Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
D		2112648

Don't Stop the Music, Don't Stop the Evolution: The Music Evolution Lying Under Math

Summary

The interdisciplinary research of music functions as a perspective of looking at the collective human experience. With facilitate access provided by the IT industry , now we can quantitatively investigate deeper into the evolution of music than ever before.

To process the data we find useful, firstly, we establish Network Model, which defines the Attributes of the Network, and capture the music influence by determining the Measures of Music Influence based on *influence_data* data set. More precisely, we adjust the measures according to detected external environment, i.e., the number of artists began their career per annum. Secondly, we propose Similarity Model, composing Selected Properties of Music into Three Levels of Indicators . Then our model integrates Euclidean distance and cosine similarity, calculate correlations of various music characteristics.

As for the results, subnetwork of Crowded House reveals that, most of the artists have weak impact or attention compared with few masters, which is in accordance with Pareto principle. However, they tend to list many influencers, so that nobodies can sometimes affect big names. After that, artists are more similar within genres according to the Similarity Model. Moreover, the likeness and influence between genres very: Avant-Garde performers are most similar, and the Pop/Rock are most influential. We also prove that influencers of a particular artists only accounts for 2.05% of the characteristics of the singer generally, which reclaim the statement that followers are likely to overstate influencers. Unlike the claimed influencers, speechiness is very contagious.

With time many interesting discoveries occur: by comparing the difference, we detect major leaps in 1929s and 1946s, which may due to the end of Great Depression and World War II. Moreover, Miles Davis *et al.* plays an important role in the revolutionary. Looking at the overall, we investigate the way music changed over time with dynamic adjustments to data sets. Then we compared the differences in influence and musical characteristics with and without environmental factors. In the end, we write to the ICM Society.

Keywords: Weighted PageRank; Vector of characteristics; Ordinary least squares.

Contents

1 Introduction

Many factors (*e.g.*, artists innate ingenuity, access to new tools) can influence artists when they create a piece of music, and hence these elements gradually constitute the evolution of music. During this process, computers and network record a detailed data base consisting of some of the factors mentioned above. To explain it further, not only does the data helps to quantify musical evolution, but it also contains the possible way a music genre affects the culture. Therefore, it is of vital importance for us to make the most use of it and analyze the evolution of music.

1.1 Problem Restatement

Our team was required to identify the influence that is measured by:

1. Establish a music influence network and analyze the influence of artists.
2. Measure the similarity of artists, and judge the difference between similarities within and between genres.
3. Find out the characteristics of the genre, study the trend of its characteristic change and the relationship between the genres.

4. Judge the influence of influencer on follower and explore the influence factors.
5. Look for major leaps and analyze possible causes.
6. Consider dynamically changing and the influence of the external environment.

1.2 Literature Review

In 2012, Joan Serrà *et al.* [1] quantified western popular music, and unveiled a number of music characteristics (*e.g.*, general usage of pitch, timbre and loudness) that have been consistently stable for more than fifty years. While other more detailed qualities are changing (*e.g.*, pitch progressions varying less, loudness level rising higher, timbres becoming more frequent).

1.3 Data Cleaning & Data Preprocessing

We discover that there are 3 items marked as *unknown* in terms of genre. Moreover, there are different artists share the same id. We then omit these items. We then apply Min-max normalization to properties provided in *full_music_data* to normalize the variables for further development. To simplify our model, we evaluate the correlation coefficients of these properties in figure 1.

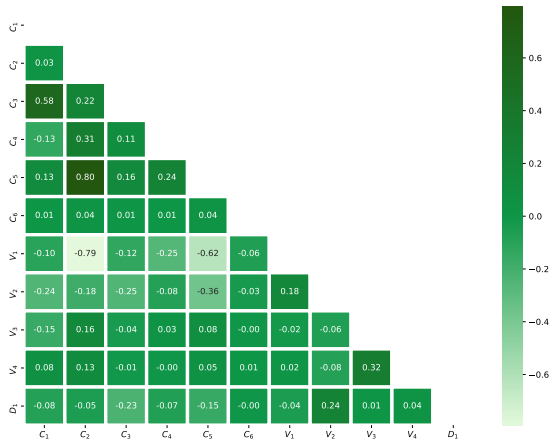


Figure 1: The correlation coefficients of the properties

The results shows that the absolute absolute value of most of the correlation coefficients are below 0.10. There are only one value reaches 0.80. Hence, the variables are not closely related to each other. Thus, we didn't process the dimensionality reduction.

1.4 Modeling Framework

Our modeling Framework is illustrated in figure 2.

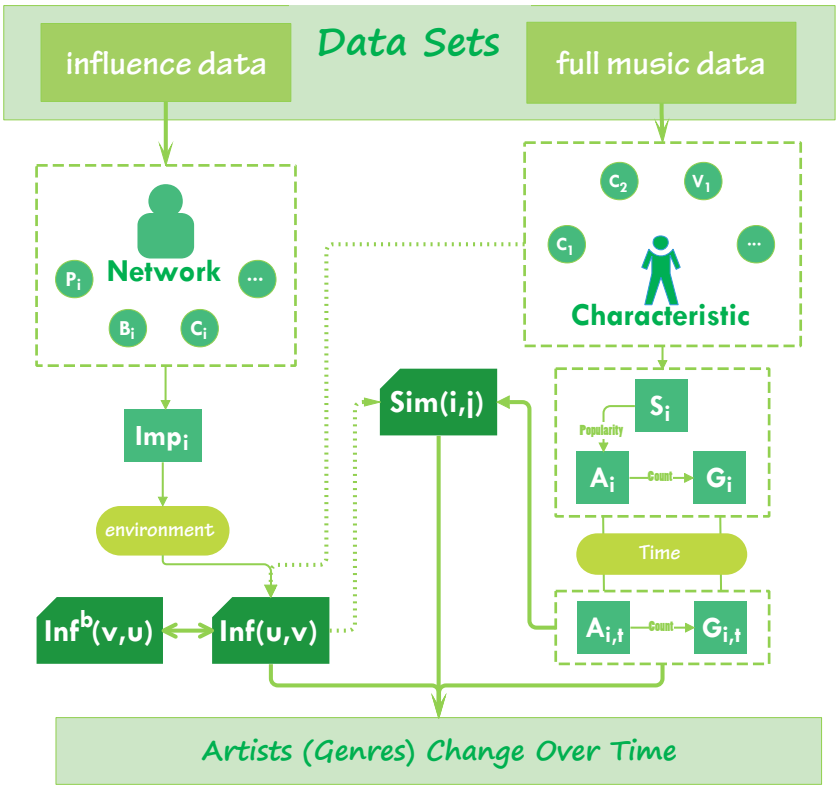


Figure 2: Modeling Framework

1.5 Assumptions & Nomenclature

To simplify our modeling, we make the following assumptions:

- Provided data include all the music around the world.
- The influence of an artist can be measured according to the network.
- The importance of each artist is related to the number of songs created and the popularity of the songs.
- Some of the effect of the properties' influence is not linearly increased.

The main notations of Network Model and Similarity Model are listed in table 1.

2 Network Model

2.1 Model Overview

Based on the *influence_data* data set, we create a weighted directed graph. The nodes of the graph are composed of artists. The edges of the graph are defined as the relationship between artists, and the direction is from influencer to follower. Note that the music genre of

Table 1: Notations we use in further modeling

Sybols	Description
i_n	Songs, $n = 1, 2, 3, ..., I$, default as i
j_n	Artists, $n = 1, 2, 3, ..., J$, default as j
k_n	Genres, $n = 1, 2, 3, ..., 19$, default as k
$\mathbf{S}_{i_n(11\times1)}$	The characteristic of song i_n, default as $\mathbf{S}_{i(11\times1)}$
$\mathbf{A}_{j_n(11\times1)}$	The characteristic of artist j_n, default as $\mathbf{A}_{j(11\times1)}$
$\mathbf{G}_{k_n(11\times1)}$	The characteristic of genre k_n, default as $\mathbf{G}_{k(11\times1)}$
$Inf(u, v)$	The influence of u on v, $u, v = \mathbf{S}_{i(11\times1)}, \mathbf{A}_{j(11\times1)} or \mathbf{G}_{k(11\times1)}$
$Sim(u, v)$	The similarity of u on v, $u, v = \mathbf{S}_{i(11\times1)}, \mathbf{A}_{j(11\times1)} or \mathbf{G}_{k(11\times1)}$

the affected artist is not completely consistent with that of the influencer. Therefore, the initial weights of edges are specified as:

$$w_{uv} = \begin{cases} 1, & genre_u = genre_v \\ 0.6, & genre_u \neq genre_v \end{cases} . \tag{1}$$

After that, we consider that one's attention is limited, and the style of the work is also limited. When an artist learns from many artists, we think the influence from each artist is relatively small. The influence among artists is not a linear accumulation, so consider using the logarithm of the in-degree of the node to modify the weight:

$$w'_{uv} = \frac{w_{uv}}{1 + \ln\left(deg_v^{in} + 1\right)}, \tag{2}$$

where deg_v^{in} represents the in-degree of node v.

2.2 Attributes of the Network

Connectivity The whole music influence network is non-connected. By searching the connected components, we get three weakly connected subnets, which are composed of 5599, 2 and 2 nodes respectively, showing a great imbalance. The small network of 2 nodes is not of research significance, so the discussion in this article is based on the larger subnet here.

Aggregation The average clustering coefficient of the network describes the degree of clustering between vertices in a graph, n is the number of nodes, and c_v is the local clustering coefficient of node v.

$$C = \frac{1}{n} \sum_{v \in G} cluster_v. \tag{3}$$

Scale-free Network We calculate the degree distribution in the entire network, and discover that the degree of most nodes is small, while the degree of a few nodes is large (see figure 3).

The results show an obvious fat-tailed distribution characteristic, which means that few top artists have a great impact and influence on the followers, while most of the artists have little

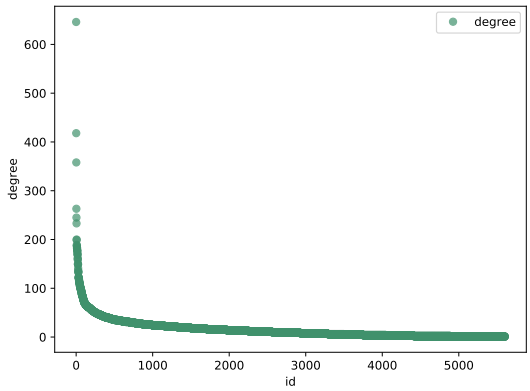


Figure 3: Degree distribution

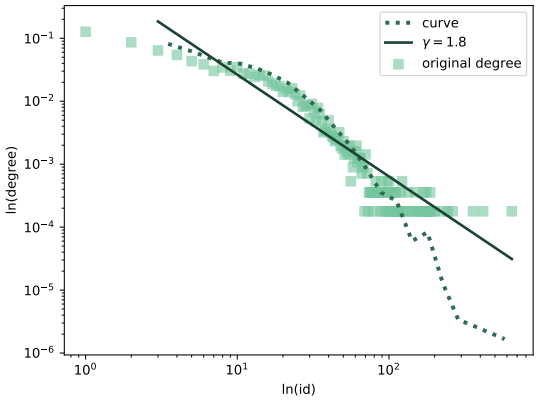


Figure 4: power law curve

effect on others. This network can be roughly considered as a scale-free network.[1]

We quantitatively analyze of the power-law distribution characteristics of its scale-free network. Here, maximum likelihood is used to estimate the parameters and KS (Kolmogorov-Smirnov) test is performed instead of least squares method and R^2 testing test.

The power law curve are generally drawed on the double logarithmic axis. As shown in figure 4, the maximum likelihood estimation curve and the fitting curve are also drawn.

It is generally considered that the network whose power law degree distribution index satisfies $2 < \gamma < 3$ is a strictly defined scale-free network. Here, the network whose $\gamma = 1.8$ can be approximated as scale-free network.

2.3 Measures of Music Influence

2.3.1 Centrality Composition

The influence of artist reflects in centrality of node. Based on the node degree, calculate the degree centrality and clustering coefficient of the node; Based on the shortest path, calculate the closeness centrality and betweenness centrality of the node; Based on the random walk, calculate the PageRank index of influence.

Clustering Coefficient The clustering coefficient is used to describe the degree of interconnection between adjacent points of a point, that is, in the music circle, the degree to which the artist's friends or admirers know each other.

For directed weighted graphs, consider the weights of edges, and since the music influence discussed here should be one-way, the out-degree is used. deg_u^{out} is the out degree, and the edge weights are normalized as $\hat{w}_{uv} = w_{uv}/\max(w)$.

$$Cluster_u = \frac{1}{deg_u^{out}(deg_u^{out} - 1)} \sum_{vw} (\hat{w}_{uv}\hat{w}_{uw}\hat{w}_{vw})^{1/3}. \tag{4}$$

[1]The scalefree network is a concept put forward by A.-L. Barabási in 1999, which is simply a network whose degree distribution obeys a power law distribution.

Degree Centrality Degree centrality indicates the direct influence of nodes in the network.

$$Degree_u = deg_u^{out}. \tag{5}$$

Closeness Centrality Closeness Centrality is an indicator that detects nodes that can effectively spread information through subgraphs. In this network, it reflects the direct degree of influence between artists by measuring distance:

$$Closeness(u) = \frac{1}{\lambda \sum_{v \in V} d(u,v) + (1-\lambda) \sum_{w \in V} d(v,u)}, \lambda \in [0,1], \tag{6}$$

where, $n-1$ is the number of nodes that can be reached by node u in the directed graph, $d(v,u)$ is the shortest path from node u to node v along the direction of influence propagation; The weighting factor λ reflects the preference for direction.

Betweenness Centrality Betweenness Centrality detects the bridge node connecting the two parts of the graph. For the music influence network, it may reflect the existence of cross-style artists, or those artists who learned much from others and became accomplished.

$$Betweenness_u = \sum_{s,t \in V} \frac{\sigma(s,t \mid u)}{\sigma(s,t)}. \tag{7}$$

Weighted PageRank PageRank is a way to measure the importance of website pages. Here, it measures the influence considering of both the influence of artists themselves and their neighbors.

$$PR(u) = \frac{1-d}{N} + ds_u \sum_{v \in M(u)} \frac{w_{vu} PR(v)}{L(v)}. \tag{8}$$

The above formula 8 is the weighted and revised PageRank formula. $M(u)$ is the collection of artists that influence artist u, $L(u)$ is the number of artists affected by artist u. d is the damping coefficient. w_{uv} represents the edge weight (reflect the intensity of influence), s_u represents the direct influence weight of node u, calculated as follows:

$$s_u = \frac{deg_u^{out}}{\sum_{v \in V} deg_v^{out}}. \tag{9}$$

2.3.2 Total Influence

The importance of an artist in the network should be considered comprehensively. The indicator is weighted, and the weight is $\boldsymbol{W} = [0.6, 0.1, 0.1, 0.1, 0.1]$. We think Pagerank is the most representative, so its weight is larger, and other features are given equal weights.

2.3.3 Specific Influence

The influence $Inf(u,v)$ is determined by the importance of the artist Imp_u and the weighted shortest path distance from the u to v which is calculated with ***bellman ford*** algorithm. $Path(u,v)$ represents the shortest path node set:

$$Path(u,v) = \{(u,k),(k,i),...,(j,v)\}, \tag{10}$$

then the influence of u on v can be calculated as follows:

$$Inf(u,v) = Imp_u \cdot \prod_{(i,j)\in Path(u,v)} w_{ij}, \tag{11}$$

and we define $Inf^b(v,u)$ also as the influence of u on v, *i.e.*, $Inf^b(v,u) \equiv Inf(u,v)$. The effect of influence is related to the importance of the artist, and decreases with the distance of transmission distance.

Considering the large scale of the network (up to several thousand nodes), the calculation amount according to the above formula may be too large; and considering the cumulative diminishing effect, when the number of passes is large, the effect will approach 0, so the scope is limited. When $Dis(u,v) > MAX_L$, it is considered to have no effect.

By establishing a music network, we can get the influence $Inf(i,j)$ of the ith artist on the jth artist in the network, assuming that the lifetime of the artist i will produce n other artists In order to calculate the influence, the total influence of this artist can be calculated Inf_i^A.

$$Inf_i^A = \sum_{j=1}^{n} Inf(i,j). \tag{12}$$

Similarly, according to the division of genres in the data set *influence_data*, we can calculate the influence of each genre. Suppose there are n artists in genre i, among which the artist j has an impact on other m_j artists, and the impact on the genre k is $Inf(j,k)$. The influence of the genre can be obtained by adding up:

$$Inf_i^G = \sum_{j=1}^{n} Inf_j^A = \sum_{j=1}^{n}\sum_{k=1}^{m_j} Inf(j,k). \tag{13}$$

2.4 Analysis of Subnetwork

2.4.1 Construction of Subnetwork

The previous network is too complex to analyze, and thus, we select the subnetwork for further research. Based on the node k of the network, we adopt *DFS* algorithm[2] to determine the subnetwork of node k, (*i.e.*, the impact network of node k). We then choose the impact network of node 130173 (Crowded House[3]) for analysis. The results are in table 2 and figure 5 .

2.4.2 Analysis of Crowded House

This is a large network of 467 artists and 894 impacts on them. Each artist build relationships with about 2 other musicians on the average. The diameter of the network is 12, indicating

[2]Depth-first search (DFS) is an algorithm for traversing or searching tree or graph data structures. The algorithm starts at the root node (selecting some arbitrary node as the root node in the case of a graph) and explores as far as possible along each branch before backtracking.

[3]Crowded House are a rock band, formed in Melbourne, Australia, in 1985.

Table 2: Influence measurements of subnetwork of Crowded House

Page rank	Cluster index	Degree centrality	Closeness centrality	Betweenness centrality	Total score
0.059296	0.114919	0.048062	0.486172	0.004734	0.11

Figure 5: Subnetwork of Artist Crowded House

that the artist's influence can be transmitted along 12 people as far as possible. The density of the network is approximately 0 (0.004), which implies the structure of sparse network.

We also discover some other interesting facts that through layers of networks, even a unpopular artists may affect numerous people indirectly. For example, Crowded House (rank 1446/5603, 25% in our ranking) indirectly affected over 500 people, and had an impact on world famous band, Cold Play (rank 44/5603, 0.8%), though being unpopular according to the search engine. After further research, Crowded House directly affected Travis (134/5603, 2.4%), then Travis directly affected Cold Play, announced by the artists themselves. Besides, there are two-way edges within artists (*e.g.*, Pearl Jam & Hurt), which indicate that affected each other mutually.

3 Similarity Model

3.1 Model Overview

In this section we are required to develop a measure of music similarity, based on this indicator, we can discover whether artists within genre are more similar than artists between genres. In order to represent the similarity, we adopt Similarity Model.

In section Selected Properties of Music and Three Levels of Indicators below, this model

evaluates the properties of music by constructing 3 vectors $\mathbf{S}_{i(11\times1)}$, $\mathbf{A}_{j(11\times1)}$ and $\mathbf{G}_{k(11\times1)}$ on the scale of songs, artists and genres.

Then in Measurements for Similarity, it offers a way to calculate the pairwise similarity $Sim(i, j)$ of songs, artists and genres.

The results enable a serious of t tests on artists between a particular genre and corresponding all other genres.

3.2 Selected Properties of Music

In order to select the properties that represent the characteristic of a song precisely, we omit *popularity*, *count* and *explicit* 3 attributes. From the perspective of the song itself, we believe that the *popularity* is more suitable for characterizing singers, and *count* is more suitable for characterizing musical genres. Moreover, whether the music are explicit may be due to the artists growth environment and education level.

The rest properties of music are listed as below in table 3.

Table 3: Selected Properties of Music

Symbol	C_1	C_2	C_3	C_4	C_5	C_6
Description	danceability	energy	valence	tempo	loudness	key
Symbol	V_1	V_2	V_3	V_4	D_1	
Description	acousticness	instrumentalness	liveness	speechiness	duration_ms	

3.3 Three Levels of Indicators

Based on the data provided and the properties selected above, we define 3 indicators represent the music characteristics of a song, an artist and a genre respectively.

In order to represent the characteristic of a song i, based on the variables listed in the in table 3 above, we develop a 11-dimensional column vector $\mathbf{S}_i$ as a song indicator:

$$\mathbf{S}_{i(11\times1)} = [C_{1i}, \cdots, C_{6i}, V_{1i}, \cdots, V_{4i}, D_{1i}]^{'}. \tag{14}$$

Subsequently,

$$\mathbf{Pop}_{i(s\times1)} = [p_1, p_2, \cdots, p_s]^{'}, \tag{15}$$

$$\mathbf{A}_{i(11\times1)} = \frac{1}{s} \cdot \left[\mathbf{S}_{1(11\times1)}, \mathbf{S}_{2(11\times1)}, \cdots \mathbf{S}_{s(11\times1)}\right]_{(11\times s)} \cdot \mathbf{Pop}_{i(s\times1)}. \tag{16}$$

Similarly,

$$\mathbf{Count}_{i(n\times1)} = [c_1, c_2, \cdots, c_n]^{'}, \tag{17}$$

$$\mathbf{G}_{i(11\times1)} = \frac{1}{n} \cdot \left[\mathbf{A}_{1(1\times11)}, \mathbf{A}_{2(1\times11)}, \cdots \mathbf{A}_{n(1\times11)}\right]^{'}_{(11\times n)} \cdot \mathbf{Count}_{i(n\times1)}. \tag{18}$$

From the equation 14, 16 and 18, $\mathbf{S}_{i(11\times1)}$, $\mathbf{A}_{i(11\times1)}$ and $\mathbf{G}_{i(11\times1)}$ are vectors of the same dimension, and thus they are mathematically interchangeable.

3.4 Measurements for Similarity

Michel Deza *et al.* [2] recorded two way to calculate the distance of two vectors, *viz.*, Euclidean distance and cosine similarity. The former focuses on the absolute similarity, while the latter focuses more on the relative similarity. Accordingly we make our measures of similarity contains both two functions. Specifically, we define our similarity between song i and song j as:

$$\begin{aligned} Sim(\mathbf{S}_{i(11\times1)}, \mathbf{S}_{j(11\times1)}) &= \alpha Sim^c(\mathbf{S}_{i(11\times1)}, \mathbf{S}_{j(11\times1)}) \\ &+ (1-\alpha) Sim^e(\mathbf{S}_{i(11\times1)}, \mathbf{S}_{j(11\times1)}) \end{aligned} \in [0,1], \tag{19}$$

where α is the weight of $Sim^c(i,j)$, and $(1-\alpha)$ is then the weight of $Sim^e(i,j)$; $Sim^c(i,j) \in [0,1]$ represent cosine similarity; $Sim^e(i,j) \in [0,1]$ is the transformation of Euclidean distance. cosine similarity is calculated as:

$$Sim^c(\mathbf{S}_{i(11\times1)}, \mathbf{S}_{j(11\times1)}) = \frac{\mathbf{S}_{i(11\times1)} \cdot \mathbf{S}_{j(11\times1)}}{\left|\mathbf{S}_{i(11\times1)}\right| \left|\mathbf{S}_{j(11\times1)}\right|}. \tag{20}$$

The original Euclidean distance is:

$$Sim^{eo}(\mathbf{S}_{i(11\times1)}, \mathbf{S}_{j(11\times1)}) = \left|\left(\mathbf{S}_{i(11\times1)} - \mathbf{S}_{j(11\times1)}\right)\right|, \tag{21}$$

however, the dispersion of different properties varies significantly (*e.g.*, the standard deviation of energy is 0.2152 while the duration_ms' is 0.0519). We thus define weighing matrix $\mathbf{\Omega}_{(11\times11)}$ to balance the dispersion. $\mathbf{\Omega}_{(11\times11)}$ is a diagonal matrix, and the $diag(\mathbf{\Omega}_{(11\times11)})$ is measured as:

$$diag(\mathbf{\Omega}_{(11\times11)}) = [\omega_1, \omega_2, \cdots, \omega_{11}]', \tag{22}$$

as mentioned above, $\mathbf{\Omega}_{(11\times11)}$ indicate the weight of properties, particularly, ω_i indicates the weight of property i, and its value equals to the reciprocal of the standard deviation of property i.

Moreover, we make the adjustment below, so as to enable that $Sim^e(i,j)$ increases while similarity grows:

$$Sim^e(\mathbf{S}_{i(11\times1)}, \mathbf{S}_{j(11\times1)}) = 1 - \frac{\left|\mathbf{\Omega}_{(11\times11)} \cdot \left(\mathbf{S}_{i(11\times1)} - \mathbf{S}_{j(11\times1)}\right)\right|}{\left|diag(\mathbf{\Omega}_{(11\times11)})\right|}. \tag{23}$$

Please note that equation 19 can also be adopted to calculate the similarity between artist i and artist j, genre i and genre j, artist i and genre j, *etc.*, since $\mathbf{S}_{i(11\times1)}$, $\mathbf{A}_{i(11\times1)}$ and $\mathbf{G}_{i(11\times1)}$ are vectors of the same dimension.

To examine whether artists within genre are more similar that artists between genres, we hence consider $Sim(\mathbf{A}_{i(11\times1)}, \mathbf{G}_{j_1(11\times1)})$ in the equation above as the similarity within artists of genre j_1 , and $\frac{1}{18}\sum_{j\neq j_1}^{18} Sim(\mathbf{A}_{i(11\times1)}, \mathbf{G}_{j(11\times1)})$ as the similarity between performers of genre j_1 and of all other genres.

3.5 Model Results

Artists similarities within and between genres are presented in figure 6.

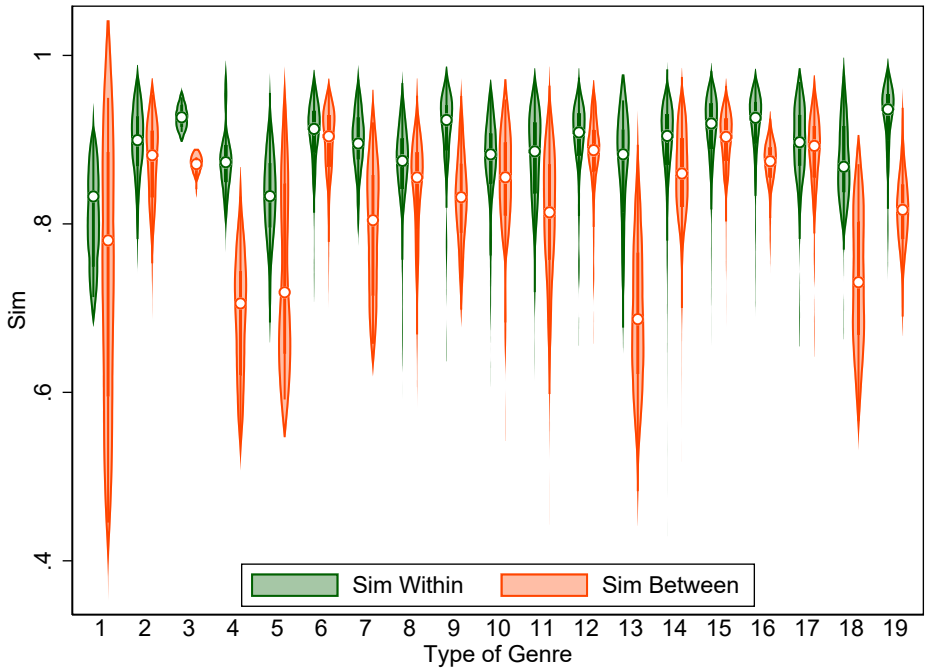


Figure 6: The violin plot of artists similarities

The green boxes represent similarities within a genre, while the orange boxes represent similarities between genres. Every green box represent the similarity within a genre, from genre 1 on the left to genre 20 on the right. And each of the green boxes has its corresponding orange box next to it, representing the similarity between genre n and all others. As a result, each of the green boxes has higher similarity compared with its orange box, and hence the artists within genres are more similar than artists between genres.

Table 4 below lists the results of t test of the difference between the similarities of artist within and between genres.

Table 4: T-test of similarity difference

genre$_i$	Sim^e		Sim^c		Sim		genre$_i$	Sim^e		Sim^c		Sim	
	T	P_u	T	P_u	T	P_u		T	P_u	T	P_u	T	P_u
g_1	1.5693	0.0738	1.9726	0.0384	1.7360	0.0566	g_{11}	16.8052	0.0000	16.1314	0.0000	16.8430	0.0000
g_2	6.0073	0.0000	5.7101	0.0000	6.0318	0.0000	g_{12}	9.6299	0.0000	8.8280	0.0000	9.8428	0.0000
g_3	5.3034	0.0065	6.7355	0.0033	5.9521	0.0047	g_{13}	8.3017	0.0000	7.5682	0.0000	8.0971	0.0000
g_4	11.9488	0.0000	10.5865	0.0000	11.7245	0.0000	g_{14}	35.1403	0.0000	31.5129	0.0000	34.5242	0.0000
g_5	4.1746	0.0001	4.6120	0.0000	4.3142	0.0000	g_{15}	21.3520	0.0000	17.3230	0.0000	20.8934	0.0000
g_6	9.6492	0.0000	8.7429	0.0000	9.5714	0.0000	g_{16}	15.8625	0.0000	12.1818	0.0000	15.3980	0.0000
g_7	5.0815	0.0000	4.7230	0.0001	5.0124	0.0000	g_{17}	5.7373	0.0000	5.6427	0.0000	5.8425	0.0000
g_8	9.5292	0.0000	9.0066	0.0000	9.5458	0.0000	g_{18}	8.8073	0.0000	8.1420	0.0000	8.6331	0.0000
g_9	12.0012	0.0000	10.7541	0.0000	11.7925	0.0000	g_{19}	22.0144	0.0000	19.8819	0.0000	21.8166	0.0000
g_{10}	3.9721	0.0001	4.2910	0.0000	4.1121	0.0000							

Apart from the genre$_1$, all p value are significant at the 1 percent level, indicating a stable discrepancy between the similarity of artists between and within genres. Thus it is concluded that generally speaking, artists shares a common characteristics among artists of same genres, and different styles of artists very much.

4 Similarities, Influences & Genre Indicators

4.1 Similarities & Influences

In section Network Model and Similarity Model, we calculated the influence and similarity of artists. According to the rules of data conversion between different levels in Models 1 and 2, the similarity and influence between genres can be calculated. According to the data, figure 7 can be made.

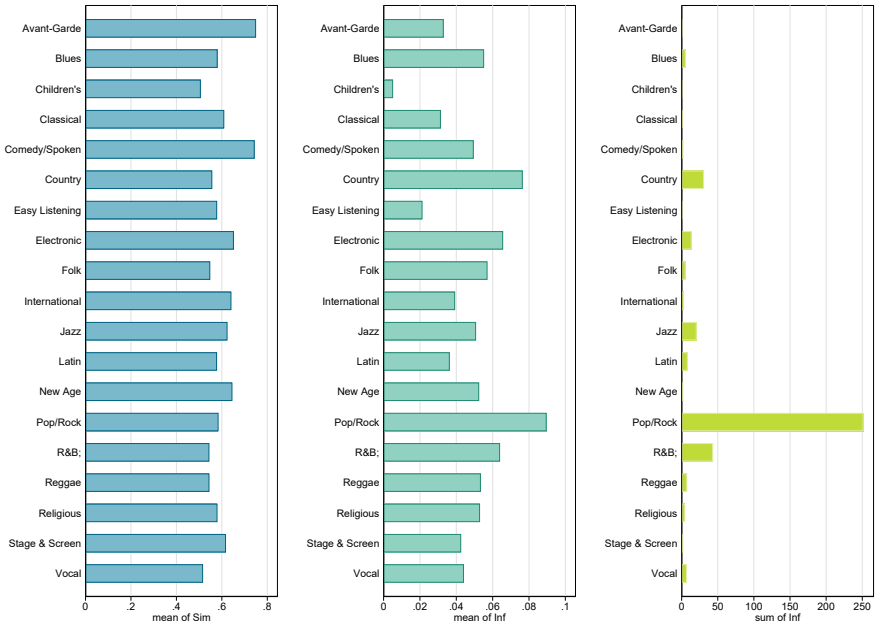


Figure 7: Distributions of similarity and influence of different genres

It can be seen from the figure that there is a huge difference in influence between different genres. Whether it is total or average, Pop/Rock has the strongest influence; from the similarity point of view, the difference in similarity between the genres is not as good as the influence The force is so obvious, but it still exists.

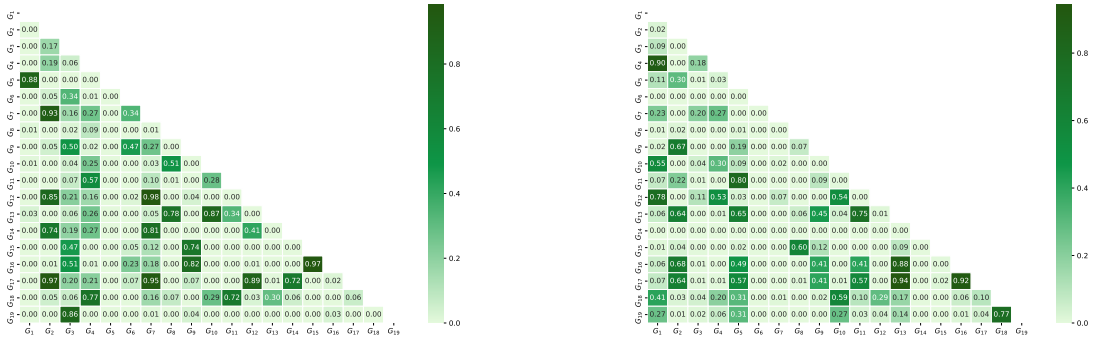
In order to more accurately measure whether the differences between the genres are significant, we use the genre as the categorical variable to conduct a one-way analysis of variance on the similarity and influence. The results of the two are listed in table 5.

Table 5: Results of one-way analysis

Variable	F-statistics	P-value	χ^2-statistics	P-value
Inf	64.68	0.0000	168.07	0.0000
Sim	24.30	0.0000	53.51	0.0000

The result shows that the data does not satisfy the assumption of equal variances, so we need to perform a pairwise independent sample t test. The pairwise difference level of the p value of the test is used to plot the heat map as shown below.

The depth of the color indicate the discrepancy between 2 genres. It can be viewed that



(a) Heat map of t test of Sim (b) Heat map of t test of Inf

Figure 8: Heat map of t test

the similarity of G_1 and the influence of G_{14} are quiet different from the others. And we can determine the certain genres that cause the differences.

4.2 Genre Indicators via 2 Measures

Based on the properties defined in Selected Properties of Music, we pick some variables as indicators for genres. When determining the attributes, we believe that if the genre g_i, compared with other genres, has some characteristics that are significantly different , then the variable (*viz.*, C_i, V_i, D_i) is the main characteristic of the genre.

To determine significant difference, we adopt 2 measures for a comprehensive determination. Firstly, based on $\mathbf{G}_{i(11\times1)}, i = 1, ..., 19$ (*i.e.*, 19 vectors that represent 19 genres' features) in Three Levels of Indicators composing a matrix $(\mathbf{G}_1, ..., \mathbf{G}_{19})_{(11\times19)}$, each of the 11 properties are matched with 19 genres. Then, we test outliers and deviations of each propertie in figure 9.

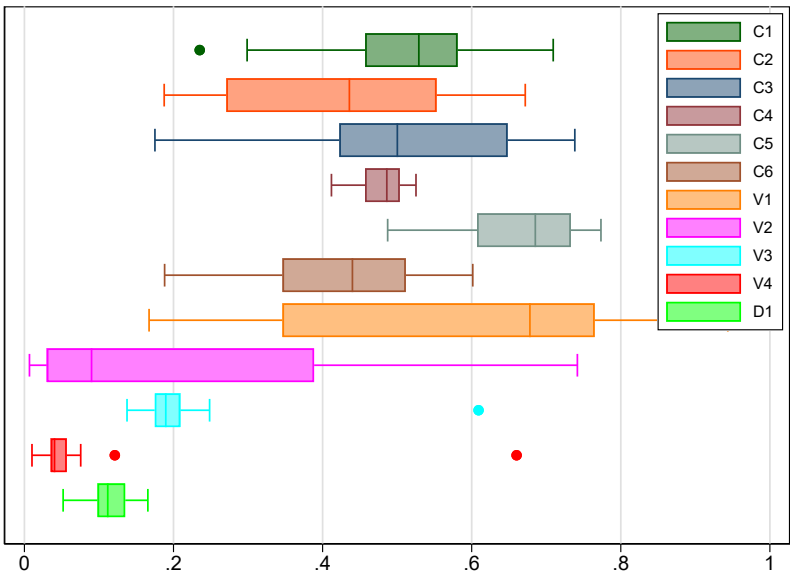


Figure 9: Box plot of properties, each of the boxes (*i.e.*, properties) have 19 values (*i.e.*, genres)

The results mark four outliers, which represent danceability (C_1), liveness (V_3), and speechiness (V_4) respectively. This means that these features have the potential of indicating a music

genre. Secondly, we adopt another measure for a comprehensive determination: the data outside $[\mu-3\sigma, u+3\sigma]$ is considered as abnormal points, and two abnormal points are discovered, which belongs to liveness (V_3), and speechiness (V_4). That is to say, danceability (C_1) is detected as perceptible with fist determination, and liveness (V_3) and speechiness (V_4) are discovered as conspicuous with both methods, the indicators are shown in the following table 6.

Table 6: Genre indicators discovered by two means

Genres	Genre Indicators		
	Danceability	Liveness	Speechiness
Stage & Screen	$\downarrow^{\#}_{**}$	-	-
Comedy/Spoken	-	$\uparrow^{\#}_{****}$	$\uparrow^{\#}_{****}$
Reggae	-	-	$\uparrow^{\#}$

\# indicates the ouliers disclosed via first measures;
$\uparrow$ indicates whether the value of genres are lower or higher than means;
the number of $*$ represents the times the value of genres outreach the means.

4.3 Changes of Genre

4.3.1 Overall Evaluation

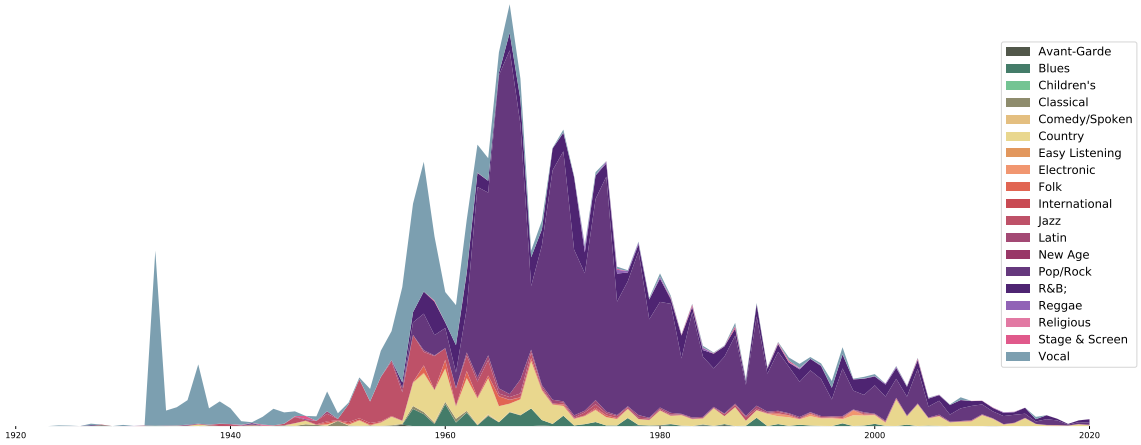


Figure 10: Overall count

As is shown in figure 10, the amount of songs sprouted from 1950s, reached its peak in 1970s, and slowly decreased after 1980. Before 1960s, Vocal has been the main stream of the music, while Pop/Rock began to flourish in 1960s and quickly dominant the field.

4.3.2 Intricate Evaluation

In consideration of the fact that there are too many genres and properties to present and analysis, we choose few long lasting genres for detailed discussion (*viz.*, Jazz, Reggae and Vocal) in figure 11.

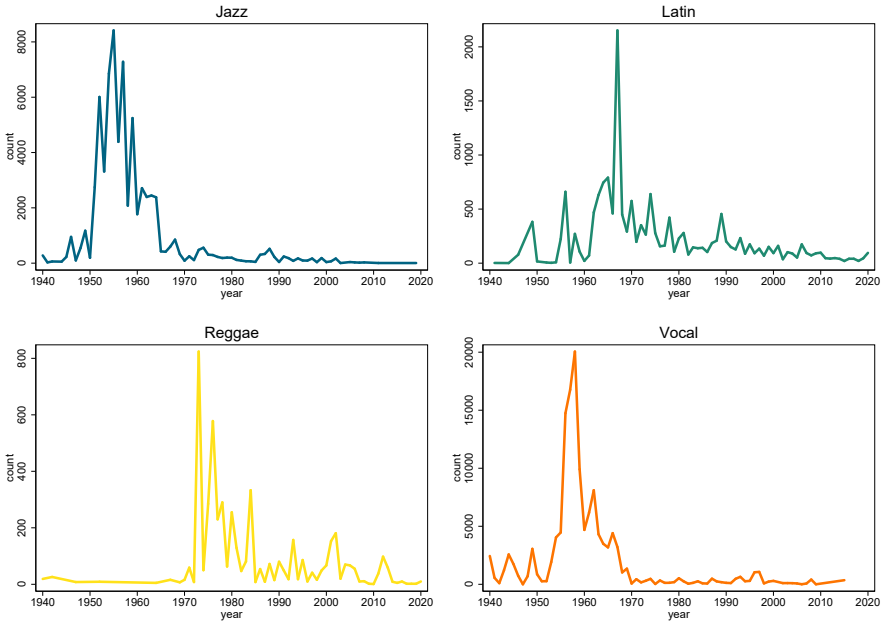


Figure 11: Count varies in different genres

As for the evolution of genres, there is a golden age of music development in 1950s – 1970s, which reaches to its peak in around 1960s. Therefore, Jazz, Latin and Vocal shared a broom during that time, and faded after 1970s. However, Reggae thrived for 10 years after 1970s.

After the analysis at the macro level, we include the evolution of 11 properties along with *popularity* in figure 12.

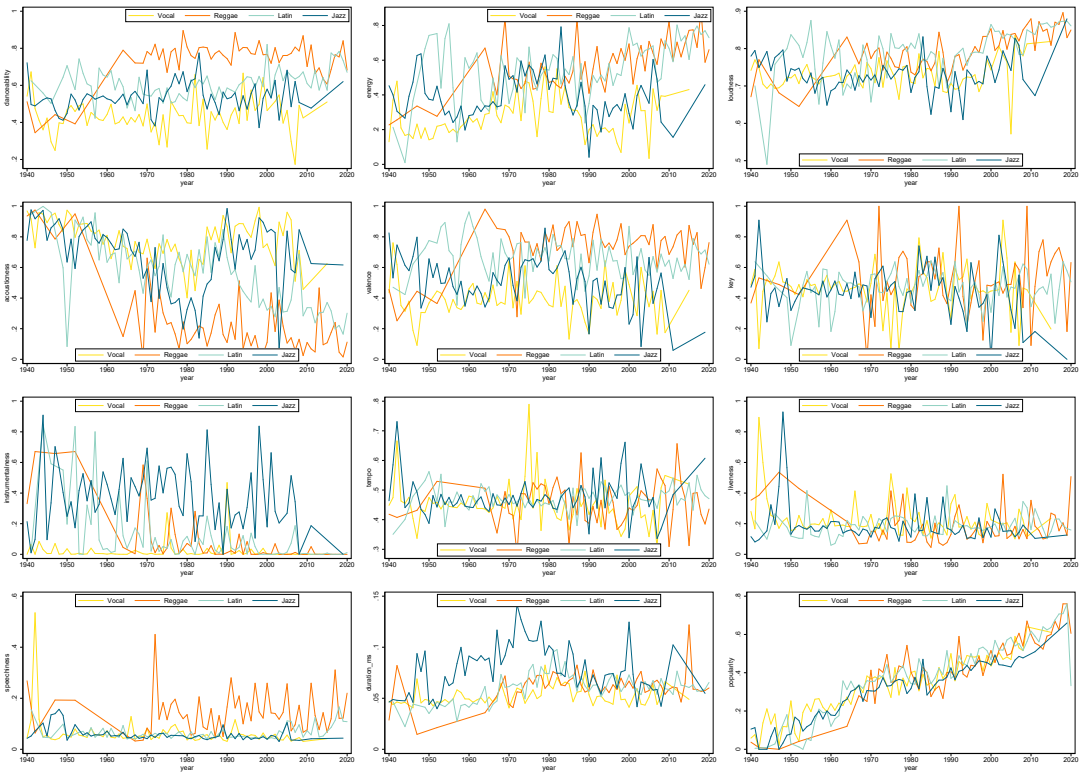


Figure 12: Characteristics vary in different genres

Pictures of figure 12 in the first column from left vary considerably, which reflect the individual differences within genres. To explain it further, we can discover that the *danceability* of Reggae is higher than other genres, and its *acousticness* is lower than others. Figure 12(1, 3) manifests that Jazz has higher *instrumentalness*, and Figure 12(1, 4) illustrates that Reggae has higher *speechiness*, which is in accordance with former research.

The data in the middle column is a bit jumbled, indicating that different genres have their own characteristics and styles in periods. To make further judgments, we need more detailed data bases or measurements.

In the rightmost column, the characteristics of each genre are relatively concentrated, which can be considered to reflect the overall trend of changes in the music industry to a certain extent. Upon observation we noticed that regardless of the genre, the *popularity* of music has a clear upward trend, which may be related to the increasing popularity of music. Further more, after the dramatic fluctuations in *liveness*, the fluctuations begin to decrease, indicating that over time, different genres have reached a unity in this regard. Additionally, the change of *key* revolves around an average value, fluctuating continuously, and there is still no trend of convergence till 2020.

4.4 Relations between Genres

The relations between genres, each of the similarities between genre k_1 and all other genres in year t, *i.e.*, $Sim(\mathbf{G}_{k_1(11\times1),t}, \mathbf{G}_{k_n(11\times1),t}), k_n \neq k_1$ can be measured by Similarity Model. For further research, we determine genres which share the highest similarities with *Avant-Garde*, and plot the scatters chronologically in the left half of figure 13 Then to compare all the genres, we define eclectic rate and present it on the right half of figure 13 .

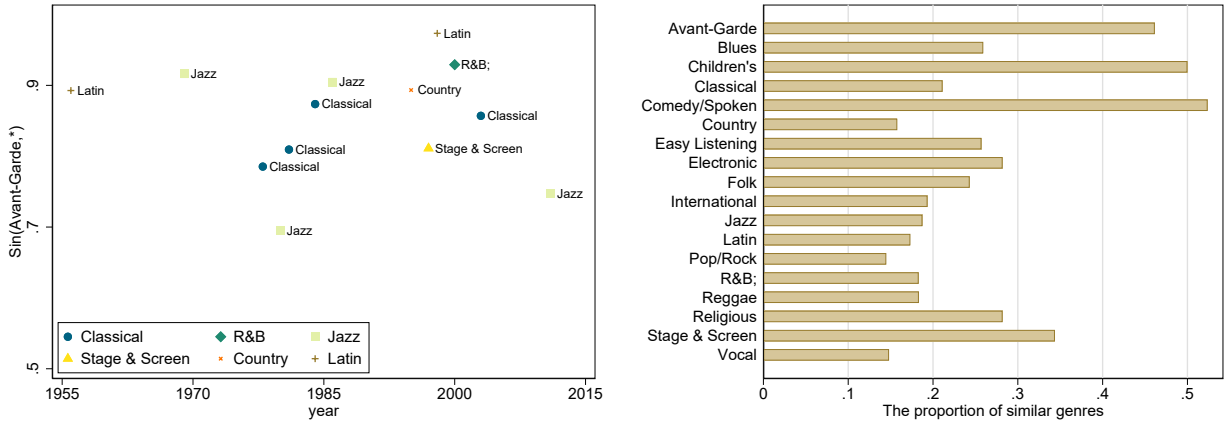


Figure 13: Relations between genres chronologically

Analyzed from the left of the figure 13, *Avant-Garde* are closely related with *Classical* and *Jazz*. From the right of the figure, we can conclude that *Avant-Garde*, *Children's* and *Comedy/Spoken* uphold other styles of music, while *Country*, *Jazz*, *Latin*, *Pop/Rock* and *Vocal* are less influenced by other music.

5 Influence of Artists and Characteristics

5.1 Authenticity of Influence

In this section, we answer whether the data suggest that the identified influencers in fact influence the respective artists, by conducting a regression of the influencers and similarity of artist v. In detail, according to the *influence_data* data set, we determine the influencers of artist v within the same genre, *e.g.*, genre 1. Then we sum up the influence as $\sum_{i=0}^{I} Inf(i, v)$. After that, we adopt $Sim(\mathbf{A}_{v(11\times1)}, \mathbf{S}_{1(11\times1)})$ as the similarity between artist v and genre 1. The results of the regression are addressed as below:

$$\hat{Sim} = \underset{(0.0033)}{0.0357 Inf^b} + \underset{(0.0023)}{0.6017} \qquad F = 117.21 \quad R^2 = 0.0205,$$

where $F = 117.21$ shows that the influencers do have a stable influence on the similarity between the artist and the corresponding genre. However, the correlation is 0.0357 and the $R^2 = 0.0205$, which reveal that the influence announced by the followers is of little trust worthiness in general.

5.2 Contagious Inspection

To quantify the properties of the music, contrary to the way we do in Data Cleaning & Data Preprocessing, we didn't normalize the data, for we are not interested in dimensional difference. Instead, we add 60 to *loudness* for latter logarithmization. In addition, we list $Count_L$, $\ln Popularity$ and $Mode$ as control variables.

The selected results are listed in table 7.

Table 7: Influence degree of each propertie

Variable	Model							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Count_L$	0.006***	0.005***	0.006***	0.005***	0.005***	0.006***	0.005***	0.005***
ln Popularity	0.011***	0.021***	0.013***	0.022***	0.020***	0.016***	0.021***	0.020***
Mode	0.001	0.002	0.003**	0.000	0.002*	0.002	0.001	0.002
Acousticness	-0.033***		-0.041***					
Acousticness^2	0.018**		0.038***					
Danceability	-0.021***			-0.040***				
Danceability^2	-0.079***			-0.085***				
ln Duration$_{ms}$	0.003*				0.007***			
(ln Duration$_{ms}$)^2	-0.005**				-0.011***			
ln Loudness	0.036***					0.095***		
(ln Loudness)^2	0.059**					0.124***		
Speechiness	-0.051***						-0.022**	
Speechiness^2	0.080***						0.021	
Valence	-0.023***							-0.021***
Valence^2	-0.023**							-0.061***
Constant	0.013***	-0.024***	-0.004	-0.026***	-0.019***	-0.008**	-0.024***	-0.018***
Observations	5,018	5,018	5,018	5,018	5,018	5,018	5,018	5,018
R-squared	0.245	0.124	0.197	0.148	0.129	0.159	0.125	0.143

Note: Standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1

The effects of all variables in the table are significant, most of the variables are significant at

a 1 percent level, and the effects of variables outside the table are not significant. The influence of *Speechiness* is the most significant, because it has the largest coefficient of 0.08.

6 Analysis of Revolution

6.1 Determination of Revolutionary Point

Based on the provided *data_by_year* data set and the Similarity Model, we calculate the similarity of musical characteristics of two consecutive years as $Sim(\mathbf{Y}_{t(11\times1)}, \mathbf{Y}_{t-1(11\times1)})$. The result are shown in figure 14.

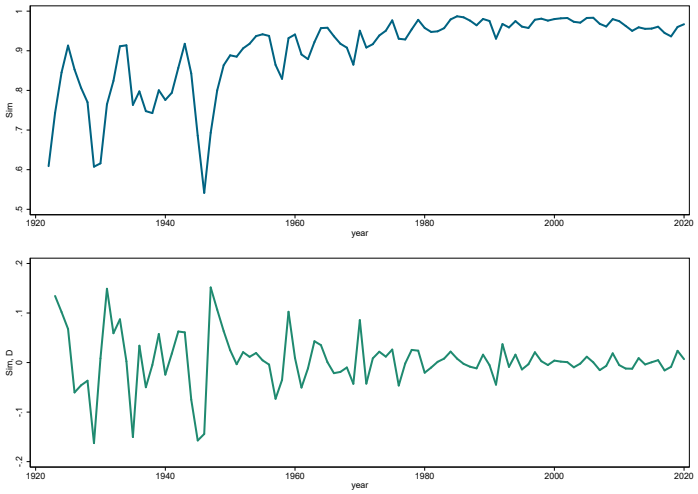


Figure 14: Yearly Similarities

There are major leaps at the point of time. The reason for this sudden change may be because the sudden change of the external environment leads to the change of people's preferences, or it may be because of the appearance of certain revolutionaries, which leads to more obvious changes in the style of music.

6.2 Determination of Revolutionaries

Because of the lack of data related to the external environment, we can only make guesses based on crucial moments:

- 1929: Possible cause, out of the Great Depression;
- 1946: Possible cause, the end of World War II.

Since the earliest data of musicians is 1930, it is impossible to explore the situation in 1929. Therefore, we mainly analyze the extent to which the mutation of 1946 was influenced by the artist. Specifically, we measures the style of artists, the similarities in 1946 and the influence conditions. The revolutionaries are Miles Davis, John Coltrane and Hank Williams, and there contributions are 10.09%, 6.15%, 5.91%.

7 Dynamic Evolution

7.1 The Indicators of Dynamic Evolution of Genres

According to the analysis in Similarity Model, the characteristics of an artist are determined by the characteristics of the works he creates. Similarly, when we talk about the characteristics of a genre, we are actually referring to the sum of the characteristics, of all the artists in the genre.

However, this mean shows no consideration for the effect of time, which means that, we believe that works and artists of different periods have the same influence on this genre. We define this similarity as $\mathbf{G}^0_{(11\times1),t}$. The changes are made so that the more timely the works are, the greater impact there will be. Accordingly, the distant works have little impact, namely:

$$\begin{aligned} G_{(11\times1),t} &= \rho G^o_{(11\times1),t} + (1-\rho) G_{(11\times1),t-1} \\ &= \rho \sum_{i=0}^{t-1} \left[(1-\rho)^n \rho \times G^o_{(11\times1),t-n} \right]. \end{aligned} \tag{24}$$

Among them, $\rho \in [0,1]$is the time coefficient, indicating the degree of emphasis on the near future, here $\rho = 0.7$.

7.2 Dynamic Changes of Genre Characteristics

According to the idea in Changes of Genre, we study the dynamic change process of the genre characteristic. We take Jazz as an example here to explore the changes in its characteristics.

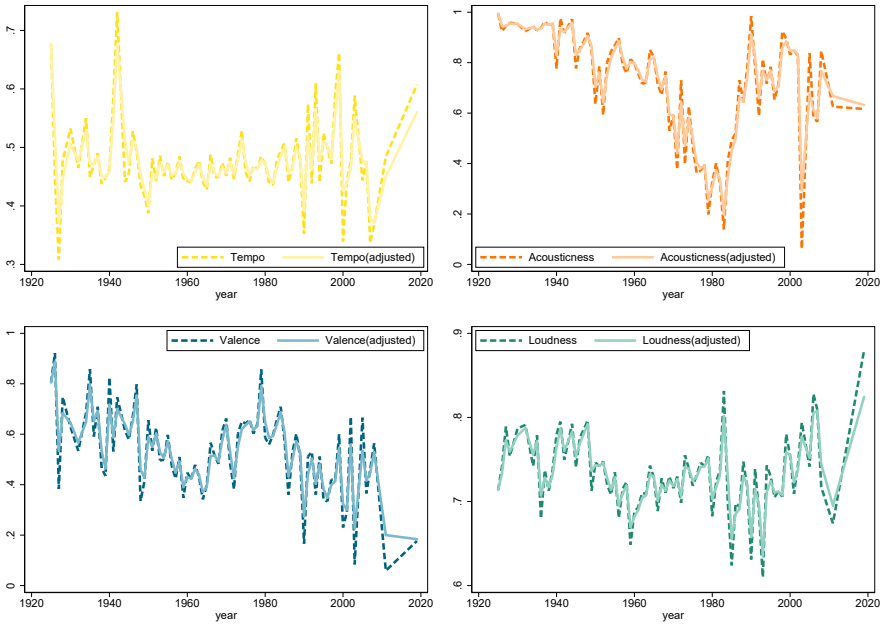


Figure 15: Characteristics after dynamic adjustments

It can be seen from the figure that compared with the original data, the fluctuation degree of the characteristics after dynamic processing becomes smaller, and the trend remains unchanged. Therefore, we can keep the original conclusion of the evolution of Jazz features. In

the last 100 years, Jazz's *Tempo* has been relatively stable. There was a large increase in 1940s and 1990s – 2010s, and then it returned to normal levels. The changes in *Acousticness* and *Valence* have obvious trends. The former declined first and then increased, while the latter has been showing a downward trend during the 100 years. *Loudness* changed around the mean value of fluctuation around 1950s, and showed an increasing trend after 1980.

8 Impact of External Factors

8.1 Cultural Factors

We repute that the characteristics of mainstream music $\mathbf{Main}_{(11\times1)}$ in period t can represent the style of the musical culture during this period. Therefore, in any period of time, the style changes of music genres can be decomposed into the influence of internal and external factors. And the impact of external factors can be considered as changes in the musical cultural environment:

$$\Delta\mathbf{Main}_{(11\times1),t} = \mathbf{Main}_{(11\times1),t} - \mathbf{Main}_{(11\times1),t-1}. \tag{25}$$

As a result, when researching the evaluation of one genre in period t (*i.e.*, $\mathbf{G}_{(11\times1),t}$), we are able to calculate the internal factors:

$$\begin{aligned}\Delta\mathbf{G}_{(11\times1),t} &= \mathbf{G}_{(11\times1),t} - \mathbf{G}_{(11\times1),t-1} - \Delta\mathbf{Main}_{(11\times1),t} \\ &= \left(\mathbf{G}_{(11\times1),t} - \mathbf{Main}_{(11\times1),t}\right) - \left(\mathbf{G}_{(11\times1),t-1} - \mathbf{Main}_{(11\times1),t-1}\right).\end{aligned} \tag{26}$$

After the extraction of external factors, we present the original characteristics and adjusted value of Jazz in figure 16 and figure 17.

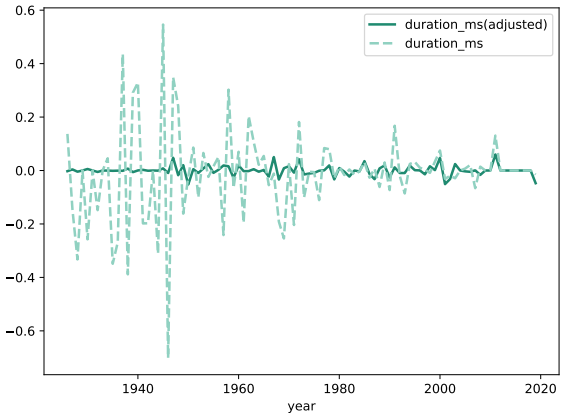


Figure 16: Original and adjusted duration

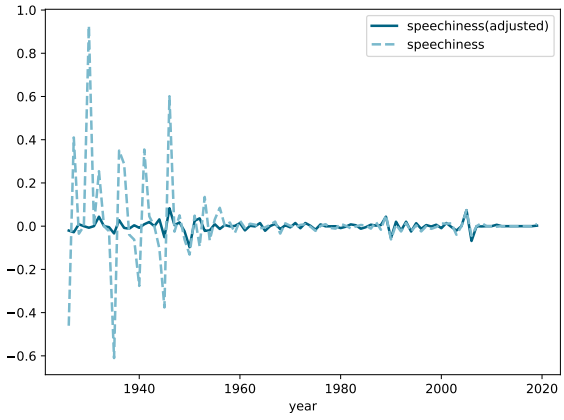


Figure 17: Original and adjusted speechiness

The dotted line in the figure represents the original feature, and the solid line represents the adjusted feature. It can be seen that after filtering the external influence, the smoothness of the characteristics has been greatly improved.

8.2 Detection for Social & Technology Changes

External factors such as the popularity of music and the internet will boost people's enthusiasm. Besides, their demand for music will increase during peace time. All these good external environments make it easier for artists to enlarge their influence.

Therefore, in order to account the changes in the external environment, we set the the number of artist active start n_t as an indicator to measure the quality of the external environment. And we made dynamic adjustments to the influence of the network as:

$$Inf_t^*(i,j) = \ln(n_t) \times Inf_t(i,j). \tag{27}$$

After the adjustment above, here are the top 10 popular musicians on the internet in table 8.

Table 8: top 10 popular online musicians

Artist name	Artist id	Rank1	Artist name	Artist id	Rank2
The Beatles	754032	1	The Beatles	754032	1
Bob Dylan	66915	2	Bob Dylan	66915	2
The Rolling Stones	894465	3	The Rolling Stones	894465	3
David Bowie	531986	4	David Bowie	531986	4
Led Zeppelin	**139026**	**5**	**Andy Black**	**3495279**	**5**
Jimi Hendrix	354105	6	Led Zeppelin	139026	6
Sex Pistols	418740	7	Jimi Hendrix	354105	7
Miles Davis	423829	8	Miles Davis	423829	8
The Beach Boys	41874	9	The Byrds	631774	9
The Byrds	631774	10	Sex Pistols	418740	10

Rank1 on the left is the original rank of influence, and Rank2 on the right is the influence after the adjustment. Generally speaking, it can be concluded that the ranking is almost unchanged: The Beatles, Bob Dylan, The Rolling Stones are top 3 of all rankings. And there are 9 artists in both 2 rankings. As our inspection go deeper, Andy Black entered the top ten after the adjustment, while Led Zeppelin faded. Interestingly, the number of search engine results of the former is about an order of magnitude higher than that of the latter[4].

9 Sensitivity Analysis

In this chapter, we need to discuss the impact of different weights on the results. We make certain adjustments to the weights and observe the changes in results.

We view the top 200 musicians (top 8.53%) in ranking of influence as high-influence musicians, and record the survivals of them under different weights. The higher the proportion of survivals (η_1), the more stable the model. The results are presented in figure 18. Similarly, we selected the 200 artists closest to the mainstream music industry and observed the changes of surviving musicians (η_2) under different weight (α). The results are listed in figure 19.

[4]The Google search result of Andy Black is $556,000,000$, and Led Zeppelin's is $63,400,000$

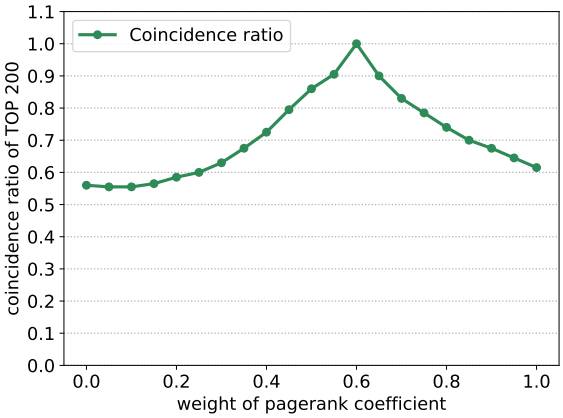


Figure 18: Sensitivity Analysis of Influence

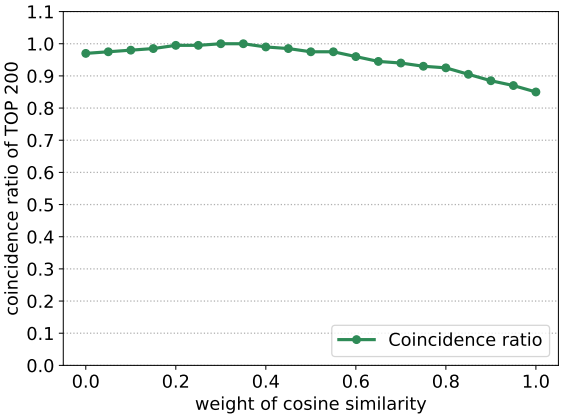


Figure 19: Sensitivity Analysis of Similarity

As can be seen from figure 18 and figure 19, the weight value in Network Model may have a certain impact on the artist's influence, but the degree of impact is generally acceptable; and the weight setting in Similarity Model also has little effect.

10 Strengths and Weaknesses

Strengths

- Comprehensive analysis of the data: We weight multiple indicators and conduct sensitive analysis.
- More developed quantitative evaluation: To stabilize our arguments, we adopt statistics beyond our subjective consideration in terms of variables and evolution trends.

Weaknesses

- Unable to elaborate on all genres: Limited by the number of pages, we only analyze some prominent and interesting phenomena, and do not comment on all the characteristics of all genres.
- The artist's music characteristics are not taken into account in our music network (but calculated separately), which may make our network information not rich enough, and influence analysis and music characteristics are not well combined.

11 The Document

Dear the Integrative Collective Music Society,

Thank you for choosing our team to deliver this meaningful research! We are informed of your specified questions and have thoroughly analyze the data provided. Based on our work, we are going to provide you with our detailed analysis.

According to the effects between artists, we established a music influence network, and fully considered the specific degree of impact.

This is a knowledge graph of music, from which we can obtain a huge amount of information. For example, we are able to study the importance of an artist in the music industry, and the number of artists directly or indirectly affected, or we can study his influence on a certain genre in a particular period; we can also quantitatively analyze the influence between any two artists, including the degree and process of influence; in addition, we can also trace the originator of music, who is imitated and learned by artists. In the entire music industry, connections are ubiquitous and intricate. When the effect of time is considered into the network, changes of influence can be dynamically analyzed, and the social backgrounds of different periods can be considered, which makes the analysis more profound.

In the process of building the model, we used all the artists, their creations and their genres to form the entire music world and participate in calculations. More accurate results will be adveloped with more relevant data. However, this will also increase the complexity of our operations. We use Sim and Inf to study the impact of music. We hence explore the computational difficulty on larger data sets.

Sim can be obtained by simple vector calculations, and the amount of computing increases linearly, which is acceptable within a considerable amount of data. Inf is obtained by calculating centrality indicators in the network, etc., a large number of graph theory algorithms and network theory are applied, and the time complexity is high. For example, the complexity of the Shortest Path Algorithm is $O(n^2)$ and the PageRank Algorithm is $O(n^2 log(n))$, that is to say, as the number of samples increases, the amount of computation will increase dramatically. When the calculation amount becomes too high with the increase of the sample, we considered reducing the calculation difficulty by setting the influence range (step size). Because the effect of the influence between artists decreases within generations, it will not cause too much disruption on the accuracy of the final result.

We explored the relationship between the artist's style characteristics and their influence. But such inspection is not very comprehensive. We did not study the interaction effects between different features, nor did we discuss the endogeneity of variables. In the future, greater exploration can be made in these areas.

As for the influence of music on culture, here we use the characteristics of the entire music industry to replace the cultural environment at that time to a certain extent. However, such behavior will lead to unavoidable endogeneity, because features cannot be used as explanatory and explained variables simultaneously. By collecting other cultural-related data, we can measure the impact of music on culture with higher accuracy.

If you want to know more details, please refer to our thesis. We will be glad to discuss with you on our details.

References

[1] Joan Serrà, Álvaro Corral, Marián Boguñá, Martín Haro, and Josep Ll Arcos. Measuring the evolution of contemporary western popular music. *Scientific reports*, 2(1):1–6, 2012.

[2] Michel Marie Deza and Elena Deza. Encyclopedia of distances. In *Encyclopedia of distances*, pages 1–583. Springer, 2009.

[3] Patrick E Savage. Cultural evolution of music. *Palgrave Communications*, 5(1):1–12, 2019.