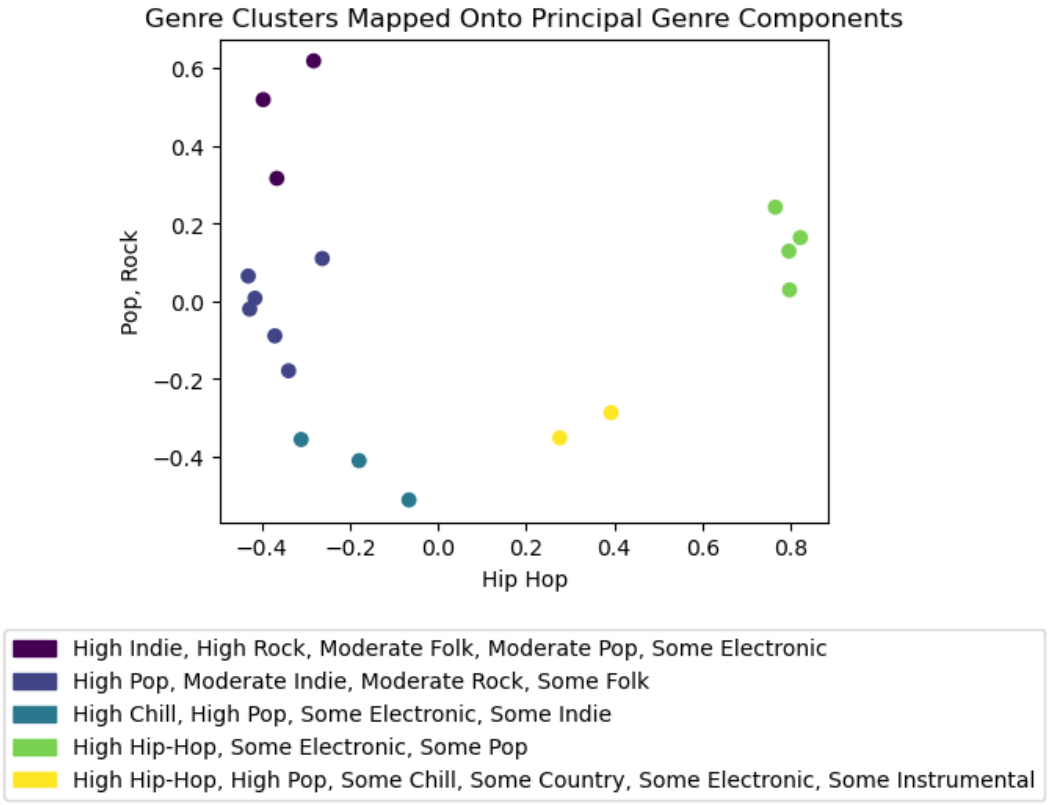
Graph D.3 Caption: *Results of k means clustering algorithm for various k with resulting clusters highlighted on the music network. Note that the outlier node 1 is not included in cluster count. k=5 clusters seem to map well onto the music graph, indicating that this algorithm is accurate.*

Graph D.4



Graph D.3 Caption: *Error results of k means clustering algorithm for various. A low inertia score and high silhouette score is desirable. k=5 clusters achieve the highest silhouette score while also having a low inertia score, and high scores have minimal improvements in inertia score.*

Graph D.5



Graph D.5 Caption: *Genre clusters are graphed based on where they lie in reduced genre vector space (measured by the principal components, roughly corresponding to "Hip Hop" and "Pop-Rock"). Note that the outlier node 1 (corresponding to the "Hip-Hop and Latin" genre) is not included in this graph.*

Table D.1

Percentage of Participants That Ranked Source as Primary Or Secondary Source Of Music For Each Genre Cluster

Genre Cluster	People In Social Network	People Outside Social Network	Algorithms	Other
High Chill, High Pop, Some Electronic, Some Indie	33.28	16.67	33.33	16.67
High Hip-Hop, Some Electronic, Some Pop	28.57	14.29	28.57	28.57
High Pop, Moderate Indie, Moderate Rock, Some Folk	25.00	25.00	41.67	8.33
High Hip-Hop, High Pop, Some Chill, Some Country, Some Electronic, Some Instrumental	20.00	40.00	40.00	0.00
High Indie, High Rock, Moderate Folk, Moderate Pop, Some Electronic	14.29	14.29	42.86	28.57
High Hip-Hop, High Latin	0.00	50.00	50.00	0.00

Table D.2

Probability of Traveling To Each Genre Cluster, Assuming Social Network Acts As Markov Chain

Genre Cluster	Probability
High Indie, High Rock, Moderate Folk, Moderate Pop, Some Electronic	0.32
High Pop, Moderate Indie, Moderate Rock, Some Folk	0.23
High Chill, High Pop, Some Electronic, Some Indie	0.17
High Hip-Hop, High Latin	0.14
High Hip-Hop, Some Electronic, Some Pop	0.08
High Hip-Hop, High Pop, Some Chill, Some Country, Some Electronic, Some Instrumental	0.05

Table D.3

Artist to Genre Category Mapping Used In Determining Genre Profile

'Bad Bunny': ['hip_hop', 'latin'],
'Beabadoobee': ['rock', 'pop', 'indie'],
'Benson Boone': ['pop'],
'Billie Eilish': ['pop'],
'Billy Joel': ['rock', 'chill', 'folk'],
'Bruno Mars': ['pop'],
'Burna Boy': ['hip_hop', 'pop'],
'C418': ['instrumental', 'electronic'],
'Chappell Roan': ['pop', 'indie'],
'Charli Xcx': ['pop', 'electronic'],
'Chonny Jash': ['electronic', 'indie'],
'Clairo': ['pop', 'indie'],
'Coldplay': ['rock', 'pop'],
'Dhruv': ['pop'],
'Don Toliver': ['hip_hop', 'pop'],
'Drake': ['hip_hop', 'pop'],
'Eagles': ['rock', 'chill', 'folk'],
'Est Gee': ['hip_hop'],
'Fiji Blue': ['pop', 'chill', 'electronic'],
'Flower Face': ['folk'],
'Frank Ocean': ['pop', 'chill'],
'Future': ['hip_hop'],
'Gracie Abrams': ['rock', 'pop', 'indie'],
'Graham': ['hip_hop', 'pop', 'chill'],
'Grentperez': ['chill'],
'Holly Humberstone': ['rock', 'pop', 'indie'],
'Hozier': ['rock', 'indie'],
'Joep Beving': ['instrumental'],
'Kanye West': ['hip_hop'],
'Kendrick Lamar': ['hip_hop'],
'Keshi': ['chill'],
'Lana Del Rey': ['pop'],
'Laufey': ['chill'],
'Lizzy Mcalpine': ['rock', 'pop', 'folk', 'indie'],
'Los Tigre'S Del Norte': ['latin'],
'Los Tucanes De Tijuana': ['latin'],
'Mac Miller': ['hip_hop', 'pop'],
'Masayoshi Soken': ['instrumental'],
'Metro Boomin': ['hip_hop'],
'Mother Mother': ['rock', 'indie'],
'Noah Kahan': ['indie'],
'Olivia Rodrigo': ['pop'],
'Paris Paloma': ['pop'],
'Penelope Scott': ['rock', 'pop', 'indie'],
'Phoebe Bridgers': ['pop', 'folk', 'indie'],
'Pi'Erre Bourne': ['hip_hop'],
'Radiohead': ['rock', 'indie'],
'Sabrina Carpenter': ['pop'],
'Sza': ['hip_hop', 'pop', 'chill'],
'Tame Impala': ['rock', 'electronic', 'indie'],
'Taylor Swift': ['pop'],
'The 1975': ['rock', 'pop', 'indie'],
'The Beatles': ['rock'],
'The Marias': ['pop', 'folk', 'indie'],
'The Paper Kites': ['pop', 'folk', 'indie'],
'The Thing': ['rock', 'indie'],
'Toby Fox (Undertale)': ['instrumental', 'pop', 'electronic'],
'Travis Scott': ['hip_hop', 'electronic'],
'Tyler, The Creator': ['hip_hop'],
'Vicente Fernandez': ['latin'],
'Westside Gunn': ['hip_hop'],
'Will Wood': ['indie'],
'Xxxtentacion': ['hip_hop', 'latin'],
'Zach Bryan': ['pop', 'country']}