Problem Chosen	2021	Team Control Number
D	MCM/ICM Summary Sheet	**2102330**

Music Evolution: Analysis of Network based on Influence and Similarity

Summary

Music is one of the most essential parts of the spiritual life of human society. It is also a complex target of scientific inquiry. Directed networks are constructed and a series of algorithms are applied, so that we can describe the characteristics and interrelation of both musicians and genres. Most importantly, their role in promoting the music evolution is also measured.

In Task I, we first establish and visualize the directed network. Then, **PageRank Algorithm** is applied to obtain $PR\ value$ as a measure of the artist influence. A subnetwork is established based on partial data, indicating that the ranking of high-impact musicians is basically unchanged over time, but its influence decreases with time due to the enrichment of music.

In Task II, we construct a weighted undirected network in terms of Whole-Feature-Based Similarity (WFBS) using **Mahalanobis Distance**. Furthermore, **Infomap Algorithm** based on Huffman coding and random walk is utilized to obtain 209 communities. It can be concluded that the same genre is not the inevitable condition of high similarity between musicians.

In Task III, we extend the influence and similarity of individual gained in Task I&II to the overall genres. Then, **Linear Discriminant Analysis (LDA)** is used to obtain genre distinguishment and change trend over time. Hence, we accurately explore the merging trend of some specific genres.

In Task IV, we respectively calculate Influence-Based Similarity (IBS) by **Centralised Cosine Measure** and Single-Feature-Based Similarity (SFBS). On the basis of **Kullback-Leibler (KL) Divergence**, we discover the consistency between similarity and influence of artists. Meanwhile, $speechiness$, $loudness$, and $tempo$ are more contagious than others.

In Task V, we speculate the possible time for musical revolution (1928, 1946, 1979, and 2009) by **Moving Average Method**. **Kernel Density Estimation** is applied to obtain the main factors that reflect the music revolution, among which $valence$ is the most significant one. Thus, a $valence$-based similarity network is constructed to identify the musicians signifying music revolution.

In task VI, we choose a specific genre $Electronic$ and observe how exactly the features investigated before changed over time. Then, we describe the similarity of their tendency by correlation coefficients and discover that features inside one genre may have a different performance. In addition, a simple dynamic influencer detection method is developed based on popularity of artists in a given period.

In Task VII, we point out how our work shows that cultural, social and religious changes have influence on the music evolution. These external influences on music may lead to changes in music evolution.

To sum up, we consider the extension and future improvement of our model. Additionally, we analyze the strengths and weaknesses of our model.

Keywords: Music influence; Network; PageRank Algorithm; LDA; KL Divergence; KDE

Contents

1 Introduction

1.1 Background

Music is one of the most essential parts of the spiritual life in human society. It is also a complex and fascinating target of scientific inquiry. In the early days, music can only be measured by subjective perception. However, with advances in the science of data and information, data on the characteristics of musicians and their works, collected with modern technology, provides the basis for a more accurate assessment of the influence and interrelationships. Therefore, using these data to build mathematical models to analyze it more accurately has become a key link in the technical analysis of music evaluation. At the same time, the analysis of the quantitative measurement of music revolution is also helpful for us to analyze the teamwork in many other aspects of art.

A number of researchers have previously contributed to analysis for music revolution. Typically, these authors use descriptive narratives and music manuscripts analyses, which obviously limit the exploration of the further development [1]. In today's world, **Network Science** has been investigated and utilized in solving music problems. The study of Georges [2] proposes a statistical analysis that captures similarities and differences between classical music composers applying the centralised cosine measure as an index of association. Zadel and Fujinaga [3] combine cultural meta-data with a metric of similarity to generate a network of related artists. Georges and Nguyen [4] applies clustering techniques and multi-dimensional scaling analysis to a composers' similarity matrix, in order to visualize the similarity into maps of classical music composers.

Unfortunately, most of these researches are best to be seen as preliminary background, since there is a need to improve this framework using a finer analysis [5]. This paper is a complement to these approaches by proposing a network analysis that captures similarity, influence and characteristics.

1.2 Restatement of the Problem

The mathematical models we build should finish seven tasks:

- Construct the directed network and its subnetwork of music influence and determine the specific index.
- Establish the measurement index of music similarity and judge whether the same genre is the decisive parameter affecting similarity of music.
- Establish the similarity and influence relationship between genres, to explore their differences, correlations, and the changes with time respectively.
- Determine whether similarity is positively related to influence, and analyze the ways in which highly affected musicians influence the specific features of their followers.
- Identify the determinant characteristics and musicians that signify music revolution.
- Investigate the changes of features selected over time in a certain genre and propose a simple method to reveal the dynamic influencers.
- Explain how our work shows the extenal influence on the music.

2 Assumptions and Notations

2.1 Assumptions

- Suppose that 80/20 Rule can be followed in the study of music, which means the key few often determine the results of the entire research.
- The effect of the artists before 1930 on the following artists can be ignored due to the lack of the data before 1930.

2.2 Notations

Table 1: Notations

Symbols	Definition
$degree$	The number of nodes linked to the node.
$PR\ value$	The influence of the artist calculated by PageRank algorithm.
P	Similarity measure produced by IBS
Q	Similarity measure produced by WFBS
q	Similarity measure produced by SFBS

3 Task I: Directed Network of Music Influence

This section we first construct a complete directed network on the basis of the whole data in the set $influence\_data$, and apply the PageRank algorithm to acquire $PR\ value$ as the main measure of musician influence. Moreover, we select some of the data according to the year index to build subnetworks, on the purpose of exploring and describing their interactions.

3.1 Directed Network

3.1.1 Construction of Network

In this section, we create a directed network for influence between artists. Using this directed network, we can not only identify common interactions, but also find out the classical music evaluation indexes and present these results visually.

The nodes in the network represent the musicians and influencers are connected to followers. The depth of the node color represents the $degree$ **(the number of nodes linked to the node)** in the directed network. The larger the $degree$ is, the darker the node is.

3.1.2 PageRank Algorithm

That an artist is influential in the music world means the node corresponding to him/her is more crucial in the network. Thus, we use the **PageRank algorithm** [6] to measure the significance of a certain node, which is the basis for google to determine the importance of this page and to rank web pages. The basic idea is that the importance of a page depends on the number and quality of the other pages that point to it. Details are shown in the Alorithm1.

And we take **PageRank value (PR value: the influence of the artist)** to capture its $music\ influence$ in this network. The size of a node circle depends on its $PR\ valu$e of nodes.

The larger the $PR\ value$ is, the larger the circle is. $PRvalue$ ranges from 0 to 1 and its value is positively related to the importance of the musician.

Algorithm 1 PageRank Algorithm

1: $\mathbf{Initialstep}$: Given the initial $PR\ value$ of all nodes $PR_i(0),\ i = 1, 2, \ldots, N$, which satisfies

$$\sum_{i=1}^{N} PR_i(0) = 1. \tag{1}$$

2: $\mathbf{The\ basic\ correction\ rule}$: If node i has an output of k_i^{out} , then each node pointed to i is assigned. If a node i has an $out - degree$ of k_i^{out} , then the $PR\ value$ is $PR_i(k-1)/k_{i'}^{\text{out}}$ The new PR value is calculated as following.

$$PR_i(k) = \sum_{j=1}^{N} a_j \frac{PR_j(k-1)}{k_j^{\text{out}}}, \quad i = 1, 2, \ldots, . \tag{2}$$

3: $\mathbf{Revised\ regularization}$: give a scaling constant $s \in (0, 1)$. $PR\ value$ is calculated according to (2) and the sum of $PR\ values$ is reduced to s. Then divide the $1 - s$ to keep the total $PR\ value$ 1. Then the formula can be constructed.

$$PR_i(k) = s \sum_{j=1}^{N} \bar{a}_{ji} PR_j(k-1) + (1-s)\frac{1}{N} \quad i = 12\ldots N \tag{3}$$

By using the above pagerank algorithm, we get the $PR\ value$ of all the musicians and sort them successfully. Therefore, we get the influence ranking result according to the $PR\ values$ and the top 10 musicians and their $PR\ value$ are listed in the table 2.

3.1.3 Result

$PR\ value$ is positively correlated with the $degree$ of node as shown in Figure 1, but some differences still exist. For instance, $Cab\ Calloway$, as the musician with the first $PR\ value$ ranking (the most influential musician), does not own the largest $degree$ in the initial directed network. In order to explore this phenomenon, we build a sub-influence network that is oriented by $Cab\ Calloway$, first we find the artists directly influenced by him, such as $Billy\ Eckstine$, $Frank\ Sinatra$, $T - Bone\ Walker$, etc., and then dissect the artists influenced by them layer by layer to finally complete the whole sub-influence network.

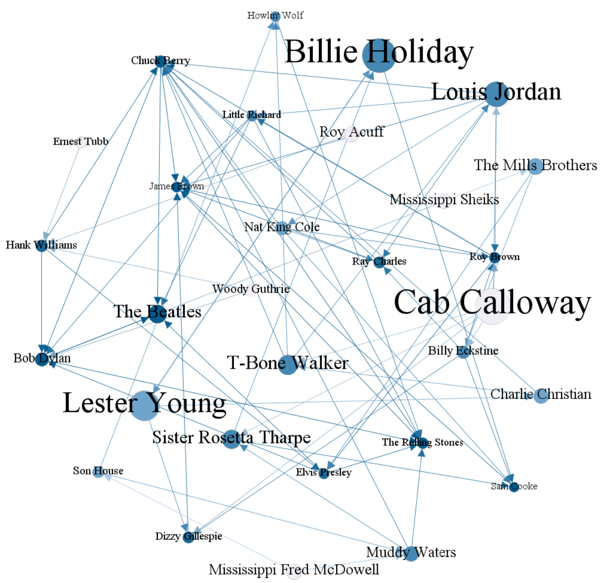


Figure 1: Directed Influencer Network

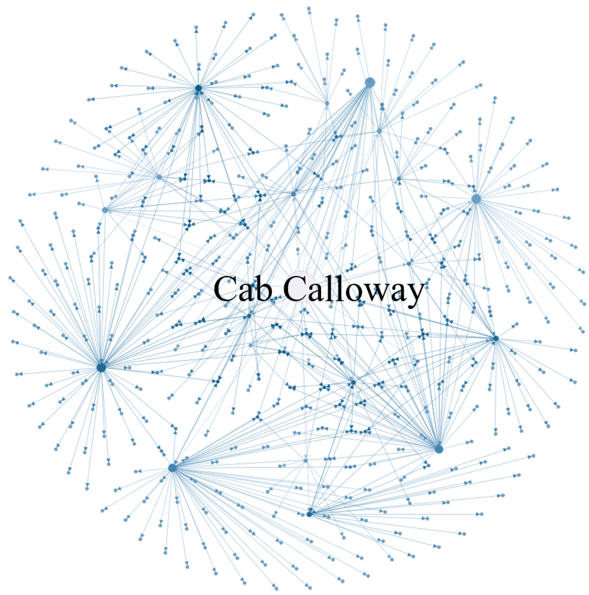


Figure 2: Cab Calloway Subnetwork

It can be indicated that the sub-influence network is scattered in the form of dots as shown in Figure 2, in which $Cab\ Calloway$ is undoubtedly the biggest focus, and the artists influenced by him have influenced many later musicians from generation to generation. Since then its influence has grown stronger over time based on this cascading relationship.

The discovery of this result changed the traditional perceptions to some extent. We used to think that the influence of a node in a network is only dependent on the $degree$ of its node, in other words, the more people it directly influences, the greater the artist's influence tends to be. However, the number of people directly influenced is just one of the important reference factors of their overall influence, there is also the long-term nature of their influence, that is, whether the influenced people can continue to influence future generations, which can be described as the quality of a single influence should also be taken into account.

In general, through the construction of the sub-network and its comparison with the complete network, we can conclude that the size of a musician's influence needs to be described by both the quantity and the quality of the influence. The quantity of influence can be represented as the $degree$ of the node, while the quality represents the $degree$ of influence diffusion of the previously influenced person in the future time. This conclusion is contrary to the results we obtained through emotional perception, but fits perfectly with the theory that the number of nodes and their quality are considered simultaneously in the network algorithm.

3.2 Subnetwork based on data in 1930-1990

In order to further explore the changes in the influence of musicians over time, we explore a subset of musical influence by creating a subnetwork based on data from 1930 to 1990 in your directed influencer network. The above PageRank algorithm is also utilized to sort the influence according to the $PR\ value$, and the results are indicated below.

In Table 2, there are slight differences in the ranking of the top 10 influence obtained by directed network and subnetwork, mainly reflected in the small ups and downs of the ranking, but the top musicians basically no change. The subnetwork also accords with the conclusion that the influencer needs to be described by both the number and the quality of the influence.

Table 2: Comparison of Top 10 Influencers

Top 10 Influencers in 1930-2010			Top 10 Influencers in 1930-1990		
Ranking	**Artist Name**	PR *value*	**Ranking**	**Artist Name**	PR *value*
1	Cab Calloway	0.0205	1	Cab Calloway	0.0218
2	Billie Holiday	0.0196	2	Billie Holiday	0.0208
3	Lester Young	0.0169	3	Lester Young	0.0183
4	Louis Jordan	0.0134	4	Louis Jordan	0.0142
5	T-Bone Walker	0.0099	5	T-Bone Walker	0.0104
6	Sister Rosetta Tharpe	0.0091	6	Sister Rosetta Tharpe	0.0097
7	The Beatles	0.0091	7	The Beatles	0.0092
8	The Mills Brothers	0.0076	8	The Mills Brothers	0.0081
9	Roy Acuff	0.0068	9	Charlie Christian	0.0072
10	Charlie Christian	0.0067	10	Roy Acuff	0.0072

Meanwhile, we found that the $PR\ value$ of 1990 is generally higher than the $PR\ value$ of 2010, which means that the impact of influential musicians declined over time during these two decades. By comparing Figure 1 and Figure 3, we show that the number of musicians increases after 20 years, the complexity of the influence relationship between musicians increases significantly, and the number of directional lines reflected in the network increases significantly.

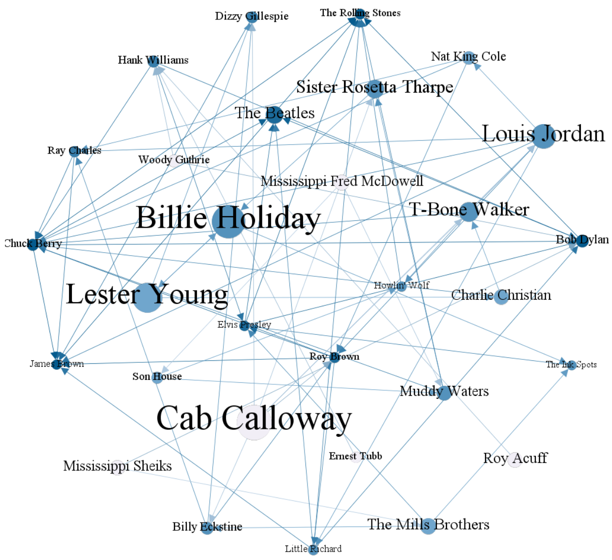


Figure 3: Subnetwork (1930-1990)

Therefore, the ranking of eminent musicians remained basically unchanged between 1990 and 2010, but there is a significant spread of influence over the past two decades. This conclusion accords with the increasingly rich situation of music revolution in today's era, which will be further discussed in the below section.

4 Task II: Feature-Based Similarity

In this section, we utilize Markov distance to measure all musicians in the $data\_by\_artist$, and then establish a complete measure of music similarity. Then the weighted undirected network is constructed on the basis of the similarity matrix. The **Infomap Algorithm** based on Huffman coding in entropy theory and random walk in information theory is utilized to obtain 209 new communities. Whether artists within the genre are more similar than artists between genres are further judged by thermal diagram.

4.1 Measure of Music Similarity

4.1.1 Selection of Data Sets and Variables

Since we need to compare artists within the genre and artists between genres later, here we considerate all the musicians in the $data\_by\_artist$. According to the Kalaganis [7], we choose ten characteristics $danceability$, $energy$, $valence$, $tempo$, $loudness$, $mode$, key, $acousticness$, $instrumentalness$ and $speechiness$ as the indicators of similarity. Since $liveness$, $duration\_ms$, $popularity$, $count$ play a smaller role in music similarity measurement , these four characteristics have been excluded.

4.1.2 Mahalanobis Distance

We choose **Mahalanobis distance** [8] d_{ij} as the distance measurement of music between i and j. The criterion of choosing Mahalanobis distance is as following.

- The music features selected in the data acquire their own ranges and are not dimensionally homogeneous. Mahalanobis distances are dimensionless and scale-independent, so it can be used to process musical features and to depict musical similarities.
- The correlation coefficient indicates that some music characteristics own strong correlation. For instance, energy have strong positive correlation with loudness (correlation coefficient is 0.797) and strong negative correlation with acousticness (negative correlation coefficient -0.793). Mahalanobis distance can consider all kinds of indexes synthetically and is not readily disturbed by the correlation between variables. It can be used to calculate without prior processing of each musical feature.

Then all distances are mapped to interval $[0, 1]$ as a measure of similarity by the exponential function $f(d_{ij}) = \exp^{-\frac{d_{ij}^2}{2}}$ of distance square.

4.2 Infomap Algorithm

The similarity between all musicians in the $data\_by\_artist$ is calculated according to the Mahalanobis distance function and the similarity matrix is constructed. The i line and j column represent the similarity of the two musicians in line i and column j. Since only 1.8% of the similarity is above 0.2 as shown in the Figure 4, and too small similarity is of no significance to the study of this problemwe select similarity more than 0.5 to construct a weighted undirected network.

Musician communities with a strong correlation can be used to reveal whether similar relationships and genre relationships are related. We obtained 209 minor communities in the new network using a informap algotithm. Infomap is a clustering algorithm derived from the principle of minimum entropy. More accurately, it is a community detection algorithm on graph networks. Specific steps are shown in Algorithm 2. Nine communities with more than 100 musicians is selected form the results. So far, we have established and screened high-related communities without considering genre.

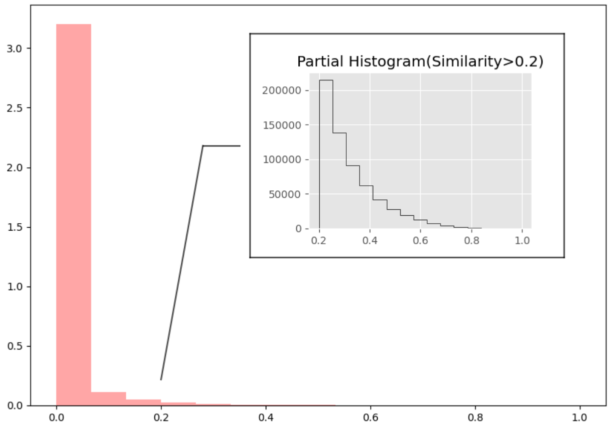


Figure 4: Histogram of Similarity

Algorithm 2 Infomap Algorithm

1: **Random walk** : abstract our sample set into a directed network, start from a point j, jump to the next point according to probability $p(i|j)$, then jump to the next point according to transfer probability again, repeat this process to get a long sequence.

2: **Huffman coding** : construct huffman coding based on random walk probability.

3: **Hierarchical coding** : by doing hierarchical coding of the sequence, minimize the target, so as to complete the clustering.

Next, the number of musicians of each genre in these communities is counted and the results are drawn in the thermal diagram. From the Figure 5, artists of the $Pop/Rock$, $Country$, and $R\&B$ have a higher number distribution in the above nine societies. It is worth noting that instead of being concentrated in one community, these three kinds of music artists are distributed evenly in different societies, which indicates that even if the two artists belong to the same genre, it does not mean a higher similarity between them. This is consistent with our actual life, the same genre of music still shows a high degree of difference, and the communication between music genres is very crucial to promote the development of music.

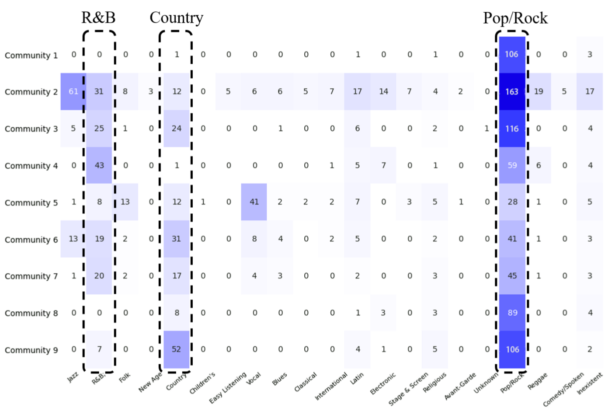


Figure 5: Genres in Each Community

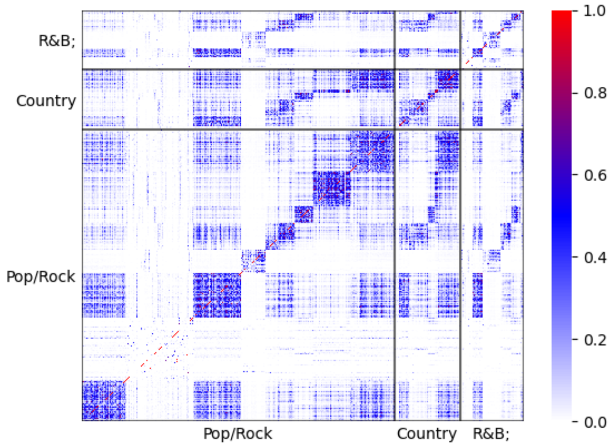


Figure 6: Group Similarity

In order to further observe the similarity between genres, we visualize the similarities between music artists of $Pop/Rock$, $Country$, and $R\&B$ in the nine groups. In the similarity thermodynamic diagram (Figure 6), it can be observed that:

- **The existence of the same group is a sufficient condition for high similarity.** The darker square area near the diagonal line means that there are many small groups among musicians of the same genre, referring to musicians of the same genre with higher similarity.
- **The same genre is not the inevitable condition of correlation.** There are a large number of rectangular areas with relatively light colors in the thermal map, most of which exist in the cross regions of two small group of diffierent genre, which shows that there is a certain similarity between the music artists of different genre, hence it is not completely irrelevant due to the different genres.

Therefore, it can be inferred that the same genre can indeed speculate on the similarity between musicians to a certain extent. If the two musicians belong to one small group, they must have a high correlation with each other, that is, the existence of the same group is a sufficient condition for high similarity. But the same genre is not the inevitable condition of correlation, since there exists similarities between different genres.

5 Task III: Genre Distinguishment

The influence $PR\ value$ and similarity of individual musicians respectively have been evaluated in Task I and Task II. In this section, we extend it to the overall influence and similarity between genres, and then apply LDA to obtain genre distinguishment, relationship and change over time.

5.1 Data Processing

We count the $PR\ values$ of all musicians as shown in Figure 7. The results show that most artists (more than 90%) have very small $PR\ values$. In order to better reflect the overall influence of a genre, we ignore the contribution of musicians with smaller $PR\ values$ to his genre, then select and average the top 10% artists with higher $PR\ values$ in each genre as the overall influence. The overall similarity of genres is obtained in the same way due to the uniform distribution characteristics.

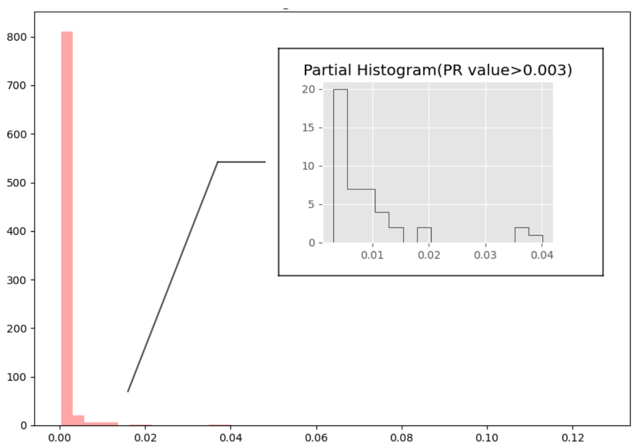


Figure 7: Histogram of Influence

In addition, according to the given statistics, due to the lack of data before 1930, the influence and similarity between 1930 and 1990 changed greatly and the accuracy could not be guaranteed, so the latter data were analyzed concretely. We separately calculated the influence and similarity of each genre from the 1930s to the 1990s , 2000s, and 2010s. Unfortunately, a single indicator alone does not accurately reflect the variation of the genre over time. Therefore, we consider analyzing the influence and similarity of genres simultaneously in a two-dimensional plane.

5.2 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) [9] classifies samples by reducing the dimension, aiming at finding a linear combination of raw data to separate each group of data so as to facilitate the analysis of data features. The Steps is shown in Algorithm 3.

Algorithm 3 Linear Discriminant Analysis (LDA)

Input: A data matrix $X_{n\times p}$=$\left(X_1^T, X_2^T, \ldots, X_n^T\right)^T$. Here, X_i means the $i-th$ sample and is a p-dimensional column vector;

The label corresponding to each sample.

Output: r linear combinations of X, i.e

$$y_1 = a_1^T X, \quad y_2 = a_2 X, \ldots, y_r = a_r X \tag{4}$$

1: Given that there are k classes, calculate the mean value of each class $\bar{X}_i$ and all the samples $\bar{X}$.

2: Calculate the between-class scatter matrix H and within-class scatter matrix E.

$$H = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})^T,\ E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T. \tag{5}$$

Here, n_i means the number of samples in class $i(i = 1, 2, \ldots, k)$.

3: Calculate the eigenvalue λ_i of the matrix $E^{-1}H$, its corresponding standardized eigenvector t_i such that $t_i^T \frac{E}{n-k} t_i = 1$ and the number of eigenvalues $r = \min(k-1, p)$.

4: Get r linear combinations

$$y_i = t_i^T(X - \bar{X}), \quad i = 1, 2, \ldots, r \tag{6}$$

LDA is selected based on this dataset to distinguish different genres as far as possible. We use influence and similarity as two dimensions to examine genres, and construct data matrix

$$X^{2010} = \begin{bmatrix} x_{11}^{2010} & x_{12}^{2010} \\ \vdots & \vdots \\ x_{i1}^{2010} & x_{i2}^{2010} \\ \vdots & \vdots \\ x_{20,1}^{2010} & x_{20,2}^{2010} \end{bmatrix}_{20\times 2} \tag{7}$$

by selecting the influence and similarity of 20 genres from 1930 to 2010. Here, x_{ij}^k means values of the influence and similarity from the $1930s$ to the k, the j dimension of the i_{th} genre, $k = 1990, 2000, 2010; i = 1, 2, , 20; j = 1, 2$. The genre of data indicates its label. Matrix X^{1990}, X^{2000} is constructed similarly. Finally, we put three 20×2 matrices together to get the matrix

$$X = \begin{bmatrix} X^{1990} \\ X^{2000} \\ X^{2010} \end{bmatrix}. \tag{8}$$

The matrix X is substituted into LDA for operation.

5.3 Result of LDA

We transform all the points in three time periods, by Linear Discriminative Analysis model as follows.

$$\begin{bmatrix} y_1 & y_2 \end{bmatrix} = \begin{bmatrix} x_1 - \bar{u}_x & x_2 - \bar{u}_y \end{bmatrix} \begin{bmatrix} -89.38 & -12218.14 \\ -191.28 & 67.51 \end{bmatrix} \tag{9}$$

$\begin{bmatrix} \bar{u}_x & \bar{u}_y \end{bmatrix}$ is the average of all coordinate points. The coordinates are converted and the graphic result is shown in Figure 8.

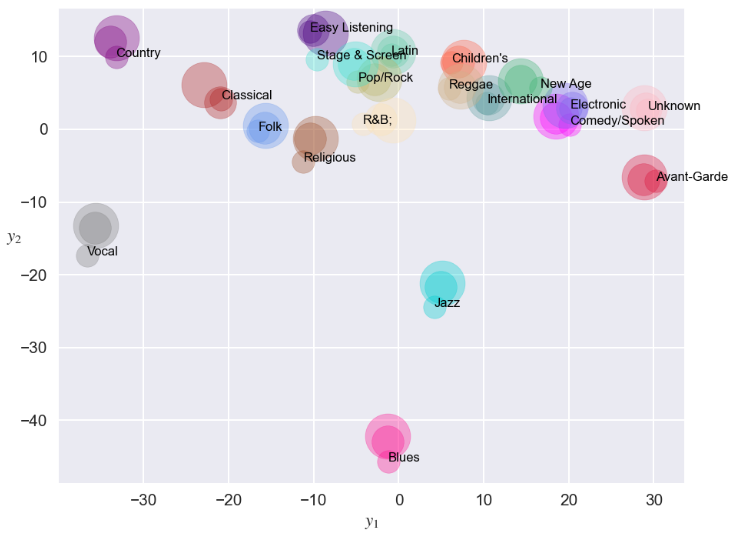


Figure 8: Genre Evolution

Genres which can be easily distinguished are $Vocal$, $Blues$, $Jazz$, $Avant - Garde$. However, even those most recognizable genres can not be distinguished easily if we focus on one dimension. For example, $Vocal$ and $Avant - Garde$ can be easily recognized when choosing vertical direction, but we cant tell the difference between $Blues$ and $Jazz$. Conversely, if we choose horizontal direction, then $Blues$ and $Jazz$ are recognizable, but it may be hard to identify the difference between $Vocal$ and $Avant - Grade$. This phenomenon suggests that it is the combination of influence and similarity which makes genres distinguishable.

Moreover, there are mainly two patterns of changes and the latter shows the relation between genres in Figure 8.

- The first type can be classified as **natural change**, which includes $Avant-Garde$, $Blu-es$, $Classical$, $Country$, $Folk$, $Jazz$, $R\&B$, $Religious$, $Vocal$, $Unknown$. On the one hand, these 8 genres have changed flatly over time, on the other hand, these genres have distinct characteristics and are relatively distant from other genres in Figure 8.

- The second type should be classified as **merging change**, there are two small groups which show the merging phenomenon, first consists of $Easy\ Listening$, $Latin$, $Pop/Rock$ and $Stage\ \&\ Screen$, second is $Children's$, $Comedy/Spoken$, $Electronic$,$International$, $New\ Age$ and $Reggae$. Their members tend to have more and more resemble structure of influence and similarity. It can be presumed that the combination of influence and similarity cant be used to recognize members in both groups after several years due to merging of specific genres.

6 Task IV: Correlation between Similarity and Influence

This section describes the methodology used in this paper to assess the relationship between similarity and influence. The influence-based similarity (IBS) is set up according to the cosine distance. To make it clear, the previous similarity measure developed in Task II shall be mentioned as whole-feature-based similarity (WFBS), and the single-feature-based similarity (SFBS) after normalization is also crucial. P, Q, q are respectively calculated by IBS, WFBS, SFBS. The results are produced by their comparisons.

6.1 Consistency of Similarity and Influence

6.1.1 Measure of Influence-Based Similarity (IBS)

In order to investigate the true relation between similarity and influence, we propose a new similarity measure based on influence relation (influence-based similarity). Intension of developing a new similar measure is if the influencer actually influences the follower, then the music features between them will be more similar than others, which indicates the consistent relationship between high similarity and high influence.

A **centralised cosine measure** [2] is used here to evaluate the new measure. This measure can be utilized to judge the statistical significance of the association between two artists. We only retain the similarity with 5% confidence for the results.

6.1.2 Kullback-Leibler Divergence

For a certain artist, we can respectively obtain two similarity vectors based on two different measures. And normalize them (the sum value of both vectors is 1) by multiplying a scalar, which can be treated as probability distribution. Then **Kullback-Leibler divergence (KL divergence or relative entropy)** is introduced to measure their difference.

$$\mathrm{KL}(P||Q) = \sum P(x) \log \frac{P(x)}{Q(x)} \tag{10}$$

6.1.3 Comparision Result between IBS and WFBS

We take artists both in $influence\_data$ set and $data\_by\_artist$ as the objective of our analysis. In Figure 9, the KL divergence distribution histogram of the IBS and WFBS comparison results indicates that most of the KL divergence are within the range of 0 to 10. When we take all artists into consideration, there is no necessary relation between similarity and influence relation. However, when we select artrists with low Kullback-Leibler divergence (<3), the group structure between these artists is found in Figure 10, which suggests a positive correlation between high similarity and high influence.

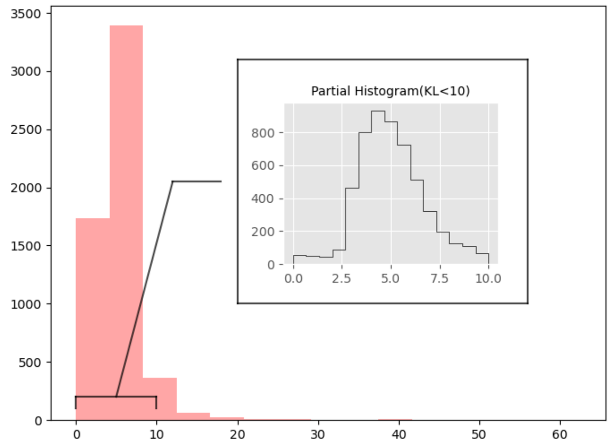


Figure 9: KL Divergence of the IBS and WFBS

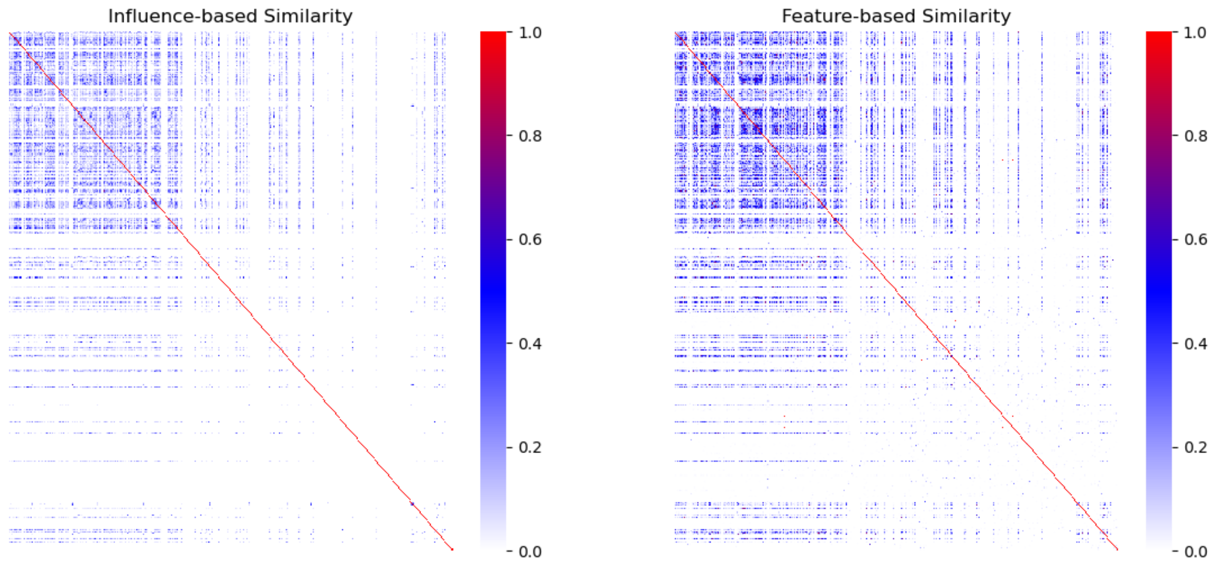


Figure 10: Artist Similarity with Small KL Values

6.2 Contagious Features

6.2.1 Measure of Single-Feature-Based Similarity (SFBS)

In terms of exploring the change trend of a single feature in similarity relation, we need to establish a measure of Single-Feature-Based Similarity (SFBS). We simply take the differential absolute value of different musicians as a distance function to measure a single feature and normalize it as following.

$$r_{ij} = \exp(-\frac{d_{ij}^2}{2}) \tag{11}$$

6.2.2 Comparision Result between IBS and SFBS

When an artist is influenced by another artist, it is likely that the influence only occurs in some contagious music features, (that is musical features that are highly similar and easy to be imitated). The total music feature of influencer and follower is not highly similar according to *section 6.1.3*. In order to discover the relationship between a certain feature and influence, we also measure the IBS and SFBS results with KL divergence to obtain the contagious features.

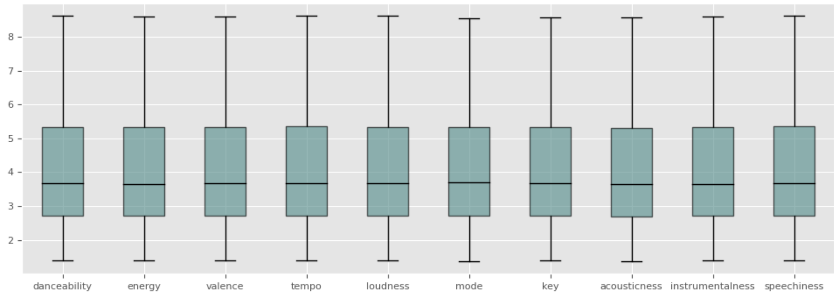


Figure 11: KL Divergence of Individual Musical Features

The distribution of Kullback-Leibler divergence of each music feature we choose are highly similar in Figure 11. There is no feature that seems more contagious at the overall level. Hence, we also only observe the thermodynamic diagram of the correlation between the characteristics of KL divergence less than 3.

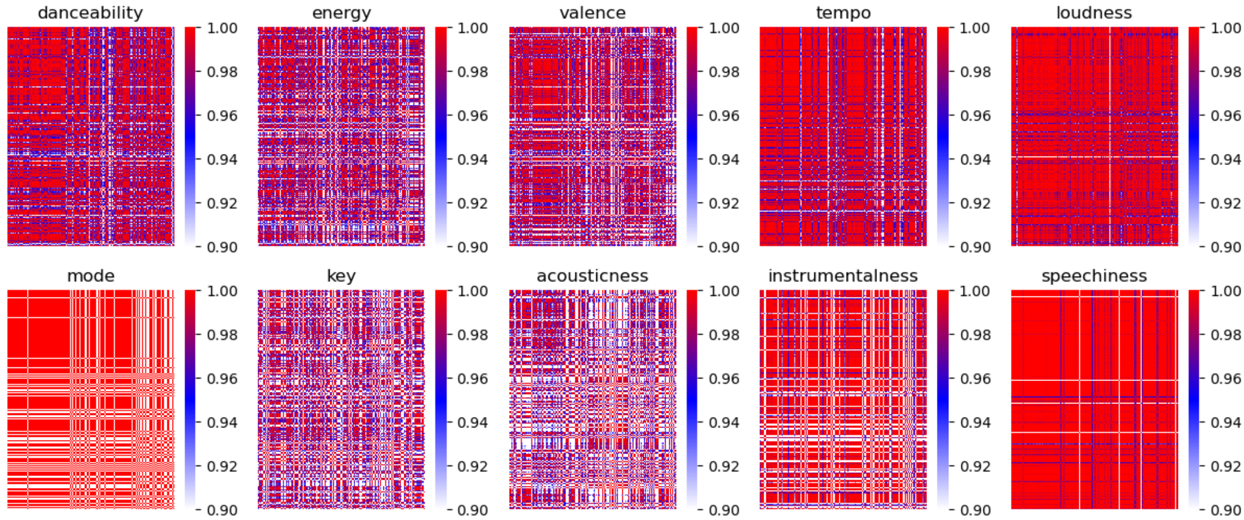


Figure 12: Contagious Musical Features

It can be observed from Figure 12 that to the similarity of 0.90, *speechiness*, *loudness*, *tempo* more contagious than others, followed by *danceability*, *instrumentalness* and *valence*. While *mode* also has large areas of red in the thermal maps, we do not regard it as contagious features, since *mode* is a 0-1 variable and most musicians have a mode of 1.

7 Task V: Features and Artists signifying Music Revolution

Before analyzing which features and musicians that might signify revolutions, we are supposed to estimate the probability of musical revolutions based on the *data_by_year* data set (since the revolution is mainly captured by time). Moreover, the feature *mode* only takes one value in the data set, which enables us to ignore it.

7.1 The Rocognition of Music Revolution

We firstly visualize the change of selected features shown in Figure 13, the curves fluctuate so much that it is difficult to identify the actual tendency of certain values. Thus, moving average method is used to show the real tendency behind each curveswhich can effectively eliminate the random fluctuation. We extract the turning point of years when the trend of each feature

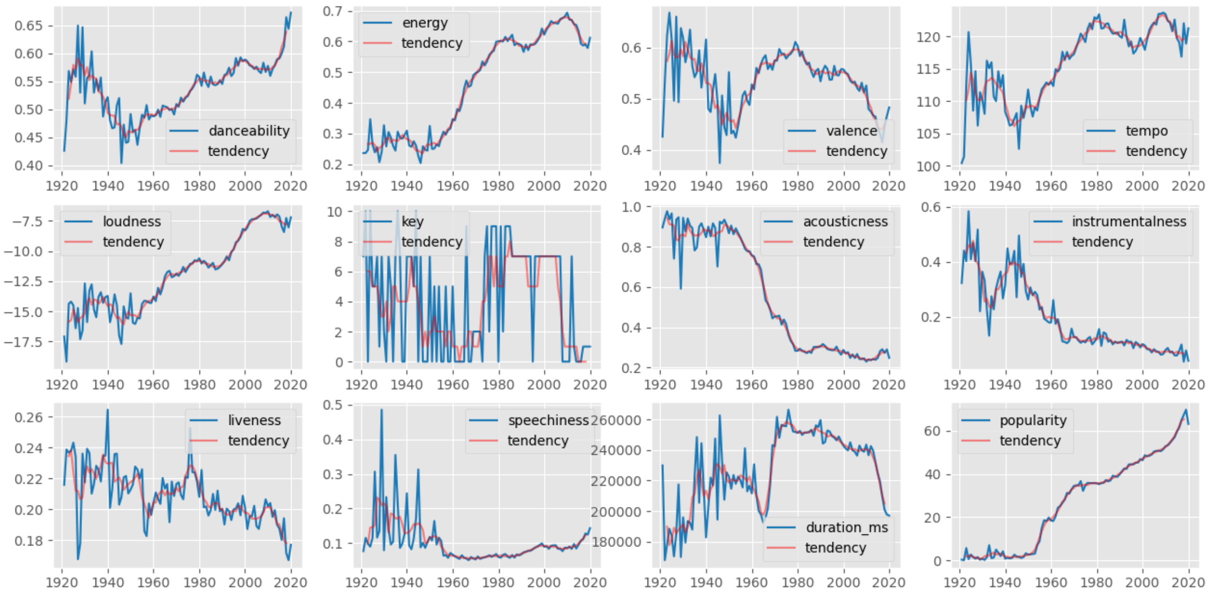


Figure 13: The Tendency Curve of Each Feature

begins to change and compare them. As showing in Figure 13, there are some similarities between different features. For axample, *danceability*, *energy*, *valence*, *tempo*, *loudness*, *key*, *acousticness* and *popularity* show almost the same change tendency during 1945-1950.

In order to figure out the possible number and the probability of music revolutions that happen during 1920-2020, we use **Kernel Density Estimation** based on the selected years to give a possible probability distribution of music revolutions. The relatively possible time for musical revolution is 1928, 1946, 1979, 2009, as shown in Figure 14.

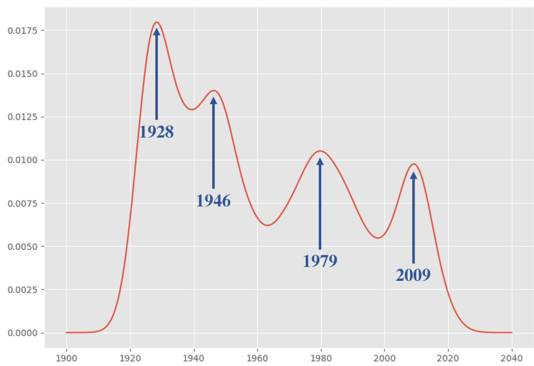


Figure 14: Result of Kernel Density Estimation

7.2 Characteristics signifying Music Revolution

By comparing the change of curves as shown in Figure 14, there are multiple features that show signs of changing tendency for years for musical revolution, such as $loudness$, $danceability$, $energy$, key, $valence$ and $liveness$. Therefore, they can be regarded as the factors that have a major impact on music revolution to some extent.

Among all the features above, the change in $valence$ (a measure describing the musical positiveness conveyed by a track.) is not only varies the most considerably, but also highly consistent with previous estimates of the musical revolution.

Generally speaking, $loudness$, $danceability$, $energy$, key, $valence$ and $liveness$ are likey to reflect music revolution to some extent, and $valence$ can significantly signify revolutions (major leaps).

7.3 Musicians signifying Music Revolution

Based on the above discussion, it is reasonable to investigate artists who represent revolutionaries through valence-based similarity network. Then, we detect the community equipped with both feature similarities and influence structures. Specifically, the network should be subjected to the following two constraints.

- **Feature similarity:** artists in the community take similar values of valence.
- **Influence similarity:** artists in our community are similar in influence-based network.

We draw the line between two artists if and only if their valence-based similarity and influence-based similarity are both greater than threshold values.

By performing Infomap algorithm mentioned in $section$ 4.2, we obtain a community with 744 artists that represent revolutionaries, their distribution in different genre is shown in Figure 15.

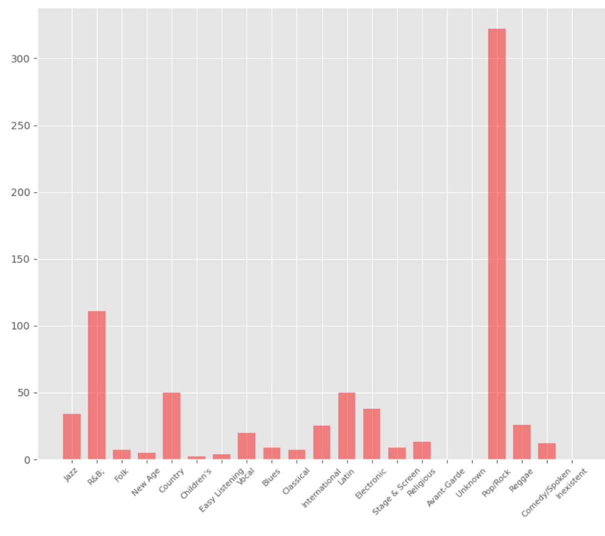


Figure 15: Distribution of Musicians Signifing Music Revolution

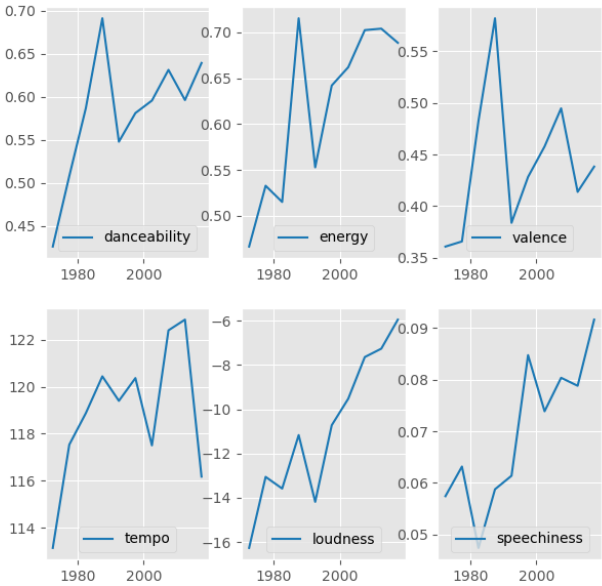
By performing Infomap algorithm mentioned in $section$ 4.2, we obtain a community with

744 artists that represent revolutionaries, their distribution in different genre is shown in Figure 15.

8 Task VI: Discussion of Evolution and Dynamic Influencer

When analyzing consistency of influence relation with feature-based similarity, we discuss the most contagious features of music, in the discussion of recognization of musical revolutions, we discovered some features which tendency over time can reflect the period of revolution. Now we focus on a certain genre and try to describe the pattern of different features.

The genre we choose to investigate is $Electronic$, we select two sets of features: three contagious feature $speechiness$, $loudness$, $tempo$ and three features of $valence$, $danceability$, $energy$, which can be used to signify the musical revolution. Their changes over time are shown in Figure 16, to measure the similarity of changing patterns, the correlation coefficient heatmap is shown in Figure 17.



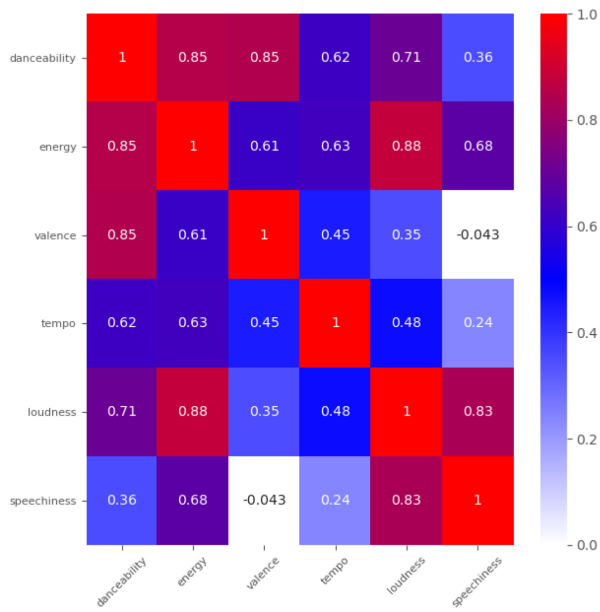


Figure 16: Indicator Changes

Figure 17: Time Series Correlation

It can be seen that high correlation occur between features that are used for revolution recognization, and their curves show a similar tendency. Since $danceability$, $energy$ are also contagious to some extent, there are some blocks with higher correlation in the cross area of two sets of features. An interesting phenomenon is that $tempo$ shows a decreasing tendency around the year 2009 while other two features of same set are likely to keep increasing, which suggests that the response ability of each features differs from each other.

Next, We want to describe the dynamic change over time of influencer in genre $Electronic$, though PageRank algorithm can be applied to track the change of each artists inside $Electronic$ genre, we would like to develop an indicator to illustrate the change of influence structure. Instead of using a cumulative method, we calculate the average value of popularity of a certain artist in a given period, average value is set to 0 if an artist has no popularity in a period of time.

The result in *Electronic* genre is shown in Figure 18. The Y-axis presents time(year,starts from1970) and X-axis presents numerical ID for artists in genre *Electronic*.

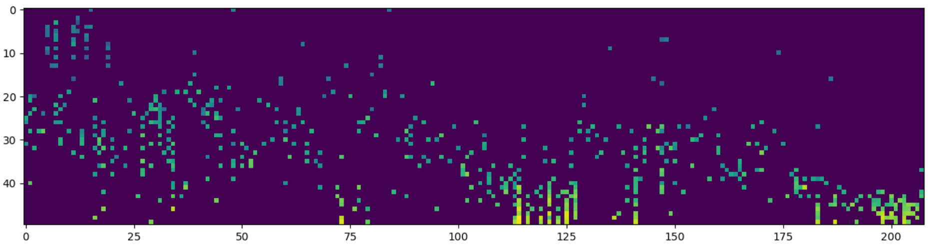


Figure 18: Dynamic Impact

At first, the most popularity artists are those whose ID is less than 15, as time goes by, major artists with higher popularity have ID between 0 and 100, and recently artists with ID close to 120 or 200 are most popular. These different distribution of popularity reveals that the influencer inside one genre will change over time instead of keeping the same.

9 Task VII: External Influences on Music

As a major form of reflecting high-class ideology about society reform, art is undergoing a great innovation. Music, as an art form susceptible to the trend of the era, has also undergone a great change. Brand-new music forms have sprung up like mushrooms in the 20th century. Consequently, initial music forms no longer exist and a royal music revolution occurred.

As a part of society, music has been influenced by social, political or technological changes. For example, the prevailing of *Electronic* largely depends on the development of electronic and informational technology.Actually, these external influences can be signified in our model.

9.1 Cultural Influence on Music

According to Nikolsky Aleksey [10], the transformation of composers' minds and ideological emancipation led to the emergence of many new genres, such as the genre New age and Electronic. Moreover, it has been pointed out in *section* 5.4 that genres tend to merge in the music evolution. These genres to be merged can be viewed as a kind of cultural fusion. Consequently, different culture leads to more various music genres and similar genres tend to merge over time.

9.2 Social Influence on Music

Through the dataset $data\_by\_year$, the Figure 19 demonstates the tendency of $popularity$ from 1921 to 2020. The value of popularity boosted in 1950s and experienced a unexpected decline in 2020.

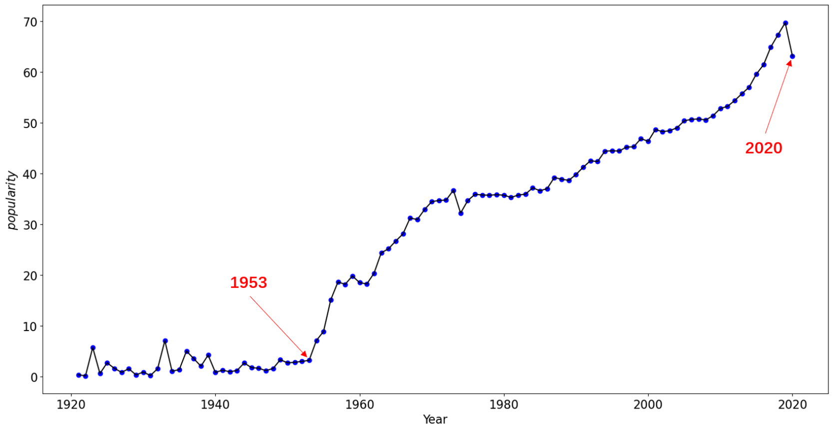


Figure 19: Tendency of *Popularity*

Noted that the world before 1950 was in conflict and confrontation, many people struggled against hunger and cold just to be alive. Naturally, they had no mood to enjoy music or even created new music styles. However, with the life goes better, we tend to pursue spiritual satisfaction and music gradually prospers.

In the same way, we underwent a severe epidemic in 2020 and people's health was under threaten. People care more about their safety and fight with the disease by all means. Under such an intense social circumstance, less music albums published and less attention was paid on music.

9.3 Racial and Religious Influence on Music

As mentioned in *section* 5.4, there exists some genres that are easily distinguished by their influence and similarity, such as *Religious* and *Blues*. Religious music is any type of music that is performed or composed for religious use or through religious influence. Blues is a music genre originated from roots in African musical traditions. The origins of the blues are closely related to the religious music of the Afro-American community. It is the religious-related feature that make these two genres easy to be distinguished from others.

10 Strengths and Weaknesses

10.1 Strengths

- Our network reveals that the relation between influence and feature appears differently at different scale.
- Mahalanobis distance is a scale-invariant measure and can be applied to process inhomogeneous data set in this problem.
- Instead of transfering similarity measure to influence infomation, we develop a new similariry based on influence ralationship, and it can be easily generalized to investigate the relation between single feature of music and influence relation of artists.
- Most analysis and results are carried out based on data and mathematical method and not much background knowlegde about music is required.

10.2 Weaknesses

- Limitied to the availability of the data, the accuracy of the result may be unstable, which needs more reliable data and further consideration.
- The thresholds are problem-specific and should be adjusted according to the problem we are handling.
- Since the interrelation is complex, the model should be used cautiously.

11 A letter to the ICM Society

Dear ICM Society,

It is our great honor to be identified by your association to do research on the evolutionary and revolutionary trends of music. We have set up several networks and developed models that quantify the musical influence. Stated below are some results and prospect of our research.

Music plays an important role in our daily life and is woven into the fabric of society. Due to the intricate links between musicians and different music style, the music world can be viewed as a complex system. Thus, it is pleased to use network science to better explain and demonstrate the music world. Based on the networks and models we build, we can obtain the following conclusions.

- The ranking of high-impact musicians is basically unchanged over time, but its influence may decrease with time.
- The same genre will not guarantee high similarity between musicians.
- Some specific genres will merge over time.
- Social events, technological advancements, political changes do affect the music evolution.

In spite of the limited time, our research through network is meaningful. With the help of network, we can study the music evolution in a multidisciplinary, quantitative and mathematical way. First, the visualization of the network makes the links between two musicians more intuitive. Second, the community in network science greatly describes the music genres. Moreover, dynamic analysis of the network enables us to better understand how different genres change over time due to some certain external influences.

Considering the two data sets are limited to only some genres, the real relationship between different musicians and genres may be much more complicated. The richer data is, the more complex it will be. But we have some ideas and prospects about the research if given more data.

- We need to explore more accurate indicators that can reflect the influence of musicians and the similarity between them.
- It will be harder to clearly visual the network since plenty of nodes and edges are included. We may use some heuristic algorithms to detect communities in the network.

- The result of our work may be greatly changed if there are some influential musicians excluded from the two data sets we get.

Thanks for taking time out of your busy schedule to read my letter. We hope your association can continually support us. Your prompt attention to my inquiry would be highly appreciated.

Yours truly

ICM 2021 Team

References

[1] Collins, N., Computational Analysis of Musical Influence: A Musicological Case Study Using MIR Tools, International Society for Music Information Retrieval Conference, DBLP, 2010.

[2] Georges, P., Western classical music development: A statistical analysis of composers similarity, diferentiation and evolution, Scientometrics, Vol.112, No.1, 2153, 2017.

[3] Mark, Z., Ichiro, F., Web services for music information retrieval, In Proceedings of the International Symposium on Music Information Retrieval, Barcelona, 2004.

[4] Georges, P., Nguyen, N., Visualizing music similarity: Clustering and mapping 500 classical music composers, Mimeo: University of Ottawa, 2019.

[5] Ben, F., Michael, C., Kurt, J., Mark, S., Do you sound like your friends? Exploring artist similarity via artist social network relationships and audio signal processing, In Proceedings of the International Computer Music Conference, 2008.

[6] Henzinger, M., Web Information Retrieval - an Algorithmic Perspective, Algorithms - ESA 2000, 1-8, 2000.

[7] Kalaganis, F., Adamos, D., Laskaris, N., A consumer BCI for automated music evaluation within a popular on-demand music streaming service, Artificial intelligence applications and innovations, 429–440, 2016.

[8] Mahalanobis., Prasanta, C., On the generalised distance in statistics, Proceedings of the National Institute of Sciences of India, Vol.2, No.1, 21-53, 2016.

[9] Alhajj, R., Rokne, J., LDA, Encyclopedia of Social Network Analysis and Mining, 785-785, 2014.

[10] Nikolsky, A., Perlovsky, L., Editorial: The Evolution of Music, Front psychol, 2020.

Appendices

Appendix A Python Code

```python
# Problem 1 - PageRank
import networkx as nx
if __name__ == '__main__':
        # edges is a list with form [(,),(,),...]
        G = build_diGraph(edges)
        pr_impro_value = nx.pagerank(G, alpha=0.85)
        print(pr_impro_value)

# Problem 2 - Similarity
import pandas as pd
import numpy as np
def Distance(artist_1, artist_2, cov_inv_mat):
        n_dim = len(artist_1)
        delta_vector = (artist_1 - artist_2).reshape((1, n_dim))
        distance_value = np.sqrt(
                np.dot(
                        np.dot(delta_vector, cov_inv_mat),
                        delta_vector.T
                )[0,0]
        )
return distance_value
def FormDistance(data_mat):
        cov_mat = np.cov(data_mat, rowvar=False)
        cov_inverse = np.linalg.inv(cov_mat)
        res_mat = np.zeros((data_mat.shape[0], data_mat.shape[0]))
        for i in range(data_mat.shape[0]):
                for j in range(i+1, data_mat.shape[0]):
                        res_mat[i][j] = Distance(data_mat[i], data_mat[j], cov_inverse)
                        res_mat[j][i] = res_mat[i][j]
        return res_mat

if __name__ == '__main__':
        # load data_by_artist
        data_by_artist = pd.read_csv('data_by_artist.csv')
        # select 'artist_id':'popularity'
        labels_list = ['danceability', 'energy', 'valence','tempo', 'loudness', 'mode', 'key',
        'acousticness', 'instrumentalness','speechiness']
        data = data_by_artist.loc[:,labels_list]
        # form Mahalanobis distance matrix
        Ma_dis_mat = FormDistance(np.array(data))

```

```
#Problem 2 - Infomap
import infomap
if __name__=='__main__':
        # create a network and run Infomap algorithm
        similar_network = infomap.Infomap()
        similar_network.read_file('SimilarNetwork.net')
        similar_network.run("-N5")
        result = similar_network.get_modules(depth_level=2)

# Problem 3 - Linear Discriminant Analysis
def LinearDiscriminantAnalysis(X_, Y_):
        n_sample, n_dim = X_.shape
        mean_all = np.mean(X_, axis=0)
        class_list = []
        for i_ in Y_:
                if i_ not in class_list:
                        class_list.append(i_)
        n_class = len(class_list)
        X_class,numbers_class,mean_class  = [],[],[]
        for class_value in class_list:
                X_class.append(X_[Y_==class_value])
                numbers_class.append(X_class[-1].shape[0])
                mean_class.append(np.mean(X_class[-1], axis=0))
        Sb = np.zeros((n_dim,n_dim))
        for i_ in range(n_class):
                temp = (mean_class[i_] - mean_all).reshape((n_dim,1))
                Sb += numbers_class[i_]*np.dot(temp,temp.T)
        Sw = np.zeros((n_dim,n_dim))
        for i_ in range(n_class):
                for j_ in range(numbers_class[i_]):
                        temp = (X_class[i_][j_] -
                         ↪  mean_class[i_]).reshape((n_dim,1))
                        Sw += np.dot(temp,temp.T)
        W_ = np.dot(np.linalg.inv(Sw),Sb)
        n_eig = min(n_class-1, n_dim)
        eig_value, eig_vector = np.linalg.eig(W_)
        eig_vector = eig_vector.T
        eig_vector = eig_vector[np.argsort(-eig_value)]
        eig_value = eig_value[np.argsort(-eig_value)]
        for i_ in range(n_dim):
                temp = eig_vector[i_].reshape((1,n_dim))
                normal_value = np.dot(np.dot(temp,
                 ↪  Sw/(n_sample-n_class)), temp.T)
                eig_vector[i_] /= np.sqrt(normal_value[0][0])
        Contribution = eig_value[0:n_eig]/sum(eig_value)
        Rotator = eig_vector[0:n_eig].T
        New_X = np.dot((X_-mean_all), Rotator.reshape((n_dim, n_eig)))
        return Contribution, Rotator, New_X
```

```

# Problem - 4 - KL divergence
import numpy as np
def Entropy(P_vector):
        new_P = P_vector[P_vector>0]
        return np.sum(new_P*np.log(new_P))
def KL(P_vector,Q_vector):
        P_vector = P_vector / np.sum(P_vector)
        Q_vector = Q_vector / np.sum(Q_vector)
        result = Entropy(P_vector) - np.sum(P_vector*np.log(Q_vector))
        return result
if __name__ == '__main__':
        Total_KL = np.zeros((11, 5602))
        influence_mat = np.load(file_name_1)
        feature_mat = np.load(file_name_2)
        for i in range(5602):
                Total_KL[index][i] = KL(influence_mat[i], feature_mat[i])

# Problem - 5 - Kernel Density Estimation
from sklearn.neighbors import KernelDensity
if __name__=='__main__':
        model = KernelDensity(kernel='gaussian', bandwidth=5)
        model.fit(Total) # Total consist all the years where tendency of a feature changes
```