Problem Chosen	2021	Team Control Number
D	MCM/ICM	2112040
	Summary Sheet	

Summary

Music is an art form that embodies aesthetic values and reveals social and cultural developments over time. Aiming to explore and quantify music evolution especially in the aspect of influences, we develop mathematical models and utilize data visualization to analyze the progress and trends on the world stage of music.

Our paper conducts a thorough and extensive approach to measure artists from the 1920s to 2010s' musical influence and to compare their composition similarity based on various audio features data. To derive the influence distribution, we calculate the artist influence index respectively based on graph structures and characteristic similarity. For instance, we gradually specify and improve our results from degree/eigenvector centrality to scaled time-smooth function and betweenness. Precisely, we narrow down the insight from artists to their specific songs and use them as representatives of their significance in the overall network.

In further, we explore the transformation of genres' characteristics over decades and calculate the inter- & intra-genre similarities of songs. We are also engaged to establish a direct relationship between genre similarity and artist influence. In addition, we project our measurement results into reality and rationalize them based on social, cultural, and technological changes.

Evolution of Modern Music: A Quantitative Approach

Team # 2112040

February 2021

Keywords: Cosine Distance, Regression Analysis, Feature Vector Space, Degree Centrality, Eigenvector Centrality, Betweenness.

Contents

1 Introduction

1.1 Problem Background

Music has always been a cultural essence that not only unfolds characteristic artistic creations but also witnesses social and technological advancements in its developing compositions. In the last 100 years, music from different genres went through rises and declines during their formation, innovation, and reinvention, and finally formed the current condition of world music. We aim to explore and quantify music evolution by the given data sets, seeking similarities among different genres and analyzing revolutions of music styles.

In the process of music formation and reinvention, artists influence their followers' composing styles, and these followers later become influencers for their posterity, so that it forms a broad network which reveals the influences between artists and shapes the similarities between genres. The remarkable characteristics in influencers' compositions affect artists in the future generation and probably construct the style of the whole genre. In fact, other than these influencing effects, there are social, cultural, and technological changes that signify major leaps and long-term trends in music evolution. To delve into these questions, we develop a model to analyze the revolution and influence in music over time.

1.2 Basic Assumptions

1. All our measurements and calculations are based on the data provided, which might not be factually correct and accurate.

2. The artist's musical influence is presented only through its compositions and causal relationship with his followers.

Explanation: The data types and amounts of the provided information are insufficient to make further detailed measurements.

3. The artist's musical influence is evenly distributed to all his compositions.

Explanation: Simply the calculation process. Incapable to simulate their complete music careers and analyze the independent role/significance of each song.

4. Artists from different decades/genres have the same inborn capacity of creation and performances (on average). Only the music features and popularity of previous artists would affect their career prospects.

Explanation: Standardize people's capacity. Eliminate all disturbing and indescribable elements. Mainly focus on the influence relationship between artists.

1.3 General Structure of Models

This mind map shows the brief structure of our model, showing our flow of thoughts from the given raw data to each stage of analysis and observation towards our conclusion. In the mind map, blue rectangles represent states, purple ellipses represent concepts we define in this model, and orange parallelograms represent the outputs we get.

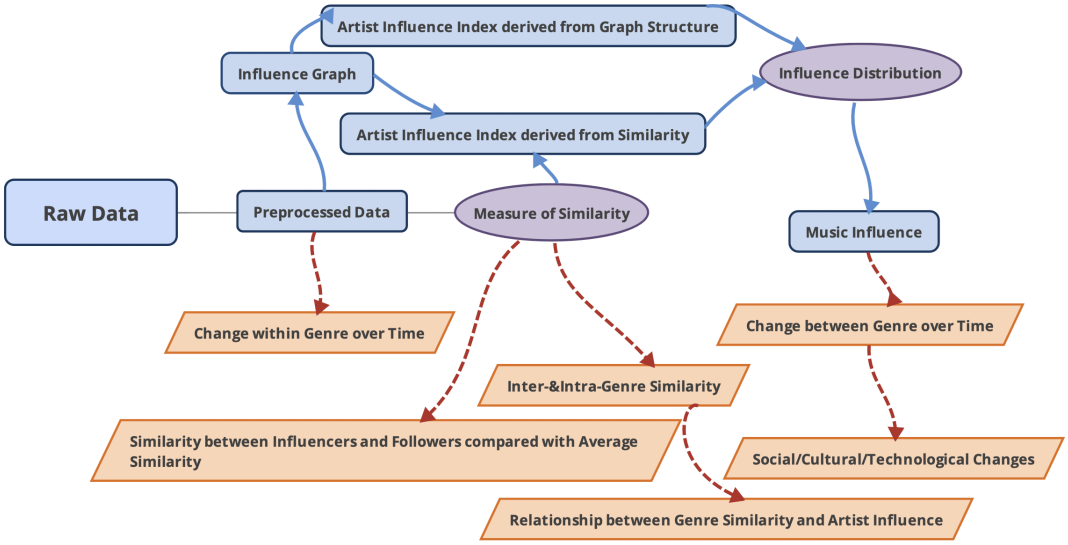

Figure 1: General Structure in Mind Map

2 Artist and Music Influence Index derived from Similarity

2.1 Measuring Music and Artist Similarity

In **data_by_artist.csv** and **full_music_data.csv**, multiple parameters of a song/artist are provided in order to better describe the song/artist. Among them, ten quantitative data (*danceability, energy, valence, tempo, loudness, acousticness, instrumentalness, liveness, speechiness, duration_ms*) in both files are selected to construct the feature vector x for each song/artist due to their quantitative and continuous nature. *Popularity* is quantitative and continuous, but it tells us nothing inherent about the song/artist, thus it is not included in the feature vector.

2.1.1 Normalizing Parameters

To prevent features with large numerical values from dominating in our similarity function[1], we normalize *tempo, loudness*, and *duration_ms* by scaling between 0 to 1 so that all parameters have the same range.

2.1.2 Distance Metric

Several distance metrics have been considered to calculate the measure of similarity. Euclidean distance is the most straightforward approach, but it suffers a lot from "the curse of dimensionality". Considering that our feature vector is 10-dimensional, Euclidean distance may not be an appropriate choice. Manhattan distance performs better when the dimension is high, but we aim to derive a similarity score such that a higher score represents more similarity between two songs/artists. It is quite difficult to determine a proper method to reverse the Manhattan distance. Therefore, cosine similarity seems to be the best approach in this scenario: like the Manhattan distance, the performance of cosine distance also suffers less from high dimentionality. Besides, all of our

parameter now ranges from zero to one after normalization, which means that the cosine distance between any two feature vectors will always falls in $[0, 1]$. Thus, given two feature vector $\boldsymbol{x}, \boldsymbol{y}$, the measure of similarity, $\text{sim}(\boldsymbol{x}, \boldsymbol{y})$ will be

$$\text{sim}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}. \tag{1}$$

Our later work will prove that knowing the range of distance that measures the similarity could be quite beneficial.

2.1.3 Similarity Score Formula

We could simply reverse the cosine distance by subtracting it from 1 and use it as the similarity score. Nevertheless, after examining the distribution of the similarity score, we find that the distribution is highly skewed. To fix that and enhance quantitative discrimination, we take the square of $1 - sim(\boldsymbol{x}, \boldsymbol{y})$ and obtain more evenly-distributed similarity scores. Thus, we define

$$s(\boldsymbol{x}, \boldsymbol{y}) = (1 - sim(\boldsymbol{x}, \boldsymbol{y}))^2. \tag{2}$$

as the similarity score. Still, $s \in [0, 1]$. Larger s indicates higher similarity between artists or songs. Later in section 4, the underlying mathematics calculation of multiple findings involves averaging and comparing the similarity of artists or songs within or between genre.

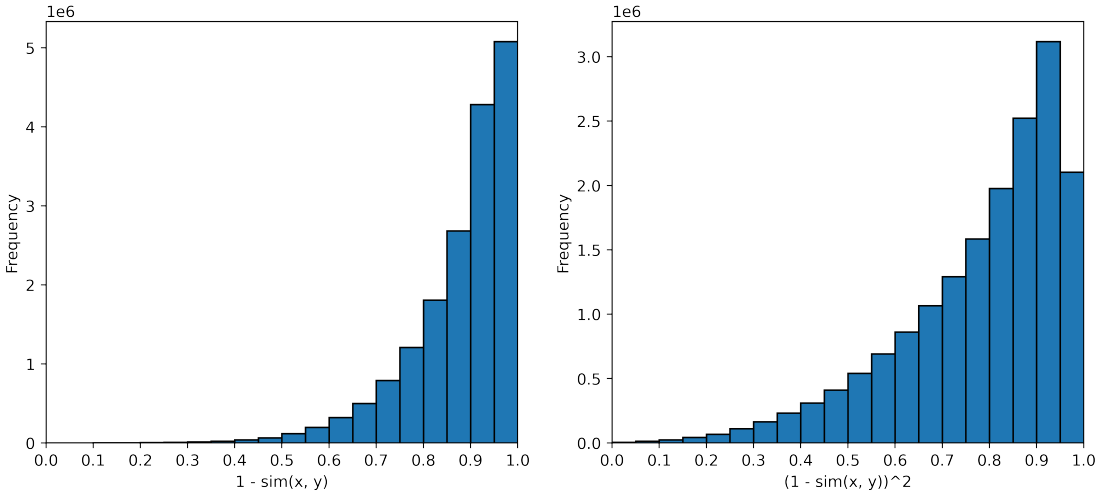

Figure 2: The distributions of similarity scores between artists after applying different formulas

2.2 Measuring Artist Influence with Similarity

To quantify the influence of an artist, we need to take both the number of followers and the similarity between the artist and their followers into consideration. An intuitive approach is to add up the similarity score between an artist and their followers. In this way, the influence of an artist $\boldsymbol{a}$ will be

$$\text{infl}(\boldsymbol{a}) = \sum_{\boldsymbol{f} \in \text{follower}(\boldsymbol{a})} s(\boldsymbol{a}, \boldsymbol{f}). \tag{3}$$

Artist	Influence Index	Starting Decade
The Beatles	567.26	1960
Bob Dylan	353.88	1960
The Rolling Stones	299.44	1960
David Bowie	221.74	1960
Led Zeppelin	208.07	1960
Jimi Hendrix	182.44	1960
The Kinks	179.31	1960
The Beach Boys	170.45	1960
The Velvet Underground	161.62	1960
Hank Williams	160.41	1930

Table 1: Influence Index derived from Similarity, Top 10

2.3 Distributing Influence of Artists to their works

Artists demonstrate their influences through their compositions. Thus, assume that an artist is the only composer of all their compositions, summing up the influence of all songs composed by an artist should give the influence of that artist. Based on the given data, it is not possible to determine the influence of every single piece of music based on its specs and its composer. Therefore, we assume that every piece of music carries equal influence from its composer. If $\boldsymbol{m}$ is a piece of music, we have

$$\text{infl}(\boldsymbol{m}) = \frac{\sum_{\boldsymbol{c} \in \text{composer}(\boldsymbol{m})} \frac{\text{infl}(\boldsymbol{c})}{|\text{music}(\boldsymbol{c})|}}{|\text{composer}(\boldsymbol{m})|}. \tag{4}$$

By our definition, suppose A is the set of all artists and M is the set of all music

$$\sum_{a \in A} \text{infl}(\boldsymbol{a}) = \sum_{m \in M} \text{infl}(\boldsymbol{m}) |\text{composer}(\boldsymbol{m})| \tag{5}$$

should hold.

Artist	Average Influence Per Composition	Starting Decade
Louis Jordan	23.07	1930
Captain Beefheart	13.52	1960
Slim Harpo	10.44	1950
Hazel Dickens	9.65	1950
Cabaret Voltaire	9.43	1970
The 13th Floor Elevators	9.06	1960
Del McCoury	9.02	1960
Steve Reich	8.97	1960
Beth Orton	8.68	1990
Spike Jones	8.12	1930

Table 2: Influence Index divided by Total Number of Compositions, Top 10

After distributing influence of artists to their works, we now get influence index for each piece of music, which could then be used to evaluate the change across genres over time in Section 4.

3 Artist and Music Influence Index derived from Graph Structure

3.1 Music Influence Measurement

3.1.1 Degree Centrality

Intuitively speaking, the musician's influence on his followers and future generations could be directly measured by its out-degree, as its follower numbers. As the parent node of its spanned network, its influence continuously pass on to the following children connected to the tree, despite the magnitude gradually decreases (in exponential speed). Such measurement is understandable and easily implemented. The result is highly consistent with our common knowledge and *100 Greatest Artists Ranking* published by Rolling Stone[2]. As Figure 3 illustrates, the three-layers subnetwork originated from The Beatles mostly covers the total graph. Its followers' genres are widely diversified, which also indicates its universally inspirational power. However, degree centrality does not capture any indirect form of influence. Rather, it equally weights all followers' significance (as in the case of sample chains) and disregards their respective inherent music influence. The recursive and dynamic process is truncated to a one-time calculation. The time factor is also ignored.

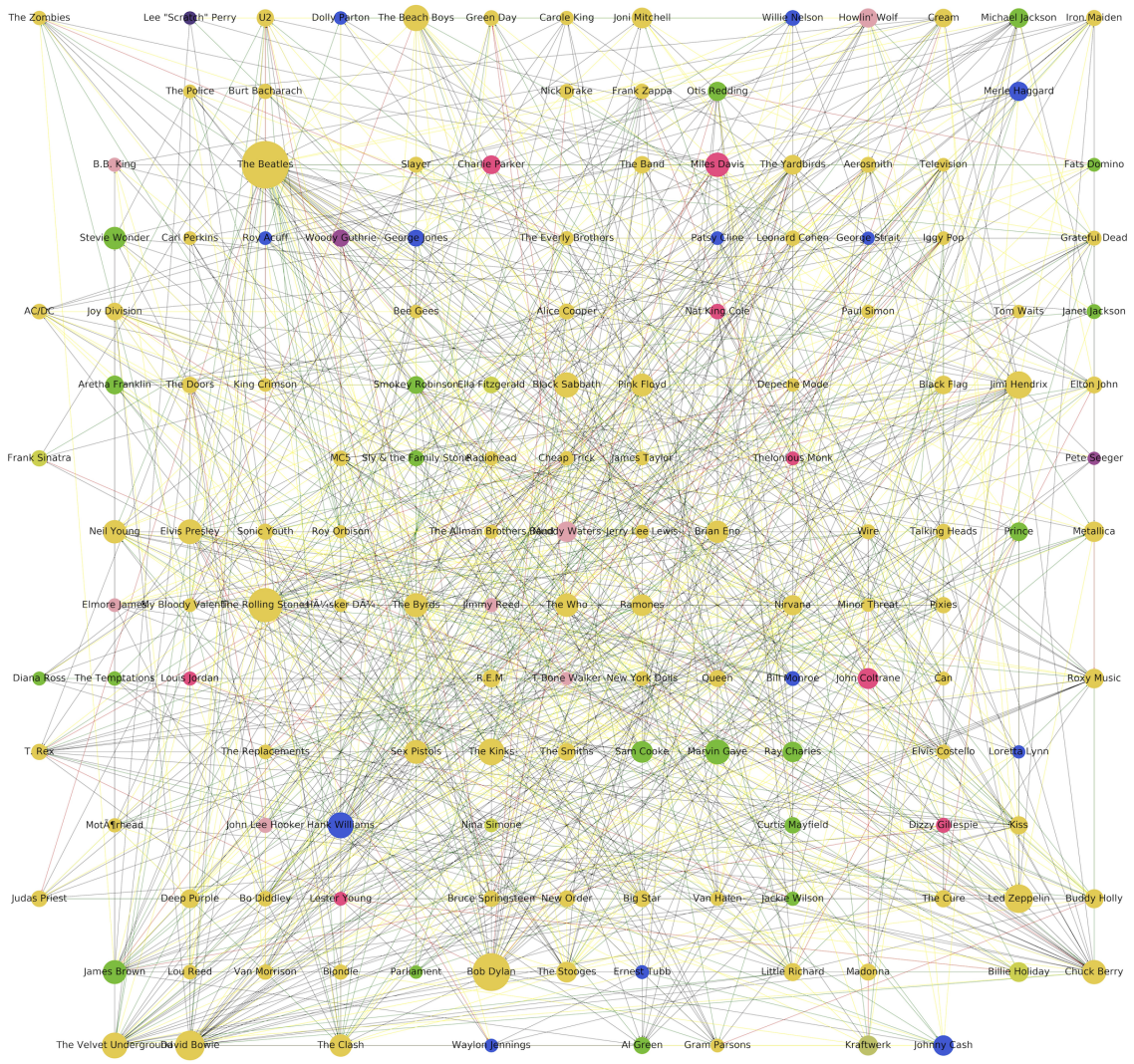

Figure 3: Reduced Network sorted by follower count. Origin at The Beatles. Node size reflects its out-degree. Node color presents its categorized genre. The Beatles and its 1st-generation followers are connected in red edges. 1st-generation and 2nd-generation followers are in yellow. 2nd-generation and 3nd-generation followers are in green. Else influential relationships are in black.

3.1.2 Eigenvector Centrality: PageRank Algorithm

A node is important if it is linked to other important nodes. For nodes with the same number of connections, nodes with higher adjacent node scores will have higher scores than nodes with lower adjacent node scores. For any given graph $G := (V, E)$, suppose $A = (a_{v,t})$ be the adjacency matrix. The relative centrality, R, the score of vertex v, could therefore be defined as

$$R_v = \frac{1}{\lambda} \sum_{k \in M(v)} R_k = \frac{1}{\lambda} \sum_{k \in G} a_{v,k} R_k. \tag{6}$$

$M(v)$ is a set of neighbors of node v and λ is a constant. After arrangement, we deduce the eigenvector equation

$$\mathbf{Ax} = \lambda \mathbf{x}. \tag{7}$$

Notice that all items in the eigenvector are non-negative and the desired centrality can only be measured when the eigenvalue is the largest.

Based on such a foundation, the PageRank algorithm is developed to roughly estimate how important the specific node is by counting the number and quality of links (edges). Hence, PageRank could better describe indirect influence between nodes, calculate an overall influence rank among each node, and demonstrate the influence of one node to another. A modified assumption is that more important nodes are highly probable to extend more direct edges. We simulate the music significance flow as the random surfer, and the overall structure of the network is a Markov chain. Nodes (artists) with no outbound links (followers) are assumed to link out to all other nodes in the collection. Their PageRank score is evenly divided by others. To solve that residual probability, we define the damping factor d to describe the probability that a specific artist's influence continues to transit at any step. Also, in reality, the parameter introduces some "unstable" and "weak" relationship that the influencer barely makes slight short-term impacts on the followers and does not change their artistic style distinctly. A large d would exaggerate the parent nodes' influence. Early-year artists are unconditionally preferred and the whole evaluation network becomes biased and skewed. Conversely, a low d would cause purely random walks. After parameter adjustments, we set d as 0.8. Hence, the equation could be written as

$$PR(a_i) = \frac{1-d}{N} + d \sum_{a_j \in A(a_i)} \frac{PR(a_j)}{L(a_j)}. \tag{8}$$

where $a_1, ..., a_N$ are all artists in **influence_data.csv**, $L(a_j)$ is the number of outbound links of node a_j.

The PageRank values are the entries of the dominant right eigenvector of the modified adjacency matrix rescaled, which could be written as

$$\mathbf{R} = \begin{bmatrix} \mathrm{PR}(p_1) \\ \mathrm{PR}(p_2) \\ \vdots \\ \mathrm{PR}(p_N) \end{bmatrix}. \tag{9}$$

Also, we define the adjacency function $l(p_i, p_j)$ as the ratio between the number of links outbound from artist j to artist i to the total number of outbound links of artist j. To solve the previously mentioned leaf node, then the adjacency function equals 0. Then implement normalization such that

$$\sum_{i=1}^{N} l(p_i, p_j) = 1 \quad \forall j. \tag{10}$$

Also, $\mathbf{R}$ is the solution of the equation

$$\mathbf{R} = \begin{bmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{bmatrix} + d \begin{bmatrix} \ell(p_1, p_1) & \ell(p_1, p_2) & \cdots & \ell(p_1, p_N) \\ \ell(p_2, p_1) & \ddots & & \vdots \\ \vdots & & \ell(p_i, p_j) & \\ \ell(p_N, p_1) & \cdots & & \ell(p_N, p_N) \end{bmatrix} \mathbf{R}. \tag{11}$$

The values of PageRank eigenvector could be accurately approximated after several iterations, despite of the large eigengap of the modified adjacency matrix [3].

Artist	Influence Index	Starting Decade
Billie Holiday	0.010598997	1930
Louis Jordan	0.010518932	1930
Lester Young	0.009437718	1930
The Beatles	0.007428771	1960
T-Bone Walker	0.007303294	1930
The Mills Brothers	0.005775355	1930
Charlie Christian	0.004730226	1930
Bob Dylan	0.004662024	1960
Sister Rosetta Tharpe	0.004533481	1930
Nat King Cole	0.004319766	1930

Table 3: PageRank Score: Influence Index, Top 10

However, some inherent problems of the PageRank algorithm consistently exist, which inspires us to pursue and combine it with further improvements and other measurements. **influence_data.csv** shows that young and talented artists normally do not have sufficient followers (like Taylor Swift) since they are still in the early stage of their career. Despite we moderately reduce d value to push the evaluation more inclined to current generations, their influence measurement is still underestimated and inaccurate. Table 3 clearly illustrates that early-period artists are highly praised and occupy the influence flow. Thus, we need the smooth function to alleviate the misleading time effect.

Also, PageRank automatically down-weights influence created by a destination node that samples more than once and from artists with lengthy careers[4]. Katz influence might be a better alternative, which equally weights each instance of sampling. As such, the influence matrix $\mathbf{I}_K$ is defined by

$$\begin{aligned} \mathbf{I}_K &= (\mathbf{I} - \alpha\mathbf{A})^{-1} - \mathbf{I} \\ &= \alpha\mathbf{A} + \alpha^2\mathbf{A}^2 + \ldots + \alpha^k\mathbf{A}^k + \ldots \end{aligned} \tag{12}$$

α is the decay factor that describes the indirect influence to propagating through the network. Also, the matrix equals to the weighted sum of the powers of the adjacency matrix. The iterative calculation is more compatible to our large network considering how intense the matrix inversion is. By doing summation respectively to the rows/columns of influence matrix, we could get the ranking of most influential/influenced nodes.

3.1.3 Time Adjustment

full_music_data.csv file provides a criterion to describe the popularity of the track based on the total number of plays and how recent those plays are. This information highly emphasizes the modern generation of artists, whose compositions are widely spread through digital media and recognized by the general public (along with the raising of average educational level). The relationship between average popularity (by years) and time is greatly intense.

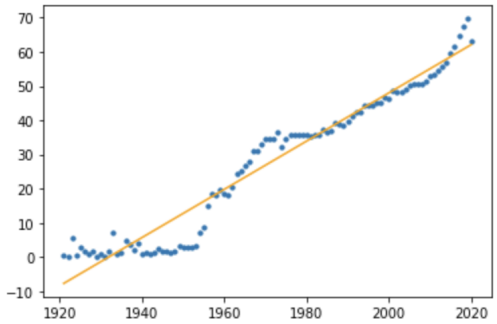

Figure 4: Time vs. Average Popularity Rating of that specific year

Through the Least Squares method, we obtain the coefficient of determination (r^2) as 0.947 and the regression function is $y = 0.7045 * x - 1360.9905$. We are mostly concerned about the artists who have exceptionally distinguishable performances comparing to their contemporary peers, i.e. to maximize the difference between the actual average popularity (by the artist) and simulated result (by year). To better present the artist's career, we do not use the active start year provided in **influence_data.csv**. Rather, we search through **full_music_data.csv** and cluster the songs by artists[1]. Then we analyze the published years for these songs and create a 95% confidence interval to highlight artists' most dynamic and productive periods. Also, the outlier effect is reduced, especially for artists with lengthy and unevenly distributed careers (i.e. large range and large standard deviation, like Wang Chung has a 30-year-long profession). It would originally exert drifting force in one direction. Also, we specify the simulated popularity result by calculating the propagation of error on the CI regression parameter and working year.

$$\frac{\Delta a}{a} = \frac{\Delta b}{b} + \frac{\Delta c}{c}. \tag{13}$$

We normalize the value of difference and denote it as Time Wrapping coefficient $\mu_i \ \forall i$ for future evaluation. This truncated result (Table 4) vaguely illustrates a fairer evaluation based on a self-adapted and dynamic standard.

Artist	Diff Value	Starting Decade
El Guincho	19.60500	2000
The Beatles	19.58620	1960
Franco Battiato	17.97449	1980
Miles Davis	17.72633	1940
The Byrds	17.68923	1960
The Kinks	17.65812	1970
The Rolling Stones	17.61626	1960
Jimi Hendrix	17.60012	1970
Elvis Presley	17.57721	1950
Led Zeppelin	17.57612	1950

Table 4: Beyond Contemporary Artists Popularity, Top 10

[1]Songs with multiple composers will be counted for several times

3.1.4 Betweenness

Generally, the more connected a node is, the more likely it is to receive new links. Preferential attachment is suggested in [5] to impose weights on edges. We use the scoring functions as features of a learning algorithm. Suppose $\Gamma(u)$ is the set of neighbors of node u. The PA coefficient is defined as $|\Gamma(u)||\Gamma(v)|$, which describes the significance of an influential relationship to the whole network evolution. The graph exhibits strong PA property, as shown by the skewed in- and out-degree distributions of the graph.

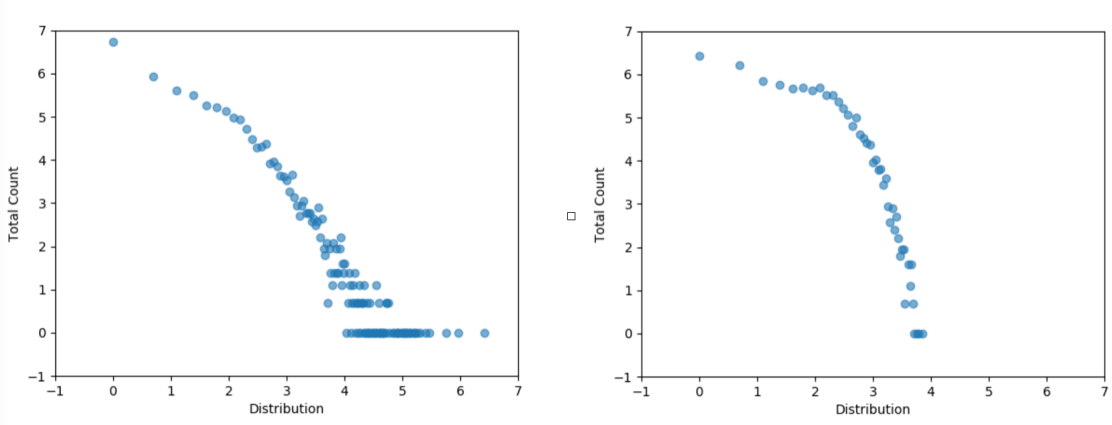

Figure 5: In- and Out-degree Distributions of the Influence Graph, plotted on a log-log scale

We take its normalized reciprocal and set the corresponding weight. Based on the weighted direct graph, we could conduct a centrality measure through betweenness to determine the node's influence on the overall network structure. It determines how often one must pass through a given node going from an origin to a destination [6]. Specifically, it describes the possibility that the composer's features are inherited and spread. Nodes located in bottom-layers and central hub positions would be preferred, which is further smooth the scale of time influence.
Betweenness of a node x is defined as

$$B(x) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}. \tag{14}$$

where V is the set of nodes. $\sigma(s,t)$ is the shortest (s,t) paths (through Dijkstra algorithm), and $\sigma(s,t|v)$ is the number of those paths passing through some node v other than s,t. If $s = t$, $\sigma(s,t) = 1$; if $v \in s,t$, $\sigma(s,t|v) = 0$.

3.1.5 Final Evaluation

We combine all factors addressed in Section 3.1.2-3.1.4 through scalar multiplication, optimal scaling, and logarithm smoothing [4]. We then generate the final evaluation ranking.

Artist	Ult Music Influence Index
The Beatles	1.00000
Bob Dylan	0.48855
The Rolling Stones	0.37669
Jimi Hendrix	0.28023
The Velvet Underground	0.26512
Miles Davis	0.25425
The Beach Boys	0.25146
Led Zeppelin	0.24848
David Bowie	0.24671
Hank Williams	0.24520

Table 5: Final Evaluation of Music Influence (normalized), Top 10

We then evenly distribute the artist's influence to each composition using the methodology mentioned in Section 2.3

4 Results

4.1 Inter- and Intra-Genre Similarity

Applying Equation 2, we calculate the average similarity of artists between or within genres.

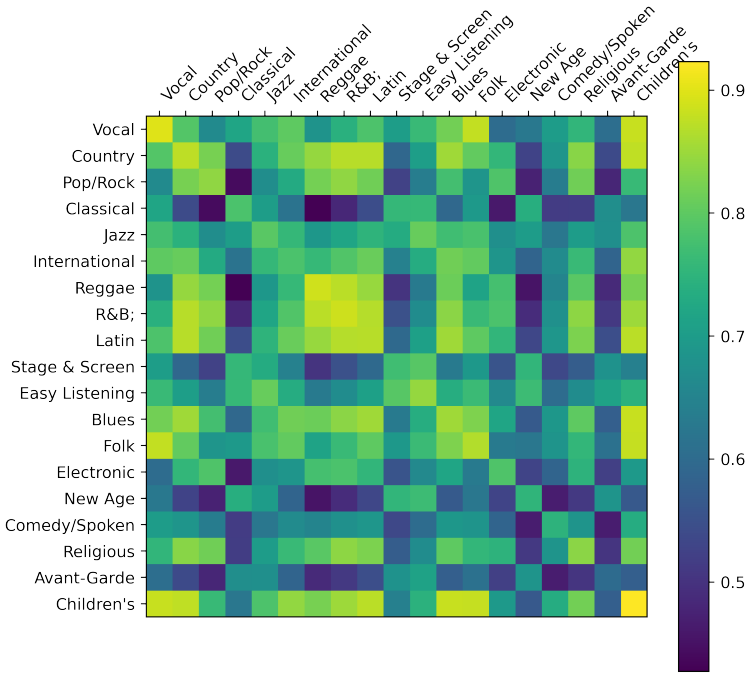

Figure 6: Average Artist Similarity Between or Within Genres

From Figure 6, one could conclude that in general, artists within one genre is more similar than

music from different genres. While the average similarity score between some genres might be higher than the average similarity score within another genre (e.g. average similarity score between folk musicians and vocalists is higher than the average similarity score among jazz players), artists from a genre is always more similar to each other compared with artists from another genre. For example, the average similarity score of classical music artist is higher than the average similarity score between classical music artists and artists from any other genre.

4.2 Relationship between Genre Similarity and Artist Influence

In section 4.1, $\binom{19}{2} = 171$ intra- or inter-genre average similarity score are calculated. On the other hand, from **influence_data.csv**, the number of interactions within or between genres could be counted. If the genre of the influencer is g_1 and the genre of the follower is g_2, the number of interactions between genre g_1 and g_2 is increased by 1. By fitting a simple linear regression model between the natural log of number of interactions and average similarity score, we get $r^2 = 0.104$, which is an acceptable value in social science field. It indicates that there might be positive relationship between number of interactions and similarity score.

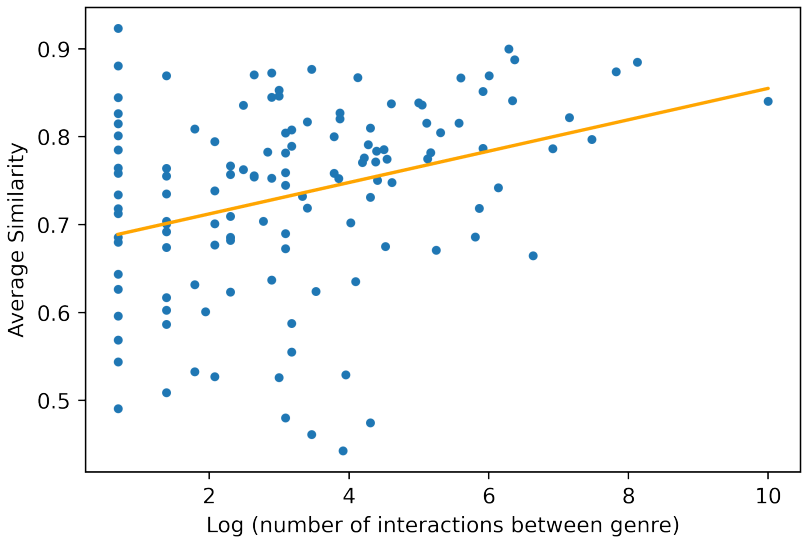

Figure 7: Relationship Between Number of Interactions and Similarity, $r^2 = 0.104$

4.3 Similarity between Influencers and Followers compared with Average Similarity

To determine whether influencer actually influence their followers, evaluating the similarity between every pair of influencer and follower in **influence_data.csv** is quite useful considering that if a follower is influenced, the characteristic of their music will become similar to their influencers. After computing ≥ 40000 similarity scores between influencers and followers and $\geq 10^8$ similarity scores between any two artists, we found that the average similarity score between influencer and followers are greater than overall average similarity score. Conducting a permutation test on two sets of similarity scores gives an extremely small p-value ($\leq 10^{-5}$), indicating that it is extremely likely

that the underlying distribution of similarity score between influencers and followers is different from the underlying distribution of similarity score between two random artists.

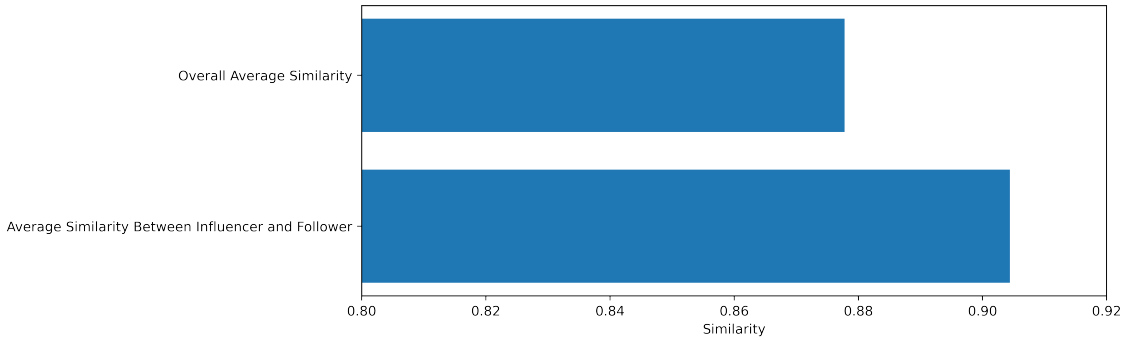

Figure 8: Average Similarity Score between Influencers and Followers Compared with Overall Average Similarity Score

4.4 Change within Genre Over Time

To better describe the inherent change of a genre, focusing on the provided characteristics of music is crucial. Therefore, we directly analyze seven key features provided in **full_music_data.csv** to observe the change within every genre over time.

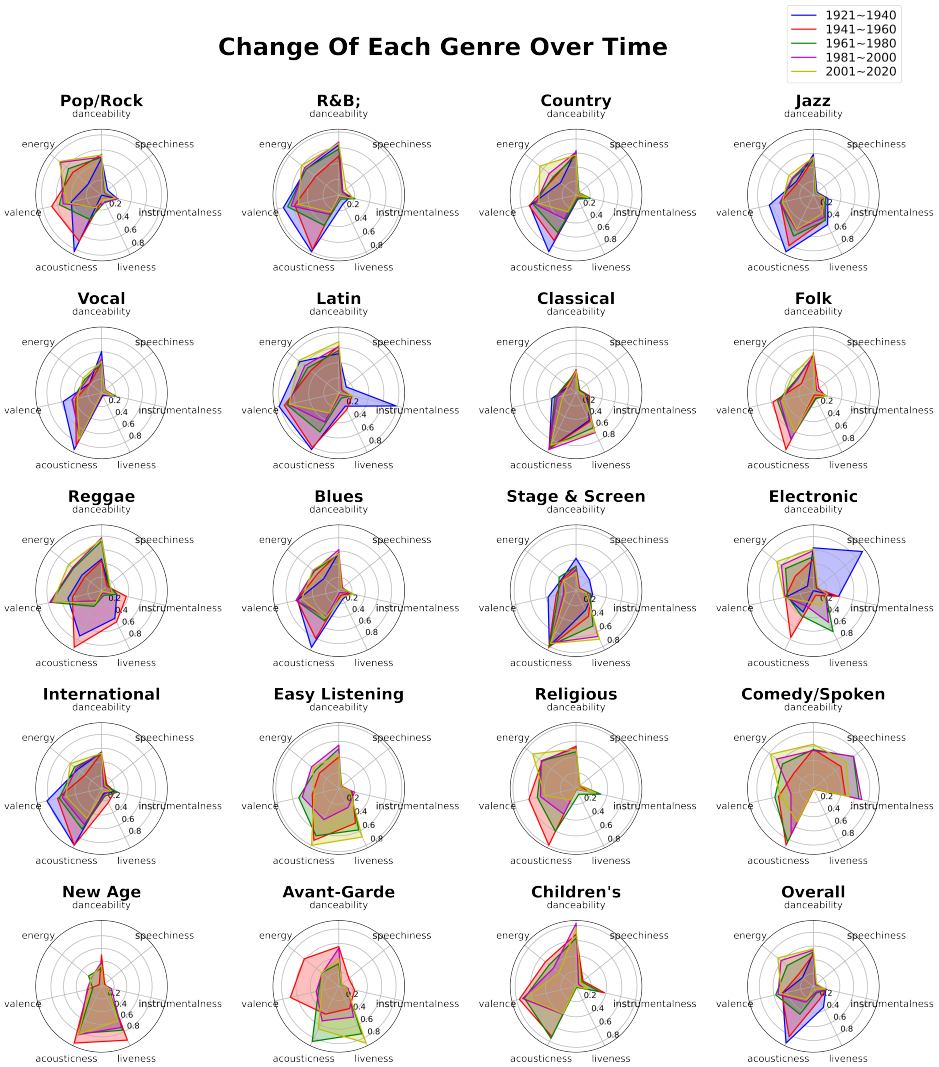

Figure 9: Change Of Each Genre Over Time

Based on the Figure 9, we find out that in 1921-1940, music genres are not so diverse: genres of Religious, New Age, Folk, Comedy/Spoken, Easy Listening, Avant-Garde, and Children's have no data before the 1940s. Starting from the 1940s, many influential artists contribute to the formation and exploration of different genres and finally promote the evolution of music styles till the present.

The *acousticness* characteristic of music decreases from the 1920s to the 2010s, meaning that technology enhancements and electrical amplification are deeply blending into music production. With the technological advancements in the last 90 years, artists add more electrical elements in their music instead of pure vocal and instrumental production in the past, and thus reduces the *acousticness* indicator in our data. Another interesting point is that the *energy* characteristic is getting more and more prominent over time. Compared with music in the early and middle 20th

century, music in the recent 20 years have higher *energy* level than ever before (except four particular genres: Classical, Avant-Garde, Stage & Screen, and Children's), generally sharing more specific characteristics of fast, loud, and intense. However, most genres show lower *valence* than ever before. That is, through the nearly 100 years of music evolution and with the background of technological advancements, the overall musical positiveness is gradually decreasing. Music before the 1960s have relatively higher *valence*, meaning that they sound more positive and cheerful. However, we find that during 2001-2020, the valence among all genres has descended to a lower value meaning the sense of sadness and depression.

For the biggest genre, Pop/Rock, in our data, we find out that stepping into the 1940s, the index of *energy* and *valence* rapidly increases, developing into more energetic and euphoric. The decade of the 1950s marks the formation of Rock music, initiated by one of the most popular artists Elvis Presley who produces more than 990 songs and pursues a high level of musical positiveness, so that our data from 1941 to 1960 shows a major leap in *valence* and *energy*. The 1960s is the most significant period of Pop/Rock music, in which *acousticness* greatly decreases and *energy* continues increasing. Influential artists such as The Beatles, Bob Dylan, and The Rolling Stones who own the three greatest numbers of followers and also the three highest final ranking of musical influence in our data are all exceptionally active in this decade. With these three artists' influences that signify a great revolution of Pop/Rock, the music characteristics in Pop/Rock genre are largely determined. The subsequent evolution is mainly because of the technological advances that stimulate the increasing electrical amplification in music production in the 1980s. Finally, there are no remarkable changes when comparing 1981-2000 with 2001-2020, since Pop/Rock music has a long history and this style gets relatively mature and steady at the end of the 20th century.

In R&B genre, a significant leap can be found from the change between 1941-1960 and 1961-1980. Influential R&B artists who have more than 100 followers such as Marvin Gaye, James Brown, Sam Cooke, and Ray Charles are all contributors in the 1960s. They enhance the *energy* characteristic, attaching more importance to their music perceptual feature of intensity and liveliness. On the o before 1960, adding electrical sounds to their previous acoustic instruments in response to the technological improvements and the rise of electronic music in the 1950s to 1960s. Evolving into the 1980s, other characteristics except *acousticness* only change a little bit, but *acousticness* continues to reduce, with further incorporation with electrical amplification strongly influenced by the flow of technological advances.

One last interesting observation is that evolved into the 2000s, the shapes of big genres showed in our graph seem fairly similar to each other. The three biggest genres Pop/Rock, R&B, and Country, together with some small genres like Latin and International, share characteristics of extremely low *speechiness* and *liveiness*, with high energy, and nearly the same level of *valence* and *acousticness*. These features are consistent with our 2001-2020's overall graph on the bottom right corner. With few major revolutions in the last 20 years, the 2000s actually see the convergence of different genres that shows more comparable characteristics among different genres in musical evolution[7].

4.5 Change in Influence across all Genre

Using two different approaches in Section 2 and 3, we generate two set of influence indexes for all artists. One set of indexes derives mainly from the graph structure, while another set of indexes derives mainly from the similarity measurements. After distributing the artist influence to music using the approach mentioned in Section 2.3 and aggregate the music influence according to the genre and decade, we are able to use stack plot to present the change in influence across all genre over time.

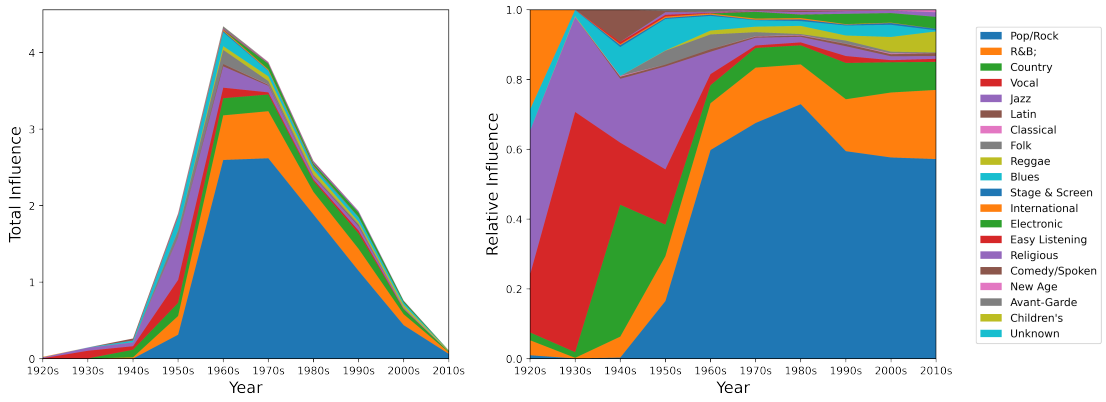

Figure 10: Change In Influence of Genres Over Time, Influence Measured By Total Number of Songs Weighted by Artist Betweenness and Smoothed Influence From Pagerank (Approach in Section 3)

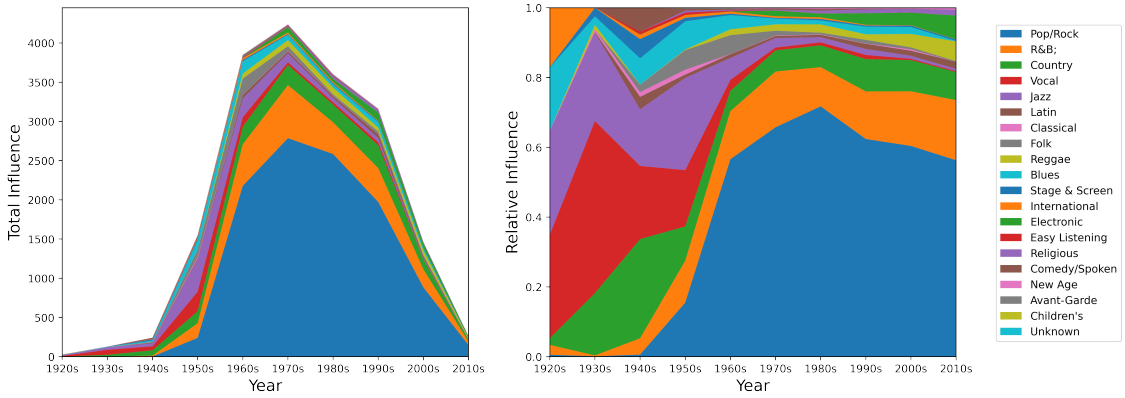

Figure 11: Change In Influence of Genres Over Time Influence Measured By Total Number of Songs Weighted by Similarity Score (Approach in Section 2)

According to Figure 10 and 11, we can visualize the musical evolution from the last 90 years and signify major revolutions by scrutinizing the changes in different genres. The overall trend of the graphs shows that the 1960s is the heyday of music around the world based on the given data, no matter whether we evaluate such prosperity by the total number of songs or by total number of songs weighted by artist betweenness and similarity.

In the 1930s, Vocal is the majority of all genres, because of the most influential pioneer Billie Holiday, who leads music around 1933 to significantly high popularity compared with other years in the decade. In the 1950s, the proportion of Jazz among all the genres markedly enhances, led by the most acclaimed artist Miles Davis in the history of Jazz. Miles Davis ranks Top 20 both in our influential rankings based on the number of followers and in those based on artist betweenness. From the above two decades, we can figure out that a prominent influencer can actually lead his genre to be a standout among other genres in a specific time period.

When it comes to the 1960s, total number of songs and their influence both reach the peak. In this decade, the proportion of Pop/Rock among all music genres soars to nearly 50%. Famous artists such as Bob Dylan, The Beatles, and The Rolling Stones all started their composition in the 1960s. These three artists have more than 1000 songs per person and represent the revolutionary age of music when many artists start music productions and Pop/Rock starts to dominate more than a half on the world stage of music. Being considered as pioneers and predecessors of the new and free stage for Pop/Rock music, Bob Dylan, The Beatles, and The Rolling Stones' songs are listened and learned over time, which again enhances their influence level and thus also elevates the weight of the 1960s in music evolution from our given data.

Another small leap is that the proportion and influence of R&B and Country music rise steadily after the 1970s. Trace to our "follower count" and "final ranking of musical influence" data, Marvin Gaye, James Brown, Stevie Wonder, Merle Haggard, George Jones are all top-ranked R&B or Country artists who are active after the 1970s. Their works signify the leap in R&B and Country music that now become the second and third biggest genres. Although R&B and Country music do not gain the high weighted influence compared with Pop/Rock, those influencers around the 1970s actually promote the style formation and work accumulation in R&B and Country music.

4.6 Social, Cultural or Technological Changes

Comparing our two Figures 10 and 11, we find out that in the 1940s, total number of songs increases than the past, while total influence measured by number of songs weighted by similarity score declines and almost hits the bottom. Artists do increase compositions, but the influence index in this period actually goes down. Such a strange situation reflects the special historical condition in the 1940s when the majority of the world was involved into the World War II. The promotion of art and the dissemination of culture were severely affected in wartime, so that the popularity and propagation of music in the 1940s were not that effective.

Based on our data, the 1960s is the peak of musical evolution. Besides the power of influencers we discussed above, some social and political changes also work together to finally bring about such prosperity of Pop/Rock music. This big revolution in music happens at the background of the Vietnam War that provides artists with broad theme spaces for creation[8]. In addition, the "British Invasion" began in the 1960s when Pop/Rock groups like The Beatles and The Rolling Stones and singers like David Bowie and Led Zeppelin from the United Kingdom became popular in the United States, which enriched Pop/Rock music with diverse cultural contexts[8]. As well as native American singers like Bob Dylan, all these artists active in the 1960s gave impetus to the

consummation of Pop/Rock that plays the dominant part of world music in the context of political events and cultural influence.

The 2000s saw the rapid technological advancements mainly in the Internet. We find out that the influence (measured by total number of songs weighted by artist betweenness) of 2000s and 2010s unexpectedly decline at a rapid rate, which has something to do with the technological changes in the 21st century. The coverage of the Internet makes it easier for people to get access to different new songs, but it disperses the influence level of each song since the convenient and abundant choices offer people approximately comparable possibilities to listen to any new songs. While for those classic songs over time, they are more likely to be known by listeners when considering influential works. Thus, these famous songs have greater possibilities to gain higher influence index, which has already shown throughout our project that our raw data inclines to pioneer artists.

Document to the Integrative Collective Music(ICM) Society

Consultant Team #2112040 - This document describes how we construct our model to better transform the provided raw data of music influence and features into intentional metrics and visualized informations to better illustrate the musical evolution.

We apply a set of rigorous and logical analysis to ensure that our final model is able to convert raw data into simple yet informative statistics, charts and graphs that perfectly respond to your specific requirements and problems. Considering that the several data files given are closely related to each other and share some columns, we decided that a graph structure could store, present and manage key information in the dataset in a most effective manner. The cosine distance metric is selected based on the high-dimentionality of music data provided. Multiple forms of data plots including heatmap, radar plot and stack plot are applied based on the nature of music-related input. All the methodologies are specifically chosen in order to better process the data, which makes our approach highly valuable for quantifying and analyzing the influence of modern music. Another advantage of out model is that increasing number of genres will not require an overhaul to the model. In our model, genre is treated as a categorical data from the beginning to the end. As a result, after expanding the genre type, our model is able to present statistics, plots and graphs with information of more genres immediately.

Nevertheless, even though in fact there are no restrictions mentioned in the previous paragraph, we still realized that there are limitations to our model during the model construction, which results from the lack of more informative dataset. For example, although **"influence_data.csv"** provides us with interactions between artists from different genres, it is still not possible to determine which subset of influencer's work actually influence the follower and the respective influence magnitude of every single composition. Therefore, we have to assume that every composition from the same influencer evenly influences its follower. Besides, the time span of the dataset is relatively short. As a consequence, the shorter influence chain will overestimate the influence of a certain group of artists in the 1920s and the 1930s when implementing some graph algorithms.

Therefore, although the model we provided serves as a satisfactory solution to fulfill all the requirement given the current dataset, if larger and more informative datasets are accessible, we recommend further study on how music in one genre influence artists from different genres, which is an expansion to our current investigation on how artists interact based on their genre. Moreover, researches that investigate the influence of music and artists over centuries are highly recommended.

In conclusion, we hope our model could effectively contribute to the analysis of modern music and help you gain deeper quantitative insights on how music industry evolves during the past 100 years.

Sincerely,
Consultant Team #2112040

References

[1] Mikhail Y. Prostov Maria M. Suarez-Alvarez, Duc-Truong Pham and Yuriy I. Prostov. Statistical approach to normalization of feature vectors and clustering of mixed datasets. *Royal Society*, 2012.

[2] Rolling Stone. 100 greatest artists, 2000.

[3] Taher H. Haveliwala. The second eigenvalue of the google matrix. *Stanford University Technical Report*, 2008.

[4] Nicholas J. Bryan. Musical influence network analysis and rank of sample-based music. *ISMIR*, 2011.

[5] E. Ravasz A.L. Barabasi, H. Jeong. Evolution of the social network of scientific collaborations. *Physica A*, 2002.

[6] Stefano Ferretti. On the complex network structure of musical pieces: Analysis of some use cases from different music genres. 2017.

[7] Digital Hits One. Major events and trends in popular music during the 2000s.

[8] Mark Levy Matthias Mauch, Robert M. MacCallum and Armand M. Leroi. The evolution of popular music: Usa 1960–2010. *Royal Society*, 2015.