Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
D		2121523

Quantification of music: Evaluation of Musical Influence to Artists and Genres Revolution

Summary

Music has been part of human societies for a long time. It has an indispensable impact on other cultures, meanwhile it is also greatly influenced by multiple factors, including political events, other artists and so on. To quantify the musical evolution, we develop several parameters to measure or describe the process. We also explore the relation between these parameters in an attempt to construct a comprehensive model.

One major factor in the evolution is the influence between artists. By constructing an influence network of artists, we can develop an influence parameter **P** based on the PageRank algorithm to measure the influence of an artist in the whole process of the evolution, in a particular group of artists, or in a particular historical period. This parameter serves as a tool for further study of the major leaps in musical evolution.

Another useful parameter is developed based on the features of music content. By data preprocessing, weight assigning and dimension reduction, we can extract the information for a higher distinguish degree and smaller size. We develop an AHP-PCA model for this process. Then by calculating the cosine distance between the vectors of features, we obtain a parameter **s** to measure the similarity between songs/artists/genres. The comparation of similarity between and within genres suggests **s** may play a role in genre classification. Also, the similarity between influencers and followers suggests **s** 's role in measuring fluence. However, our model does not have a good performance in classifying songs/artists into genres, possibly due to the diversity of music within genres. According to our model, most genres are closely related while only a few can be well distinguished.

Our failure in genre classification using feature information is understandable since genres always evolve from the earlier one. From another perspective, from the similarity between songs/artists and the influence network, we can develop indicators to signify the major leaps in the process of evolution. We develop two indicators, **N**, the portion of artists combined with their influence parameter **P** in a particular genre, and **F**， an indicator based on the feature information of music content. Though the meaning of **F** is unclear, the turning points of the curve based on these two parameters can indicate the occurrence of a major leap. By identification of the representative artists, we can extract a subnetwork of them and further analyze the cause of the major leap.

Our model is mainly based on the given data. However, as other cultural factors also affect the evolution of music, the trend of **N** and **F** in the same period can be a bit different. Under the structure of our model, more information can be used to assessing other factors.

Key words: Music Evolution, Similarity, Influence, Network, Feature Information

Contents

I. Introduction

1.1 Background

Music, as a miraculous form of art existing throughout the world, is a distinctive expression of ideas, an inspirational way of communication, and also an animated reflection of the society. Therefore, understanding the evolutionary and revolutionary trends of both artists and genres counts for a great deal. Scores of statements have offered various kinds of evidence like artists' self-descriptions, characteristics of songs, shifts of the society and so on to measure the musical influence. By combining all the data within the spectrum of certain mathematical models, we can probably get better understanding of the relationship between different genres, the systematic revolutionary process of the period, the factors that exert an enormous function and so on. Meanwhile, the analysis of the certain period also pushes forward an immense influence on our understanding of the specific impacts of social transformation, technological developments, cultural exchanges on music, thus promoting the developmental progression.

1.2 Problem Restatement

To solve the problem of measuring the musical influence, we need to address the following tasks:

Firstly, we need to build a directed network and develop specific parameters to capture "music influence". Then within the spectrum, a subnetwork should be constructed with deeper analysis.

Secondly, based on the offered data, specific measurements should be put forward to measure music similarities and deal with some comparing problems.

Thirdly, we shall use the established model to make a comparison of similarities and influences between and within the given genres. Moreover, together with the data of time, we need to come up with the classification basis of the genres, the revolutionary overview of the genres and their relationships.

Fourthly, according to the outcomes of the model, we need to demonstrate the authenticity of the influence and weigh the influence of the various impacting characteristics.

Next, we need to verify the certain characteristics in tight connection with the major leaps of the

music history and find out the representative artists as well.

Then, focusing on one genre, we need to deal with the process of musical evolution over time and use our model to identify indicators related to the dynamic influencers.

Finally, cultural influence or environmental impacts and other alternative impacting factors should be considered within the model. And we shall show the specific effects within the network.

1.3 Our work

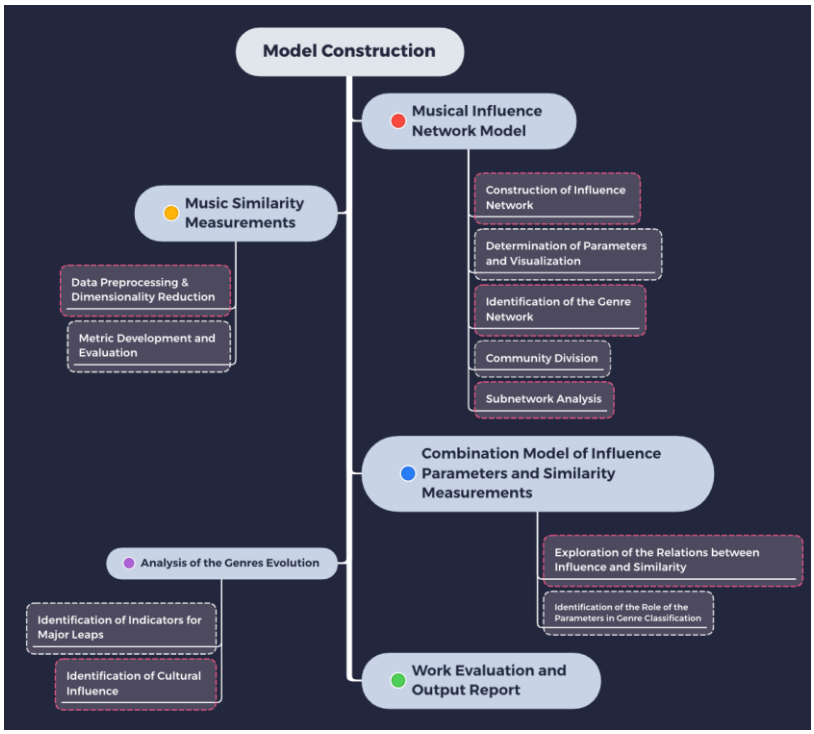


Figure 1 The structure of our work

II. Preparation of the models

2.1 Assumptions and Justifications

- Network Development Between Artists
  - With a single artist as the node, the relationship between influencer and follower is one-way, and the weight of the relationship between each pair of influencer and follower is equal.
  - The genre and year of activity between influencer and follower are not considered.
  - The in-degree calculation of nodes and the PageRank algorithm are based on the assumption of directed unweighted graphs.
- Network Development Between Genres

- Take genres as the nodes, consider the edges of two different genre, calculate the number of influencer and follower pairs belonging to the two genre, and use this number as the weight of the directed edge between the two genre.
- The influence parameter is calculated by the PageRank algorithm with edge weights.

- Community Division of the Network between Artists
  - The relationship between influencer and follower is treated as a two-way relationship, and the weights between different pairs are still equal.
  - The sub-network is taken from the result of community division.

2.2 Notations

The primary notations used are listed in Table 1.

Table 1 Main Notations List

Symbols	*Definition*
D	Degree
P	PageRank Value
I	Influence Value
a_{ij}	Comparison Value of Importance
λ_{max}	Characteristic Root
CI	Consistency Index
RI	Mean Random Consistency Index
CR	Consistency Ratio
s	Similarity Parameter
N	Indicator of Major Leaps based on the influence of artists
F	Indicator of Major Leaps based on feature information

III. Musical Influence Network Model

3.1 Construction of Musical Influence Network

Considering the relationship between influencers and followers, a directed graph which refers to

a set of vertices connected by a collection of directed edges can be used to describe the connective relations. The directed edges here are from the followers to the influencers. We define the vertex set as V, the edge set as E, thus the graph can be equated as $G = (V, E)$. Based on the graph theory, we create a directed graph to represent the network at first. To be more specific, an adjacency matrix [1] can be used to address the network analysis. We define a certain adjacency matrix A which represents a directed network with N nodes, v_i and v_j as the nodes, ω_{ij} as the number of passes between influencers and followers. The equation can be expressed as:

$$A(v_i, v_j) = \omega_{ij}$$

Then the musical influence network can be vividly presented by figures in which width of the links refer to the proportion to their weights and the size of nodes relates to the number of passes.

3.2 Determination of Network Parameters and Application

3.2.1 Degree *D*

The degree (D) determines the importance of a node by measuring the number of edges connected to it. It's a basic and direct way to evaluate the importance of the nodes in a complicated network. We define degree (v_i) as:

$$D(v_i) = \sum_{j=1}^{M} \delta_i^j, \delta_i^j = \begin{cases} 1, v_i \; is \; connected \; with \; v_j \\ 0, v_i \; and \; v_j \; are \; seperated \end{cases}$$

The degree of a node reflects the local importance of it, that is, the number of points directly connected to the node, which can identify important nodes in the network.

3.2.2 PageRank Values *P*

The high degree of a node only indicates that there are more nodes directly connected to the node, without considering the importance of the nodes connected to it. PageRank overcomes this shortcoming. While considering the number of points connected to a node, it also considers the importance of the points connected to a node. In short, PageRank works by evaluating the number and quality of links to a page to determine the importance of the website. It requires several iterations through the collection to adjust approximate PageRank values to reflect the theoretical true value more closely. Based on the relationships of influencers and followers, we can regard such relationships as the links between websites. We define the PageRank value P of a node A as:

$$P\,(A) = \alpha \cdot \frac{1}{N} + (1 - \alpha) \cdot \sum_{P_i \to A} \frac{P_i}{D_{out}(P_i)}$$

where N is the number of the nodes and α is the damping factor that refers to the probability of the links.

We compare D of the top 100 nodes and P (see Figure 2), and it can be seen clearly that the changing trends of P and D are roughly the same, but P has a more global reference value. Therefore, PageRank value can be used as an influence parameter.

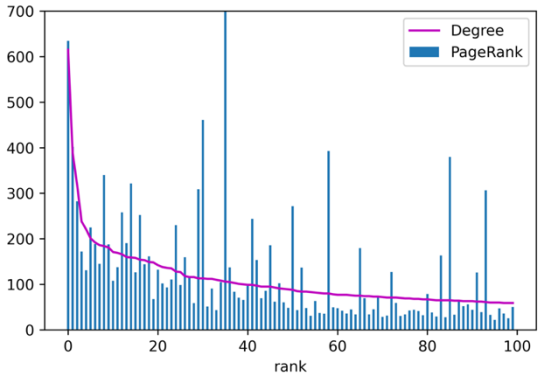


Figure 2 Comparison graph between D of the top 100 nodes and P

3.2.3 Influence Evaluation and Visualization

Using the given parameters, we calculate the data set and get the corresponding influence parameter (defined as I) representation as shown in Table 2.

Table 2 Artists' influence calculation

artist_id	artist_name	artist_main_genre	artist_active_start	influence
74	Special Duties	Pop/Rock	1980	0.000047
335	PJ Harvey	Pop/Rock	1990	0.000156
441	P.O.D.	Pop/Rock	1990	0.000069
589	Tony Furtado	Pop/Rock	1980	0.000047
1097	ToÃ±o Rosario	Latin	1990	0.000051
...	...	...	...	...
3639618	Jaira Burns	Pop/Rock	2010	0.000047
3659356	Elohim	Pop/Rock	2010	0.000047
3661296	Mika	Electronic	2010	0.000047
3661738	Rosemary & Garlic	Pop/Rock	2010	0.000047
3670556	Trinidad Cardona	R&B;	2010	0.000047

5606 rows × 4 columns

Then the network visualization results can be presented (see Figure 3). Take $N = 30$ as an example.

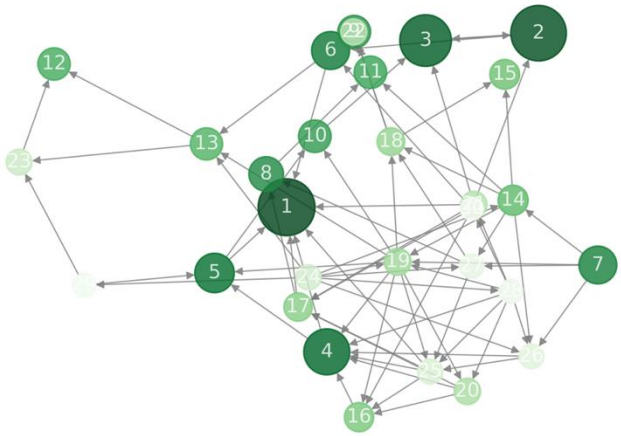


Figure 3 Visualization results of the influence network of the top 30 artists

3.3 Identification of the Genre Network

Assume that the inner factors within the genre will not affect the overall network. Based on the given parameters (D, P and I), we create a directed graph with the weight of the number of influencers in the genre as the edge and then draw the influence pie chart (see Figure 4) as well as the visualizing form of the genre network (see Figure 5).

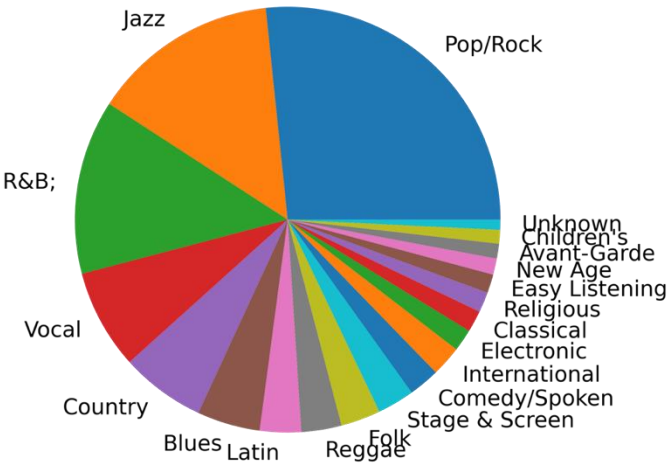


Figure 4 Influence pie chart of genres

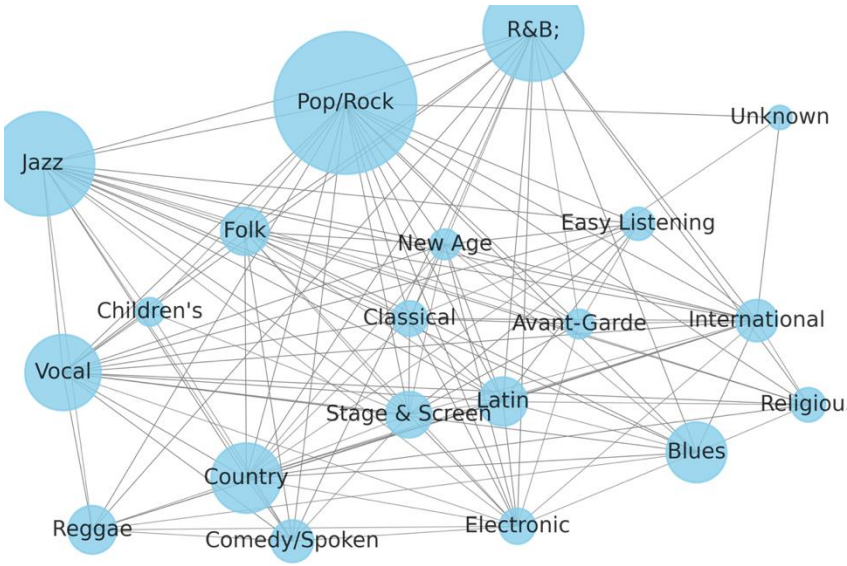


Figure 5 Visualizing form of the genre network (area represents the size of influence)

3.4 Community Division

3.4.1 K clique

K-clique is a subset of $G = (V, E)$. Each node in k-clique is connected to other nodes in pairs,

and k represents the size of the clique, which is the number of nodes in the clique that needs to be extracted. Moreover, k-clique-communities is an advanced k-clique sequence. If two k-cliques share k-1 nodes, it is said to be an advanced k-clique. A string of cliques adjacent to each other constitute the largest set, which can be called a community. Here we take $k = 6$, and 122 communities are separated.

3.4.2 Greedy modularity communities

Find communities in graph using Clauset-Newman-Moore greedy modularity maximization. This method currently supports the Graph class and does not consider edge weights. Greedy modularity maximization begins with each node in its own community and joins the pair of communities that most increases modularity until no such pair exists. Here we use the greedy algorithm to deal with the G and 29 communities are separated.

3.5 A Case Study of a Subnetwork

We choose the 10th community from the community division, which has 208 members and is relatively large. By counting the number of members belong to each genre, we found that 206 members are Pop/Rock musician which indicates that this community may be a representative of the influence network of Pop/Rock musicians, which is corresponding to the prosperity of Pop/Rock music between 1950-1960. According to the influence parameter provided by PageRank, we find that musicians like The Beatles, Bob Dylan are the core members in the community, which corresponds to their essential role in the development of Pop/Rock music.

The above analysis shows the power of analyzing subnetworks to explore the influence in the evolution of genres in a more direct and quick way. Other features of this community are a higher mean fluence parameter and a larger standard deviation comparing with other communities, suggesting the significant variety and influence power of Pop/Rock music.

Table 3 The picked subnetwork artist_data

artist_id	artist_name	artist_main_genre	artist_active_start	influence
754032	The Beatles	Pop/Rock	1960	0.005951
66915	Bob Dylan	Pop/Rock	1960	0.003720
894465	The Rolling Stones	Pop/Rock	1960	0.003023
531986	David Bowie	Pop/Rock	1960	0.002427
354105	Jimi Hendrix	Pop/Rock	1960	0.002113

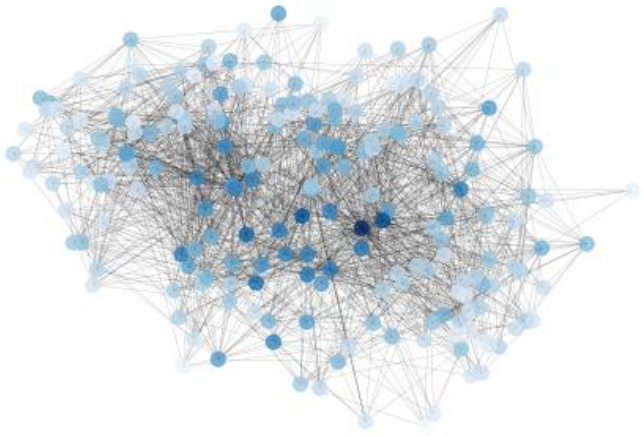


Figure 6 Representing chart of subnetwork

IV. Music Similarity Measurements

4.1 Problem Analysis and Methodology

To get a better understanding the relations between different genres, we need to develop certain measurements to evaluate the similarity between songs or artists. Based on the given data set which is the data_by_artist.csv, we first preprocess the given data, then use Principal Component Analysis (PCA) to reduce the dimensionality and establish a specific metric. At the same time, we mainly use the Analytic Hierarchy Process (AHP) model to analyze and judge the nature of each music.

4.2 Data Preprocessing

4.2.1 Data Cleaning

We first use data cleaning methods to fix or remove the incorrect, corrupted, duplicate, or incomplete data within the dataset. For outliers which refer to a set of measured values whose deviation from the average value exceeds two standard deviations, there are two main ways to deal with them: (1) Use the average to replace outliers; (2) Treat outliers as missing values and use the method of processing missing values (model calculation estimation).

4.2.2 Data Normalization

4.2.2.1 Test of Normality

Some statistical methods are only suitable for normally distributed or approximately normally distributed data. Therefore, before using these methods, normality test of the data is required. Kolmogorov-Smirnov Test (KS-Test) is a test method that compares a frequency distribution (defined as $f(x)$) with a theoretical distribution (defined as $g(x)$) or the distribution of two observations. The original hypothesis H_0: the two data distributions are consistent, or the data conform to the theoretical distribution. Define D as the actual observation value, and $D = max|f(x) - g(x)|$. When the actual observation value D > D (n, α) (n refers to the number of samples and α is significance level), H_0 is rejected.

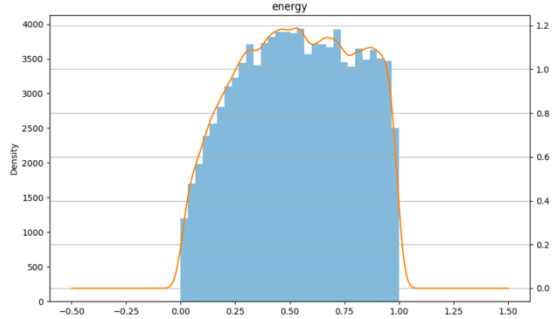


Figure 7 Data graph before normalization

We conducted KS-test on all attributes, and the results showed that the data normality of each attribute is very poor. In addition, there are some data like "mode" and "explicit" conforming to 0-1 distribution. Take the distribution diagram of energy attribution as an example (see Figure 7).

4.2.2.2 Normalization

Based on the fact that the normality of the data is very poor, we choose the square root method from the data normalization method for normalization. We define the original data as X and the normalized data as X', then the equation between the two is satisfied as: $X' = \sqrt{X}$. Then we get the data graph after normalization (see Figure 8).

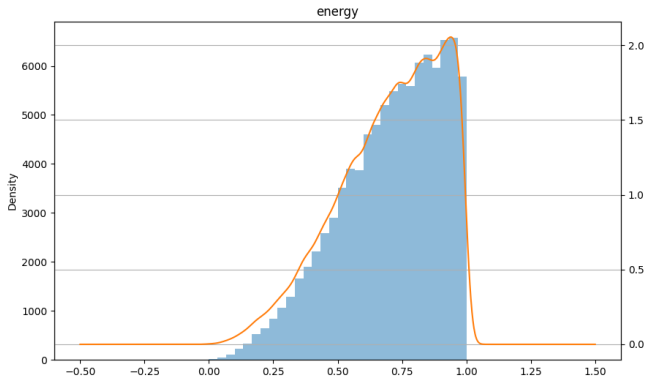


Figure 8 Data graph after normalization

4.2.3 Standardization

In a multi-index evaluation system, due to the nature of each evaluation index, it usually has different dimensions and orders of magnitude. Therefore, in order to ensure the results for reliability, we need to standardize the original data. The standardization of the data is to convert the number to a decimal between (0,1), and at the same time convert the dimensional expression to a dimensionless expression. Here we use Min-Max method to finish the process. Define x_i as the sample value and y_i as the transformed value, then we have:

$$y_i = \frac{x_i - \min\limits_{1 \le i \le n}\{x_j\}}{\max\limits_{1 \le i \le n}\{x_j\} - \min\limits_{1 \le i \le n}\{x_j\}}$$

4.3 Weight Assignment and Dimensionality Reduction

4.3.1 The AHP-PCA Model

4.3.1.1 AHP Model

The characteristic of AHP model is that on the basis of in-depth analysis of influencing factors and content of complex decision-making problems, it uses less quantitative information to

mathematicise the thinking process of decision-making, thereby providing simple methods. There are four main calculation steps in this model shown as follows.

- Build a hierarchical model

According to the offered data, the considering factors are 11 musical features. And the target layer is to analyze the weight of different features. According to recent studies[3] [4] [6] [7], genre (not included in our features) and emotion (Energy, Valence) are the two most important factors in playlist creation and management of music collections, followed by rhythm (Danceability, tempo), then acoustics (acousticness, instrumentalness, liveness and speechiness) and lastly melody (mode and key). Besides, (i). Since Danceability also contains some information about emotion, it ranks higher than tempo. (ii). Acounsticness and instrumentalness are more direct indicators of acoustics features, thus they rank higher. (iii). The feature loudness is hardly mentioned in relevant research; thus, it ranks the lowest.

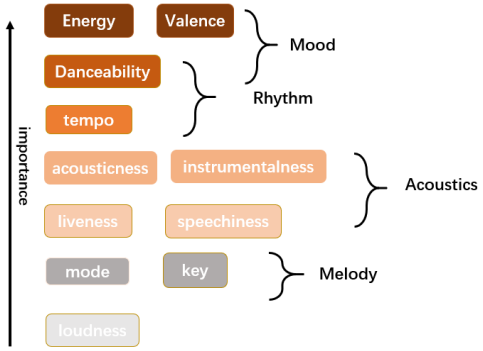


Figure 9 Rank of the importance of features

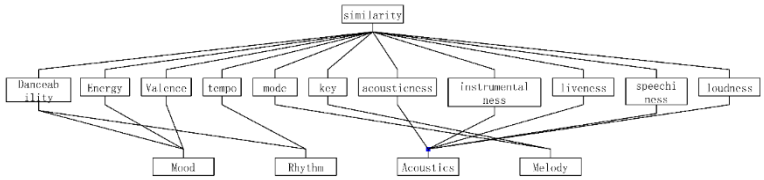


Figure 10 AHP model

- Construct a judgement matrix

The judgement matrix is constructed based on the rank of features and a value scale (1-Equal importance, 3-Somewhat more important, 5 -Much more important, 7-Very much more important, 9-Absolutely more important, 2,4,6,8-Intermediate values)

Our judgement matrix:

	Danceability	Energy	Valence	tempo	mode	key	acousticness	nstrumentalnes	liveness	speechiness	loudness
Danceability	1	1/2	1/2	2	5	5	3	3	4	4	6
Energy	2	1	1	3	6	6	4	4	5	5	7
Valence	2	1	1	3	6	6	4	4	5	5	7
tempo	1/2	1/3	1/3	1	4	4	2	2	3	3	5
mode	1/5	1/6	1/6	1/4	1	1	1/3	1/3	1/2	1/2	2
key	1/5	1/6	1/6	1/4	1	1	1/3	1/3	1/2	1/2	2
acousticness	1/3	1/4	1/4	1/2	3	3	1	1	2	2	4
instrumentalness	1/3	1/4	1/4	1/2	3	3	1	1	2	2	4
liveness	1/4	1/5	1/5	1/3	2	2	1/2	1/2	1	1	3
speechiness	1/4	1/5	1/5	1/3	2	2	1/2	1/2	1	1	3
loudness	1/6	1/7	1/7	1/5	1/2	1/2	1/4	1/4	1/3	1/3	1

Define A as the matrix, a_{ij} as the result of comparison of importance between element i and element j, $w = [w_1, w_2, w_3, \dots, w_{11}]^T$ where w is a sort vector and $a_{ij} = \frac{w_i}{w_j}$ (i, j = 1,2,⋯,11).

$$A = \begin{bmatrix} \frac{w_1}{w_1} & \cdots & \frac{w_1}{w_{11}} \\ \vdots & \ddots & \vdots \\ \frac{w_{11}}{w_1} & \cdots & \frac{w_{11}}{w_{11}} \end{bmatrix}$$

- Single-level ranking and consistency check

Define λ_{max} as the characteristic root, CI as the consistency index, RI as the mean random consistency index (a set of standard indicators generated by random methods), CR as the consistency ratio.

In our matrix, λ_{max} 11.2873

$$CI = \frac{\lambda_{max} - n}{n - 1} = 0.0287$$

CR: $CR = \frac{CI}{RI} = 0.0189, \quad CR < 0.1$, so the consistency of the matrix is acceptable.

- The result of AHP

Features	**Danceability**	**Energy**	**Valence**	**Tempo**	**Mode**	**Key**
Weight	0.1536	0.2215	0.2215	0.1047	0.0279	0.0279
Features	**acousticness**	**instrumentalness**	**liveness**	**speechiness**	**loudness**	
Weight	0.0685	0.0685	0.0434	0.0434	0.0191	

Assume $\boldsymbol{x}_w$ is the weighted feature information,

$$\boldsymbol{x}_w = \mathrm{w}\boldsymbol{x}_n$$

where w is the vector of 11 assigned weights for features.

4.3.1.2 PCA Model

PCA model is one of the most widely used data dimensionality reduction algorithms. The main idea of PCA is to map n-dimensional features to k-dimensions. In fact, this is equivalent to only retaining the dimensional features that contain most of the variance and ignoring the feature dimensions that contain almost zero variance, so as to achieve dimensionality reduction processing for data features.

Define S as the scatter matrix, m as the mean vector, $\boldsymbol{x}_k$ as the sample value, P as the eigenvector matrix, Y as the new vector space.

$$S = \sum_{k=1}^{n} (\boldsymbol{x}_k - \boldsymbol{m})(\boldsymbol{x}_k - \boldsymbol{m})^T$$

Where $\boldsymbol{m} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{x}_k$ (k = 1,2,⋯, n).

Now we implement the PCA algorithm based on the eigenvalue decomposition covariance matrix.

Step 1: De-average (decentralization) - subtract the average value from each feature.

Step 2: Calculate the covariance matrix $\frac{1}{n}XX^T$

Step 3: Sort the eigenvalues from large to small and select the largest k among them. Then the corresponding k eigenvectors are used as row vectors to form the eigenvector matrix P.

Step 4: Convert the data into a new space constructed by k feature vectors, that is, Y=PX.

By using PCA from sklearn in python, we can reduce our data to 3 dimensions, defined as PC1, PC2, PC3, the variance ratio of each is 0.47109915, 0.2374329, 0.12432578, covering 83.285783% information of our original data.

4.3.1.3 Metrix Development based on Cosine

Assume the vector of feature's values after dimension reduction is $\boldsymbol{p}_i$, then the similarity between two vectors $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ is defined as:

$$s = \frac{\sum_{k=1}^{3} p_{1k} p_{2k}}{\sqrt{\sum_{k=1}^{3} p_{1k}^2} \sqrt{\sum_{k=1}^{3} p_{2k}^2}}$$

In the end, s is the parameter we defined to measure similarity.

4.4 Analysis of the Performance of our measures

For the analysis of within genres, we randomly take the data of 30 artists from each faction (if there are no 30 in this category, use all) and combine them (C_{30}^2), use the obtained measure to calculate all similarity distances and get the average and variance results (see Figure 11). For the analysis of between genres, we randomly select 19 pairs of categories and randomly select 30 artists' data for each category, and then perform a full arrangement comparison (A_{30}^2), calculate all similarity distances to obtain the average value and variance results (see Figure 12).

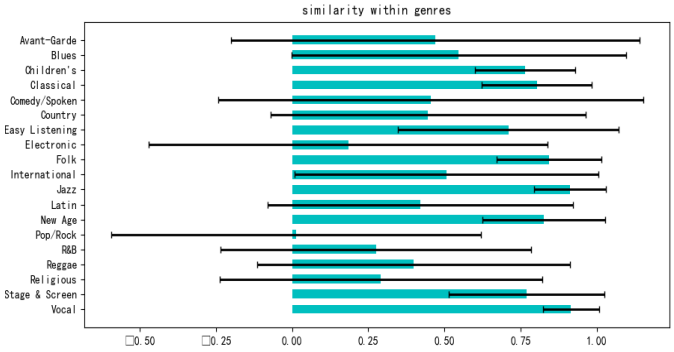


Figure 11 Similarity calculation within genres

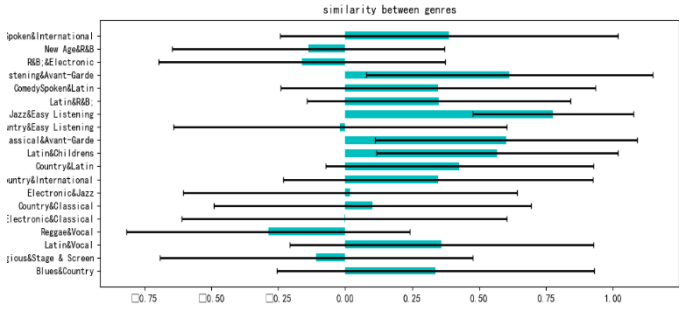


Figure 12 Similarity calculation between genres

Take some specific results for further illustration. For instance, within the Pop/Rock Genre, the similarity is quite low while the variance is quite high, indicating that the styles of the genre vary greatly. Moreover, according to the specific features of Pop/Rock music like richness, variability, fast spreading and so on, the analyzing result is consistent with the real musical styles. And the similar phenomenon may be seen within the Electronic Genre. Take Vocal Genre as another example. It's clear that within the genre, the results are just opposite to those of the Pop/Rock Genre, indicating that the styles of the genre own many similarities and the data is quite stable. Also take its features like the singing methods are relatively fixed, and there are fixed evaluation criteria into consideration, we can find the consistency between the obtained results and the real music style. From this perspective, we can also conclude that the selection of our models and metrics has a certain degree of credibility.

V. A Comprehensive Model Combining Influence Parameters and Similarity Measurements

5.1 Problem Analysis and Methodology

Based on the analysis above, we can see the importance of the influence parameter and similarity parameter in revealing the relationships between songs, artists, and genres. In this section, firstly we will have a closer look at the role of two parameters in genre classification, meanwhile answering Question 3. Next, we plan to identify and quantify the relation between the two parameters, meanwhile answering Question 4. Finally, we will try to combine the two parameters based on their relationships and their unique characteristics to build a comprehensive model for genre classification

5.2 The Relation between Influence and Similarity

5.2.1 Similarity between Influencers and Followers

In the fourth part we established a metric to measure the similarity between different songs and different singers, and in the third part we also established an influence parameter. Then, based on all the influence_data we obtained in the third part, we use the established similarity measurement method for each pair of artists to obtain the similarity measurement, and then average the obtained similarity data. At the same time, we randomly select several pairs of unrelated artist data, and then take the average of the similarity measures. The two average values obtained are then compared to determine whether the similarity of works between influencers and followers will be higher than that of artists without certain correlation and whether the influencers actually affect the music created by the followers.

To be more specific, we calculate the average value of “influence” data set and “not_influence” data set and find the average value of the former set (0.43088907) is much higher than the value of the latter set (0.01300292), indicating that the similarity between influencers and followers weigh many times more than the similarity between unrelated pairs.

5.2.2 Contagious Characteristics

After getting the correlation between influence and relevance, we further explore the appeal of various specific characteristics of music. We subtract the attribute data of all associated paired artists and music and sorted the difference values obtained to distinguish the appeal between different attributes (see Figure 13). And it is clearly that “Key” (the most contagious one), “Instrumentalness”, “Acousticness” and “Mode” are main contagious characteristics.

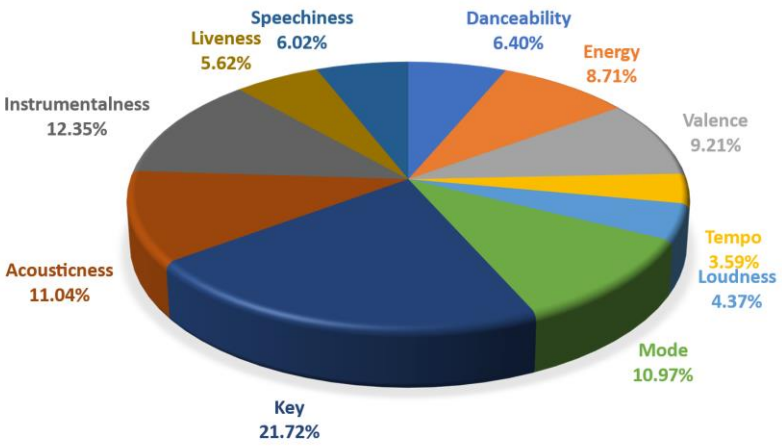


Figure 13 Pie chart of musical characteristics

5.3 The Role of Influence and Similarity Parameters in Genre Classification

5.3.1 Influence between and within genres

To qualitatively assess the relation between influence parameters and genre classification, we counted the number of influence pairs within and between genres, and found out that among 44270 influence pairs, 34253 (77.4%) happened within a genre, suggesting that if two artists are connected by the relationship of influencers and followers, they are more likely to be in the same genre.

5.3.2 Developing a measure for classification based on similarity of characteristics

By comparing the difference value of each feature between two genres, we can easily identify the significant characteristics to distinguish them. For example, instrumentalness plays a big part in distinguishing Blues and Avant-Grade.

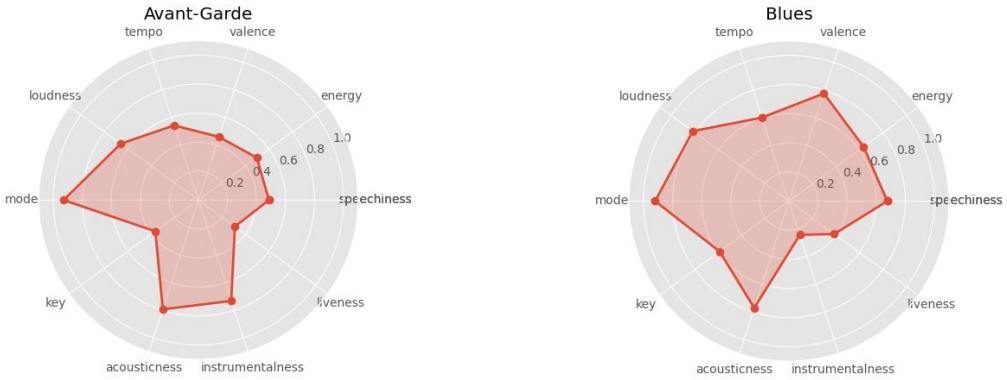


Figure 14 Instrumentalness works in distinguishing genres

However, it's hard to simply use one or two of the features to classify all the genres, therefore we need to develop a measure for classification based on the similarity parameter. Using the vector p we mention in section 4, we can plot each song/artist in a three-dimensional coordinate and use Dbscan to cluster them into genres. However, as the picture shown below, though some genres can be clustered and be distinguished from others, most songs/artists from different genres are gathering together and are not able to be distinguished into a particular genre.

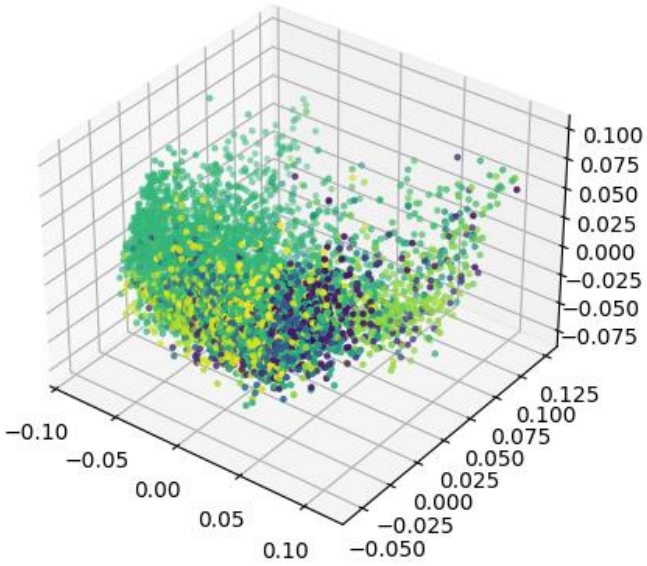


Figure 15 Genres clustering

It's not surprising that our measure doesn't work in genre classification. First, we had assigned the weight for each feature based on literature research roughly, ignoring that some features may not be significant in similarity perception, yet play an important role in genre classification. Besides, we had reduced the dimensions from 11 to 3, making it harder for points to disperse. What's more, we had used a small portion of the data with 11 dimensions and genre labels to train a decision tree model, and what we get was a tree with much more leaf nodes than the number of genres. For example, almost all subtrees have a leaf node classified to genre pop/rock, which indicates that it's hard to distinguish artists/songs between some genres.

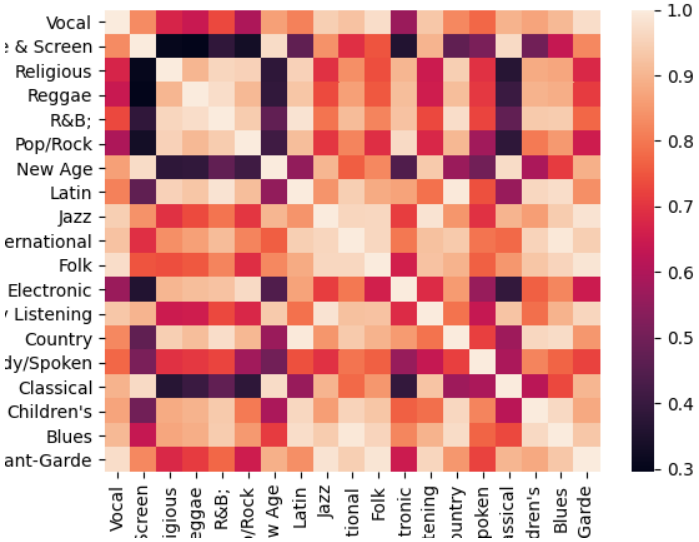


Figure 16 Similarity parameter calculating results

Though the direct clustering for songs/artists is hard, we can consider the items in a genre as a whole and assess the similarity between genres. First, we calculate the mean value for each feature for all songs/artists in one genre, then follow the steps in section 4 to calculate similarity parameter between genres. The results are shown in Figure16.

To show the similarity between genres in a more obvious way, we use PCA to reduce the dimensions of the data to 3 and 2 and plot each genre. From Figure 17 we can see that a few genres are significantly different from others, however, most genres are aggregated, which again shows that there are no obvious boundaries between these genres based on the features given.

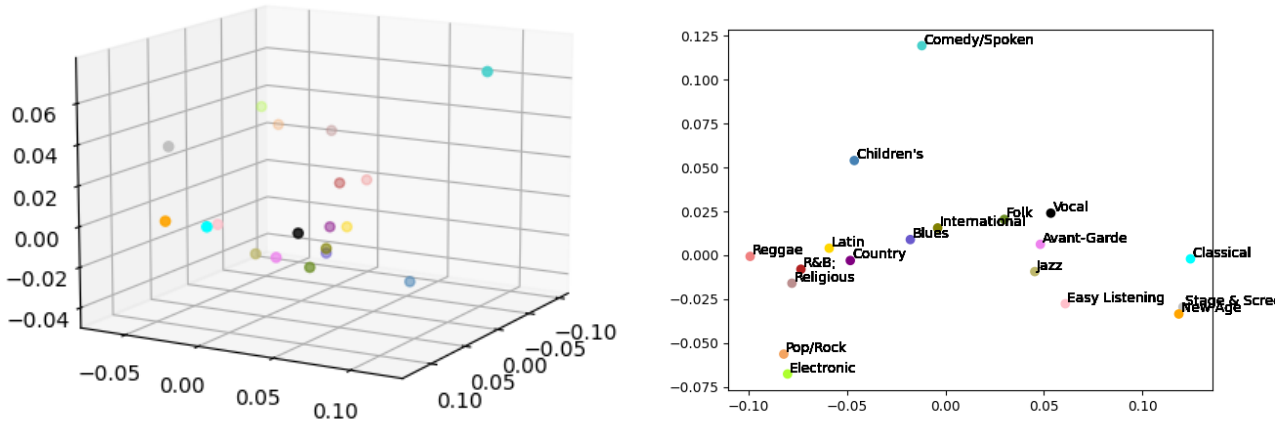


Figure 17 Genres division after dimension reduction

VI. The Evolution of Genres

6.1 Problem Analysis and Methodology

Taking time into consideration, we may get a better understanding of the evolutionary process of the period. In this section, firstly we will analyze the proportion of the number of artists of each genre in different eras since when the increase in the number of people in a certain genre in a certain year reaches a certain threshold, it may represent the evolution of the genre. Meanwhile, we will filter out the artists of the genre in that year, sort them by the influence parameters obtained before, and analyze the representative figures so as to answer Question 5. Secondly, we will develop three indicators for major leaps in musical evolution and analyze the overall changing trends of the genres so as to answer Question 6. Finally, we will take cultural factors into consideration and develop specific methods to weigh the influence of these impactors while combining the social events to present some explanation of the major leaps so as to answer Question 7 and make further exploration as well.

6.2 Indicators for Major Leaps in Musical Evolution

6.2.1 Proportion of Artists and Genre Evolution Analysis

After finishing the data cleaning which mainly deletes the data of unknown genres, we obtain the table of different numbers of artists of each genre in each era. Then based on the given data, we draw the overall figure of the numbers as shown in Figure 18.

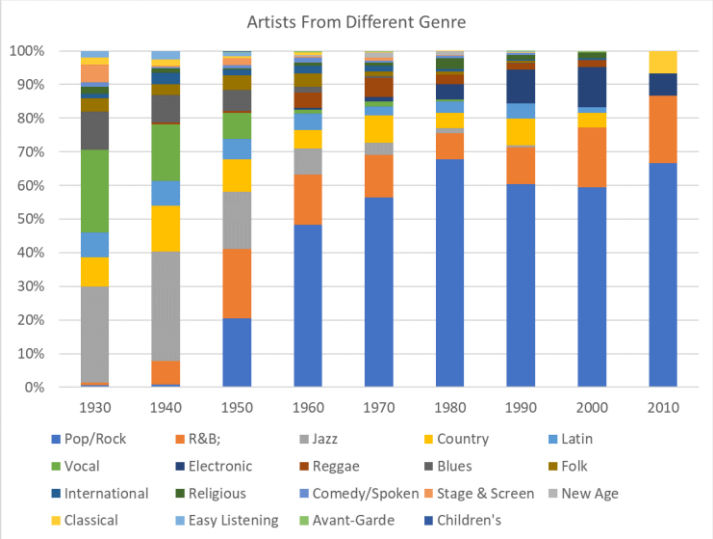


Figure 18 Overall view of different numbers of artists of each genre in each era

Next, based on the obtained parameters, we then measure the relative rate of changing of the genres as shown in Table 3 and visualize the data as shown in Figure 18. To be more concrete, as we can see clearly from the figure, Pop/Rock music has a rapid growth in the 1950s, meanwhile Electronic genre appears in the 1960s and has a stable growing speed and R&B genre grows fast in the 1940s while have some small fluctuations then. At the same time, we can also find that the declining trends can be seen mainly among the genres of Jazz, Vocal, Blues and Folks.

Table 4 Relative rate of changing of the genres

	1940	1950	1960	1970	1980	1990	2000	2010
Pop/Rock	30.43%	2262.03%	134.91%	16.96%	19.95%	-10.83%	-1.49%	12.12%
R&B;	943.48%	195.25%	-26.75%	-15.50%	-37.90%	40.51%	60.80%	12.12%
Jazz	13.75%	-47.83%	-55.02%	-53.40%	-59.00%	-66.53%	-100.00%	0.00%
Country	55.52%	-27.65%	-42.80%	44.55%	-43.78%	76.37%	-45.90%	-100.00%
Latin	0.79%	-18.60%	-20.24%	-45.08%	27.58%	30.99%	-63.18%	-100.00%
Vocal	-31.26%	-54.73%	-84.79%	32.82%	-62.28%	-72.10%	-100.00%	0.00%
Electronic	0.00%	0.00%	0.00%	168.95%	224.81%	119.57%	19.50%	-43.94%
Reggae	0.00%	43.15%	650.19%	19.53%	-50.23%	-29.53%	10.45%	-100.00%
Blues	-27.11%	-24.66%	-72.91%	-63.22%	-76.43%	11.58%	-100.00%	0.00%
Folk	-23.91%	43.15%	-10.69%	-64.14%	-37.13%	-44.21%	-100.00%	0.00%
International	160.87%	-52.28%	40.66%	-26.95%	-57.14%	-33.05%	10.45%	-100.00%
Religious	-34.78%	-68.19%	118.81%	2.46%	245.76%	-59.42%	24.26%	-100.00%
Comedy/Spoken	-67.39%	138.59%	50.04%	-70.12%	88.60%	-44.21%	-100.00%	0.00%
Stage & Screen	-91.85%	377.18%	-81.25%	139.07%	-84.28%	-100.00%	0.00%	0.00%
New Age	0.00%	0.00%	0.00%	298.45%	-33.99%	-84.06%	-100.00%	0.00%
Classical	-13.04%	-76.14%	56.29%	-76.09%	-5.70%	11.58%	-100.00%	0.00%
Easy Listening	30.43%	-52.28%	-89.58%	-100.00%	0.00%	0.00%	-100.00%	0.00%
Avant-Garde	0.00%	0.00%	87.55%	-20.31%	-100.00%	0.00%	231.35%	-100.00%
Children's	0.00%	0.00%	-100.00%	0.00%	0.00%	0.00%	0.00%	0.00%

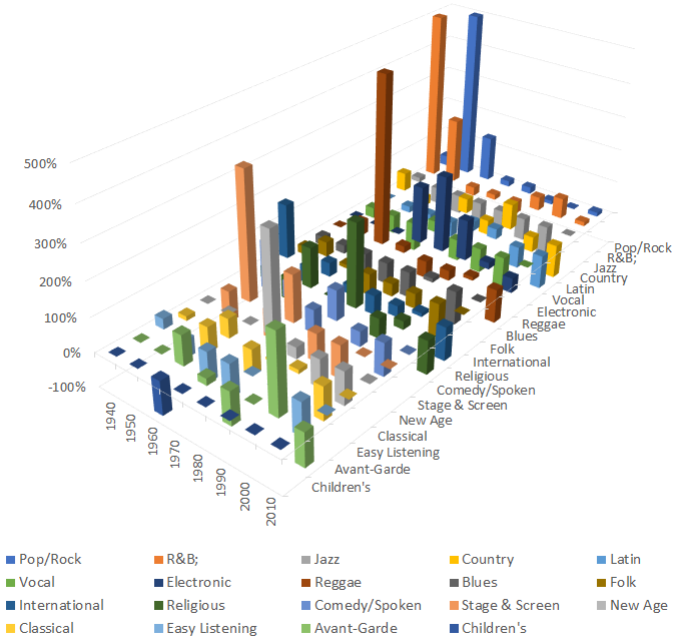


Figure 19 Changing trends of the genres

6.2.2 Feature Parameters

By using PCA to reduce the dimension of feature values to 1, we obtain a parameter F as an indicator of major leaps in the content of music. Take the dynamic process of the development of Pop/Rock music as an example, we can see that this indicator has a fair correlation with the indicator developed by the portion of the number of artists. The following picture shows that both indicators can indicate some key years when major leaps occur, like the prosperity of Pop/Rock in 1950-1960.

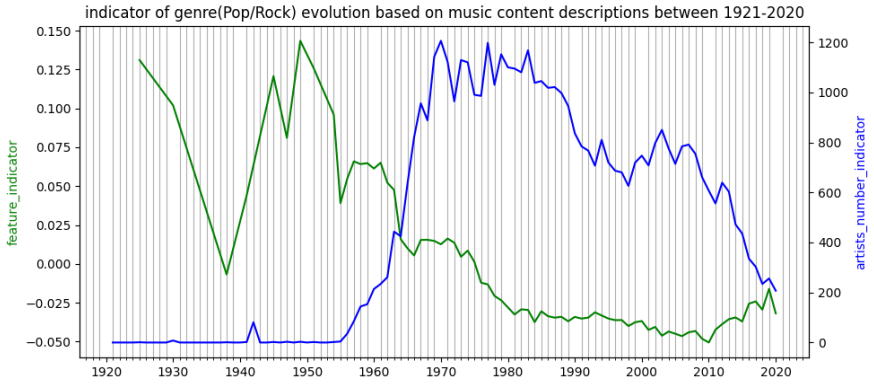
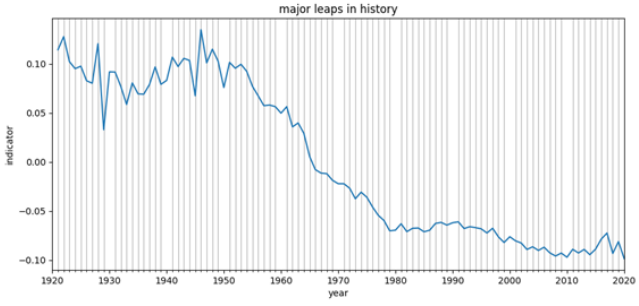


Figure 20 Indicators of genre evolution between 1921-2020

The feature parameter can also be used to analysis the overall change in the music content. We can see from the picture below that the major leap of the overall style happened around 1950 and became steadier from 1980 till now comparing with the fluctuation in 1920-1950. This leap corresponds with the prosperity of Pop/Rock and the decline of Jazz/Vocal/Blues. The trend also suggests that music nowadays is more homogeneous, which is a bit different from our general



perception.

Figure 21 Major leaps shown by the indicator

6.2.4 Representative Artists in Major Leaps and Evolution Analysis

In this section, we filter the data set of artists.csv to point out the representative artists in major leaps in both evolutionary genres and decline genres. We evaluate the obtained influence parameter of the artists in each genre and come up with following results.

In growing genres:

- Pop/Rock Music

  The top five artists are The Beatles. Bob Dylan, Chuck Berry, The Rolling Stones and Little Richard, and their further information are shown in Table 5.

Table 5 The representative artists of Pop/Rock Music

artist_id	artist_name	artist_main_genre	artist_active_start	influence
754032	The Beatles	Pop/Rock	1960	0.009067
66915	Bob Dylan	Pop/Rock	1960	0.005737
120521	Chuck Berry	Pop/Rock	1950	0.004590
894465	The Rolling Stones	Pop/Rock	1960	0.004035
824022	Little Richard	Pop/Rock	1950	0.003880

- Electronic Genre

  The top five artists' information are shown in Table 6.

Table 6 The representative artists of Electronic Genre

artist_id	artist_name	artist_main_genre	artist_active_start	influence
104714	Kraftwerk	Electronic	1970	0.001498
2411	Tangerine Dream	Electronic	1960	0.000461
387815	Neu!	Electronic	1970	0.000401
793821	Frankie Knuckles	Electronic	1970	0.000290
683750	Yellow Magic Orchestra	Electronic	1970	0.000210

- R&B Genre

  The top five artists' information are shown in Table 7.

Table 7 The representative artists of R&B Genre

artist_id	artist_name	artist_main_genre	artist_active_start	influence
46861	Ray Charles	R&B;	1940	0.004408
343396	Roy Brown	R&B;	1940	0.003988
128099	James Brown	R&B;	1950	0.003604
238115	Sam Cooke	R&B;	1950	0.003284
960674	Wynonie Harris	R&B;	1940	0.002809

In declining genres:

It can be seen that most of the artists listed are from early times and mostly concentrated in the 1930s. Compared to later generations of artists, their influence factor is greater.

Table 8 The representative artists of Jazz/Vocal/Blues

artist_id	artist_name	artist_main_genre	artist_active_start	influence
532957	Cab Calloway	Jazz	1930	0.020520
79016	Billie Holiday	Vocal	1930	0.019570
259529	Lester Young	Jazz	1930	0.016920
287604	Louis Jordan	Jazz	1930	0.013418
3829	T-Bone Walker	Blues	1930	0.009863
403120	The Mills Brothers	Vocal	1930	0.007606
805930	Charlie Christian	Jazz	1930	0.006717
898336	Mississippi Sheiks	Blues	1930	0.006687
898331	Mississippi Fred McDowell	Blues	1950	0.006649
608701	Muddy Waters	Blues	1940	0.006586

6.3 The Decline of Vocal – A Case of Study

Then we focus on the followers of these artists and make aggregation analysis (see Table 8). And the result shows that Por/Rock and R&B genres surprisingly own the highest proportion, which can probably present the evolutionary relations between growing genres and declining genres. In other words, the decline of some genres may be due to the reason that many followers transform their styles to other genres.

Table 9 The aggregation analysis

follower_main_genre	count
Pop/Rock	145
R&B;	75
Folk	11
Country	4
International	4
Reggae	4
Religious	3
Comedy/Spoken	2
Latin	2
Classical	1
Stage & Screen	1

And from the obtained data, we can clearly see that from 1940 to 1960, Vocal genre keeps declining and in the 1960s the decline rate reaches 84.79%. According to a survey on music development in the 1960s, we know that there is a period called “British Invasion”, during which many British rock bands and pop artists gained great success worldwide. Meanwhile, various kinds of music festivals which mainly consist of rock music and pop music flourished and exert a huge influence on the music styles of the period, which may lead to Vocal genre’s declination.

Then when it comes to the 1970s, a trend of relaxing music as well as dance music prevailed, which is quite an apparent change from the 1970s, which may account for the only growth of Vocal

genre. And in the 1980s, with the popularity of MTV (Music Television), pop music reached a tragic climax and many new genres like Hip Hop and New Wave appeared, which did challenge the development of Vocal music. And from the 1980s to the 2010s, the development of multiple music under the background of diversification and high-tech era has also challenged its own advantages again.

6.4 Identification of Cultural Influence in Major Leaps

Based on our existing parameters and based on the changing trends of the parameters in the time dimension, if the influence trend of human factors is different from the changing trend of the nature of music, it indicates that social factors have played an important role. Therefore, in addition to giving explanations for the trend and possible causes based on our model, we also need to develop a method to assess the impact of social factors. At the same time, because human factors and social factors influence each other, we also need to further understand the specific relationship between these two parts.

For example, from the analysis above, Electronic music emerged around 1960. We pick out the influencers of musicians in Electronic and find that Pop/Rock musicians accounts for a big part (372/504). However, the followers of Pop/Rock influencers are mainly Pop/Rock musicians as mentioned above. Thus, there should other factors leading to the emergence of Electronic music besides the influencers.

In fact, since the late 1960s, Moore's Law has promoted the development of electronic instruments and electronic music technology, which serves as a catalyst for the emergence of Electronic music.

VIII. Strengths and Weaknesses

8.1 Strengths

- We’ve made full advantage of the given data and created convincing models as well as parameters to quantify the relations within and between different data sets.
- We develop two indicators for detecting the major leaps in music evolution, and prove that they have a fair correlation, which enhances the credibility of our indicators.
- Based on the analysis, we make further analysis of the evolutionary trends and combine social events with the obtained results.
- We take the difference of the importance of features in people's perception of similarity into consideration
- We obtained visualized figures so as to present the mathematical relations in a direct way.

8.2 Weaknesses

- The mathematical methods we use have their own limitations. For instance, use degree to evaluate the importance of nodes is hard to obtain a good command of the overall structure of the network and PageRank ignores the relations between genres and does not take the time factor into account.
- The theoretical trends we obtained from the analysis are actually influenced by social impacting factors and the assumption we made underestimate their influences.

IX. One-page Document

When dealing with a dynamic process including multiple factors, it's often hard to grasp the major elements in the process of analysis. When tracing the cause of major leaps in the music evolution, we focus on people and their relations. In our influence network model, we quantify the power of influence by developing influence parameter, which helps us to find artists who play a major role in the process as the start point of our analysis. Secondly, by applying k_clique algorithm and Clauset-Newman-Moore greedy modularity maximization, it's easier and quicker for us to extract the subnetwork and identify the targets for further assessment. By combing the similarity parameter, which is examined to be positive related to the influence parameter in our model, the information about the power of influence on followers is greatly enriched, thus improving the accuracy of the information and enable us to do further calculation, like calculating the influence power per follower for an influencer and use it as another indicator for identifying important influencers.

Cultural factors, including social, political and technological changes can also have a big impact on the evolution of music. In this problem, the given data mainly contains information about musical contents and the connection between people. If more data, such as the features of the social environment of a particular period, the features of the influencers and followers, the technologies used to record and transmit music, we may apply some machine learning models to predict the influence between two artists, based on other factors, thus predict the evolution of music. Also, as the given data is mainly numeric type, if data like pictures, videos or music itself is given, we may try new methods to gain more features, such as the pitch structure, the arrange of notes in a song. As the dimension of the information increases, better methods for dimension reduction is needed. We will spend more time analyzing the relations between different parameters and choose or combine them to a more informative one.

Besides, as more relations are being analyzed, the network model will become more complex. So a better method for community division and subnetwork extraction is needed to make sure that we focus on the most important nodes.

X. References

[1] Leighton, F. Thomson. Introduction to parallel algorithms and architectures: array, trees, hypercubes. 2014.

[2] Ramin Gharizadeh Beiragh, et al. An integrated Multi-Criteria Decision Making Model for Sustainability Performance Assessment for Insurance Companies. *Sustainability* 2020, 12(3):789.

[3] Pei-I Chen, et al. Personal Factors in Music Preference and Similarity: User Study on the Role of Personality Traits. http://mac.citi.sinica.edu.tw/

[4] Bob L.Sturm. Classification accuracy is not enough: On the evaluation of music genre recognition systems. *Journal of Intelligent Information Systems,* 41, 371-406(2013).

[5] Isabelle Guyon, et al. An introduction to variable and feature selection. *The Journal of Machine Learning Research,* 2003.

[6] Yu-Lung Lo, et al. Content-based music classification. *International Conference on Computer Science and Information Technology,*2010.

[7] Maria Panteli, et al. On the evaluation of rhythmic and melodic descriptors for music similarity. *Proceedings of the 17th ISMIR Conference,* New York City, USA, August 7-11, 2016.