Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
D		2109805

A Comprehensive Study of the Influence of Music Over Time

Summary

Music has been an essential part of human society since its origination due to its unique attractiveness, and has also been a real-time image of political, economical and cultural landscape. Thus, it has been an interesting and important topic to study the evolution of music, which not only reveals a part of history of art, but also offers an insight into the landscape of real-time society. Certainly, this involves a variety of aspects: the status of a society, the influence of older generations that passed onto the young ones, and the musicians themselves. Our work stresses the influence part, providing a quantitative approach to measuring the influence of previously composed music (and the musical artists involved) on new musical and musical artists, which is illustrated as below.

In **Task 1**, we establish a **directed network** of musical influence. For measure of influence, we use a comprehensive score which takes multiple factors into consideration (e.g. degree, betweenness, closeness and eigenvector centrality, also using Google's PageRank algorithm). We then build up a subnetwork of leaders and followers, revealing the meaning of our influence measure.

In **Task 2**, we develop a **similarity function** as a measure of musical similarity. We carry out a correlation check to rule out factors that may result in collinearity, and select a modified Manhattan distance function to measure musical similarity. We then calculate the similarity function within and between genres and visualize the results, confirming the assumption "Artists within one genre are much similar compared to artists between genres".

In **Task 3**, we use models built in **task 1** and **task 2** to evaluate genres. We apply an agglomerative clustering algorithm to further analyzing the similarities. Also, we apply the network that we establish in Task 1 to analyzing the influence between genres. We also discuss the characteristics that categorizes genres and the changes of characteristics over time.

In **Task 5**, we visualize the changes of characteristics with time to give a rough definition of musical revolution. Then we define a function to quantify how revolutionary one artist can be, and check whether it is consistent with the revolutions we obtain in this problem.

In **Task 6 and 7**, we apply the results obtained from **Task 1 to 5**, identifying dynamic influencers, and discussing the influence of cultural, social, political and technological changes within our model. We conduct a case study here according to the outcomes of our model.

Contents

1 Introduction

Music has been the carrier of emotions and thoughts, reflecting the cultural, political and economic landscape in the mean time. Masterpieces with enduring tune and inspiring lyrics are immortal, inspiring people of all time and reminding people of their spirits. As music evolves over time, difference emerges and different artists have developed style of their own. For now, artists can be categorized into different **genres**, featuring their unique **characteristics**. Also, there is strong influence between previously composed music and new music and artists, as people tend to learn from what they have heard and add their own style.

There have been countless studies of the evolution of music, and influence between music has always been a focus. However, previous research tends to be qualitative rather than quantitative, since music itself is subjective. While scholars share consensuses about influence, they can also diverge as to details. New artists may claim that they are deeply influenced by previous music and artists, but we cannot tell to what extent do previous music influence new artists. Having seen all drawbacks, our group provides a quantitative approach to measuring musical influence using all the data provided by the ICM community, in an attempt to better evaluate the influence of music and boost the frontier research.

2 General Assumptions

(1) The process of influence is unidirectional, passing from influencers to followers only. Thus, we can use a directed network to characterize this process.
(2) All indexes reflecting characteristics of music represent different aspects so that they should be treated equally. Thus, they shall share the equal weight in our modeling.

3 Models

3.1 Task 1: Directed network and measures of influence

3.1.1 Model Preparation

Since the process of influence is unidirectional (from previous artist to new artist), we use a directed network for characterization. Each node stands for one particular artist, and each directed edge stands for the process of influence. It is important to consider the influence of one particular artists (i.e. the out-degree of nodes), but we also want to measure how complete traits are passed along by influencers, and how important one is over the history evolution of music (i.e. inheriting previous music and pass it along, generating huge influence).

We consider the following indexes for measure of influence. The indexes can be extracted from the directed network we develop, using **BFS or DFS algorithm**.

(1) Degree centrality:

Degree centrality is the most fundamental measure of centrality of nodes and network, which was proposed by Linton Freeman in 1979. In the network of artists, it can be used to characterize to what extent can an artist influence others and be influenced by others, the former being our biggest concern.

For a network with g nodes, the degree centrality of node i can be calculated as follows:

$$C_D(N_i) = \sum_{j=1}^{g} x_{ij}(i \neq j), \tag{1}$$

where $C_D(N_i)$ is degree centrality of node i, and x_{ij} is the weight of connectivity of node i and node j, only counting direct out-connected paths.

To eliminate the impacts brought by scale of the network, we use the standardized formula proposed by Stanley Wasserman and Katherine Faust, as follows:

$$C'_D(N_i) = \frac{\sum\limits_{j=1}^{g} x_{ij}(i \neq j)}{g-1}. \tag{2}$$

where factor $g-1$ eliminates the impact brought about by scale.

(2) Betweenness centrality:

Betweenness centrality reflects to what extent can one node controls information transmission via other nodes. In our model, we want to evaluate to what extent an artist can inherit and pass along the worthy traits of previous music, developing his style along the way. It can be calculated as follows:

$$C_B(N_i) = \frac{\sum_{i=1}^{g} \sum\limits_{j=1}^{g} N_{ij}(between\ N_i)}{\sum\limits_{i=1}^{g} \sum\limits_{j=1}^{g} N_{ij}}, \tag{3}$$

where N_{ij} is the number of shortest path from node i to j.

(3) Closeness centrality:

Closeness centrality reflects the distance between one particular node to others, which can be calculated as follows:

$$C_C(N_i) = \frac{1}{\sum_{j=1}^{g} d_{ij}(i \neq j)}. \tag{4}$$

d_{ij} is the shortest distance between node i and j.

(4) Eigenvector centrality:

Eigenvector centrality not only reflects the degree of one particular node, but also reflects the degree of all neighboring nodes. For our model, we not only evaluate the influence of one particular artist, but also consider to what extent his influencers can pass along his traits. It can be calculated as follows:

$$C_E(N_i) = ce_i = c \sum_{j=1}^{g} a_{ij} ce_j, \tag{5}$$

where c is a constant.

Let $\mathbf{ce} = [ce_1, ce_2, ce_3, \ldots, ce_n]^T$. It can be written in the following format upon reaching steady state after multiple iterations:

$$\mathbf{ce} = c\mathbf{A}\mathbf{x}, \tag{6}$$

where $\mathbf{x}$ is the eigenvector of matrix $\mathbf{A}$ with an eigenvalue c^{-1}.

(5) PageRank:

First applied by Google, PageRank (PR) algorithm reflects the importance of nodes in **a directed network**, and is calculated through the following steps:

Step 1: Assign a PR value to each node.

Step 2: Iterate until reaching a steady state, using voting algorithm as follows:

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}, \tag{7}$$

where B_u is the set of nodes connected to u, $v \in B_u$, and L_v is the out-degree of node v.

Note: In order to better illustrate our data and better further process, we have normalized all our measures so that it has a mean of 0 and a variance of 1, which is realized as follows:

$$N = \frac{X - \mu}{\sigma}, \tag{8}$$

where μ and σ stands for mean and standard deviation of the data, X is the original data, and N is the normalized one.

3.1.2 Model Establishment

(1) The derivation of comprehensive score:

We use a **comprehensive score** to illustrate the extent of musical influence using **factor analysis**. Factor analysis is a powerful tool to determine the weight of 5 contributing indexes, which focuses on different aspects.

Let $Z = (Z_1, Z_2, ..., Z_m)$ and $F = (F_1, F_2, ..., F_n)$ be the influence measure for m artists, and F be the influence index vector, where $n = 4$. Therefore,

$$Z = AF + \epsilon. \tag{9}$$

where $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_m)$ is a special factor and $A = (a_{ij})_{m \times n}$ is the **factor load matrix** and satisfies:
(1) $m \geq n$,
(2) $cov(F, \epsilon) = 0$.

Through the **SCREE SHEET CRITERION**, according to the order of factor extraction, we draw the scatter plot of the eigenvalues of the factors with the number of factors and determine the number of factors according to the shape of the graph. The shape of the graph starts from the first factor like a mountain peak. After that, the curve drops rapidly and becomes a straight line. The number of factors corresponding to the previous point after the curve flattens is considered as the maximum factor of extraction. The data is then transformed orthogonally to obtain a factor solution. To classify or evaluate the sample case using these factors as independent variables after the screening factor solution, the factors need to be scaled to give the factor score for each sample case. For the factor values can be established regression equation:

$$F = A^T R^{-1} Z, \tag{10}$$

where A^T represents the transpose of A, R is the correlation matrix of the original variable, and R^{-1} is the inverse of R.

Then the comprehensive score of each artist is M_i, which can be expressed as follows:

$$M_i = \frac{\sum_n a_i F_i}{\sum_n a_i}, \tag{11}$$

where $a_i = \frac{\lambda_i}{\sum_i \lambda_i}$ is the weight, and λ_i is the eigenvalue.

For use of factor analysis, we shall calculate R first and perform some checks to see whether the problem satisfies proper conditions for factor analysis. In our work, we perform **Bartlett's Test** to check for correlation between original values. Then we calculate the eigenvalue to decide the common factors, and decide the factor loading matrix for acquiring comprehensive score.

(2) The realization of directed network

We use **NetWorkX** based local programming to realize the directed network and visualize it. NetWorkX is a package of Python for dealing with complex network systems.

3.1.3 Results: The network and visualization of influence

(1) Factor Analysis:

We first present result of factor analysis. **Fig.1** visualizes the correlation between all factors. It is obvious that there are indeed some highly correlated factors:

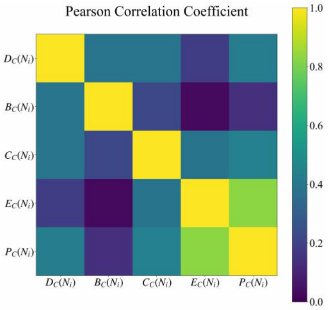

Figure 1: Pearson correlation of different factors

The result of Barrett's test (p-value) is 0.9634. Under common significance level ($\alpha = 0.05$), we infer that factor analysis works here. After using SPSS data processing system, we get a scree sheet (extract 3 common factors, data listed in the appendices) shown as **Fig.2**. It can been seen from the scree sheet that last two factors have relatively small eigenvalues (the sum of first three eigenvalues accounts for more than 85% of the sum of all the eigenvalues). So three factors are suitable for extraction.

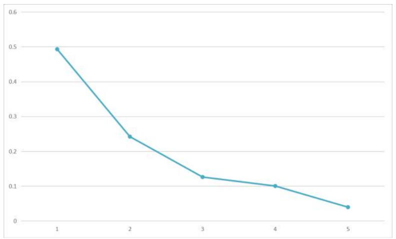

Figure 2: Scree sheet

Parameter	Factor 1	Factor 2	Factor 3
Degree centrality	0.649396	0.502796	0.081807
Betweenness centrality	0.385341	0.756722	0.000479
PageRank	0.708221	0.100189	-0.010477
Eigenvalue centrality	0.780923	-0.522564	0.234900
Closeness centrality	0.884825	-0.317558	-0.259180

Table 1: The rotated component matrix

Using SPSS data processing system, we obtain the rotated component matrix as shown in **Table.1**:

Extract the data, then we obtain:

$$\begin{cases} Z_1 = 0.649396F_1 + 0.385341F_2 + 0.708221F_3 + 0.780923F_4 + 0.884825F_5, \\ Z_2 = 0.502796F_1 + 0.756722F_2 + 0.100189F_3 - 0.522564F_4 - 0.317558F_5, \\ Z_3 = 0.081807F_1 + 0.000479F_2 - -0.010477F_3 + 0.234900F_4 - 0.259180F_5. \end{cases} \tag{12}$$

(2) The visualization of directed network

We then apply results to calculating comprehensive scores of artists' influence, and visualize our result. **Fig.3** visualizes the network we establish. In **Fig.3**, the bigger the circle, the more influential one particular artist will be.

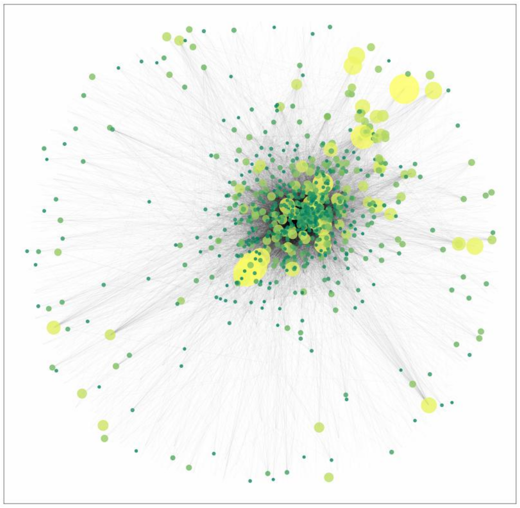

Figure 3: Visualization of directed network of influence

We have also obtained the network of musical influence categorized by year and genre, and visualize our result in **Fig.4**. All artists are identified with their IDs. In the network categorized by year, the darker color indicates that the artist was active in an earlier age, and the longer diameter indicates that the artist is more influential. In the network categorized by band, we use colors of circles to reflect genres.

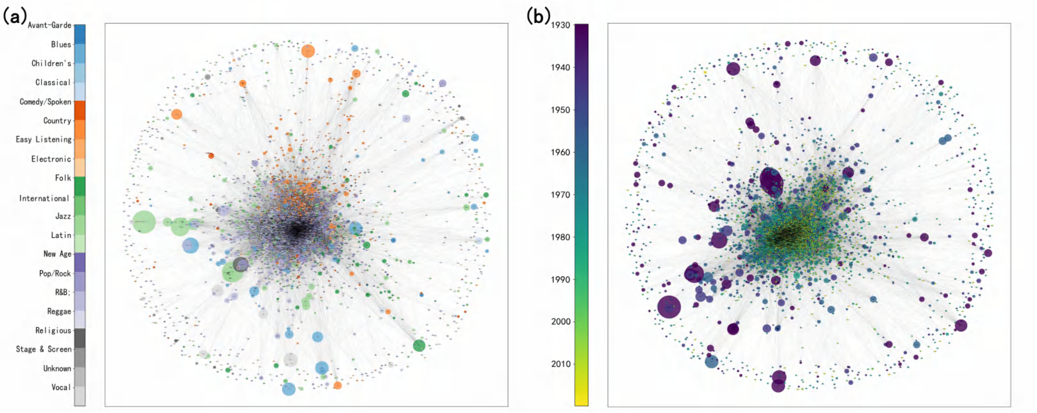

Figure 4: Visualization of network of musical influence, categorized by genre (**left**) and year (**right**)

3.1.4 The establishment and discussion of subnetwork

(1) Visualization:

We now use the data obtained from network to establish and visualize our subnetwork. From the data we have analyzed, we selected **4 artists with relatively high influence** characterized by comprehensive score and centrality.

Our selection of artists is **Cab Calloway (1930, the highest comprehensive score, eigenvalue centrality and PR value)**, **The Beatles (1960, the highest degree centrality)**, **Willie Nelson (1950, the highest betweenness centrality)**, **Bod Dylan (the highest closeness centrality)**. Comparison of their measures of influence is illustrated in **Fig.5**, with the vertical axis being standardized data of indexes measuring influence.

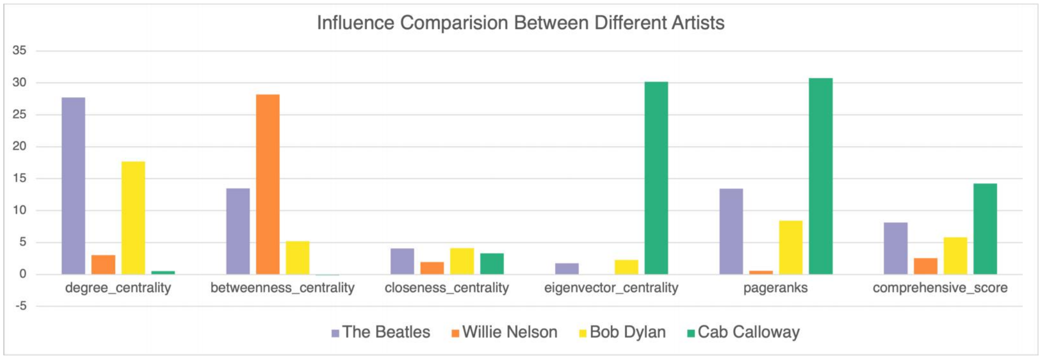

Figure 5: Comparison of measures of influence, selected artists

We visualize their subnetwork of influence as shown in **Fig.6**.

We can also locate their position in the network established beforehand, which is visualized in **Fig.7**.

(2) Description and discussion about the subnetwork:

In our subnetwork of influence, all directed edges describe the process of influence, while the diameter of the circle represents how influential an artist is.

(2.1) Cab Calloway:

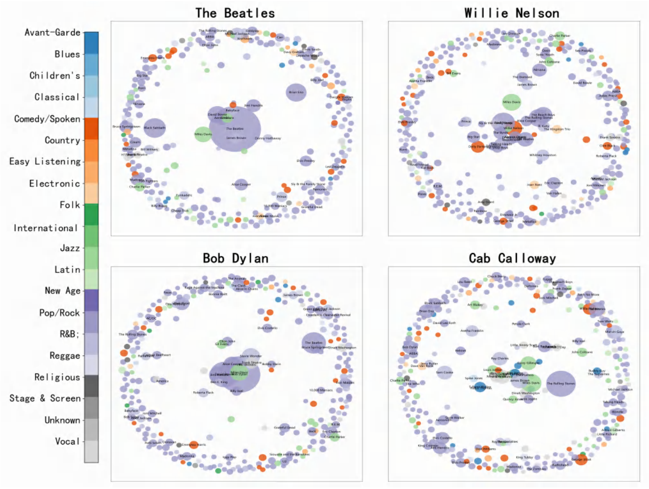

Figure 6: Visualization of subnetwork of influence, selected artists

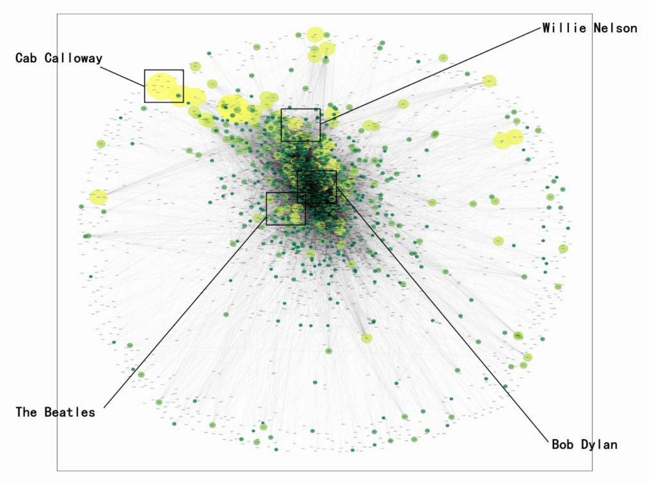

Figure 7: Position of selected artists, visualized in the network of influence

First we discuss the subnetwork of Cab Calloway, who is an influential artist active from 1930s and lies in the outer layer of our influence network. Due to his earlier active years, we record few inflow (relative low degree centrality), which makes it hard to identify him in our network. Nevertheless, he has a rather strong outflow that hugely influencing the subsequent artists, including the influential ones in the future. It is also interesting to see that most of his followers did not fully inherit his musical style, but added their understanding and formed their own. This is because Jazz is a relatively older genre of music which is not that influential in the future.

Our indicators do reflect all the possible information contained. A high PR value and comprehensive score reveals his importance in the network, while a relative low value of degree centrality combined his active years and position in the network reveal that he is **a founder inheriting less from history**.

(2.2) The Beatles:

Next we discuss the subnetwork of The Beatles, an influential band active from 1960s. The Beatles have the highest degree centrality and the second highest PR value and comprehensive score. Our measures of influence reveal that The Beatles inherit a lot from previous music, blend different genres and form their unique style, whose influence can be seen from the network (relatively high PR value and eigenvalue centrality reveals their power of influence, while the biggest degree centrality reveals their inheritance from previous music). Also, most of his followers are still in Pop/Rock genre, which may indicate that differences of genres have formed over the years.

3.2 Task 2: Measures of similarity and discussion about genres

In this part, we have developed a similarity measure that can be applied to comparing pieces, artists and the style of years by using data provided by *full_music_data*.

3.2.1 Model Preparation

We can view every piece of music as a point $C \in \mathbb{R}^n$, where n is the number of characteristics. Therefore, we can abstract the difference of music as **vectors** in $\mathbb{R}^n$. Due to difficulty in visualizing vectors, we can establish a scalar function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ to visualize the similarity or difference between music. One can easily come up with the definition of distance, which certain applies here. For this problem, we considering using the three types of distance function $d(A, B)$.

Since we are going to use linear models for this problem, it is essential to modify the weights of all indexes reflecting characteristics of music so that all dimensions share the equal weight. Thus, we use **Formula (1)** to standardize all indexes so all dimensions share a mean of 0 and a variance of 1.

(1) Choosing the Proper Indexes for Measuring characteristics:

There are 7 indexes that measure the characteristics of music, but we need to check possible correlation between all the indexes, in order to avoid collinearity. We apply factor analysis and compute the correlation matrix as shown in **Fig.8**. All Pearson correlation coefficients are presented with their absolute value.

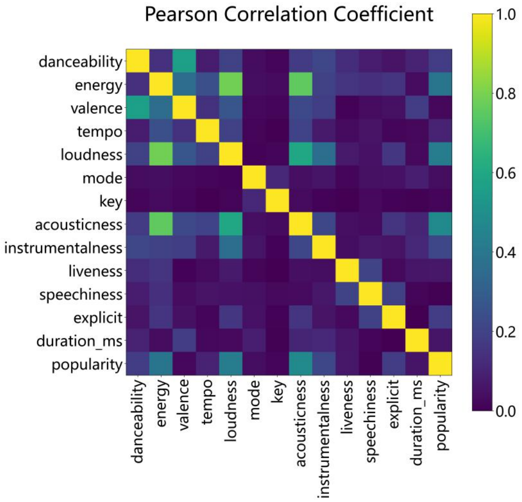

Figure 8: Correlation between all indexes measuring characteristics

Since there are some highly correlated indexes, we shall eliminate them to avoid collinearity. To find the independent variable that characterizes characteristics of music, we apply the **R-clustering**

method. We obtain the cluster dendrogram as shown in **Fig.9**, by means of **MATLAB programming**. We can conclude from the cluster dendrogram that we may eliminate **2 of the 3 indexes named energy, loudness and acousticness**, and **1 of the 2 indexes named danceability and valence**. As a result, we eliminate 3 of the 7 indexes, including loudness, acousticness and danceability.

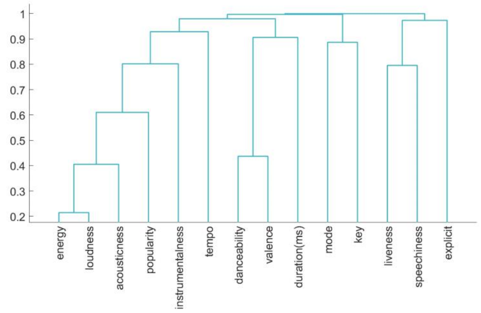

Figure 9: Cluster dendrogram of 7 indexes

(2) Measures of Similarity

(2.1) Euclid Distance:

Euclid definition of distance is the most common distance function defined in $\mathbb{R}^n$, which is easy to understand and accept as a measure of distance (or difference) between two pieces. Suppose there were n indexes that reflect characteristics of music, the distance can be defined as follows:

$$d(A,B) = (\sum_i (x_{iA} - x_{iB})^2)^{\frac{1}{2}}, \tag{13}$$

where $d(A,B)$ is the distance between point A and B, and x_{iA} and x_{iB} are the i^{th} index that reflects characteristics for A and B respectively.

(2.2) Manhattan Distance:

We can also use a modified Manhattan distance to characterize the similarity, which is originally defined as follows:

$$d(A,B) = \sum_i (x_{iA} - x_{iB}), \tag{14}$$

where the definitions of $d(A,B)$, x_{iA} and x_{iB} are the same as (2.1). For measures (2.1) and (2.2), $SIM(A,B) = 0$ indicates equivalence and the bigger $SIM(A,B)$ reflects bigger difference between genres.

(2.3) Cosine Similarity:

Another measure of similarity is **cosine similarity**, which measures the collinearity of two vectors. In $\mathbb{R}^n$, it is defined as follows:

$$SIM(A,B) = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{||\boldsymbol{A}|| \cdot ||\boldsymbol{B}||}, \tag{15}$$

where $SIM(A,B)$ is the measure function, $\boldsymbol{A}$ and $\boldsymbol{B}$ are two vectors representing point A and B respectively (starting point at origin), and $||\boldsymbol{A}||$ and $||\boldsymbol{B}||$ represent the 2-norm of vector $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively.

The similarity function defined uses the cosine value of angle between the vectors, which is a better measurement of similarity since it does not require standardization, and reflects similarity in a uniform manner. For the setting of this problem, we have $SIM(A, B) \in [0, 1]$, where $SIM(A, B) = 1$ indicates complete equivalence, and $SIM(A, B) = 0$ characterizes the maximum difference that can occur here.

The above three measurements should be combined when using, since cosine similarity only reflects the collinearity of two pieces, despite its stability and convenience of using. The distance measurement usually cannot evenly assign the same weight to all dimensions, but it can measure equivalence more precisely.

3.2.2 Model Establishment and Results

We now use the similarity functions defined in 3.2.1 to analyze the problem. The process can be explicitly stated as follows:

Consider the similarity measure between genre A and B, let U_A and U_B be the set containing all artists belonging to genre A and B respectively. Thus the similarity measure for genre A and B is calculated as below:

$$SIM(A, B) = \frac{\sum_{i \in U_A, j \in U_B} SIM(i, j)}{n_{A,B}}, \tag{16}$$

where $SIM(i, j)$ represents the similarity measure of piece $i \in U_A$ and piece $j \in U_B$. $n(A, B) = n_A n_B$ is the number of pair of artists involves in calculation, where n_A and n_B represent number of artists in genre A and B respectively. The process is equivalent to find the mean value of similarity function between genre A and genre B.

Have obtained all the similarity measures between and within genres (which can be written as a symmetric matrix), we calculate the average of all similarity measures within genres and all similarity measures between genres, which is expressed below:

$$\begin{cases} SIM_{IN} = \frac{\sum_i SIM(i,i)}{n_i}, \\ SIM_B = \frac{\sum_{i \neq j} SIM(i,j)}{n_{ij}}, \end{cases} \tag{17}$$

where SIM_{IN} and SIM_B represent similarity measurement within and between genres respectively, n_i is the number of genres, and $n_{ij} (i \neq j)$ is the number of different genre pairs.

From the data set *data_by_artist*, we obtain the similarity measure using cosine measurement. The result is visualized in **Fig.10**, where SIM 1, SIM 2 and SIM 3 represent similarity function calculated by means of Manhattan distance, Euclid distance and cosine similarity respectively.

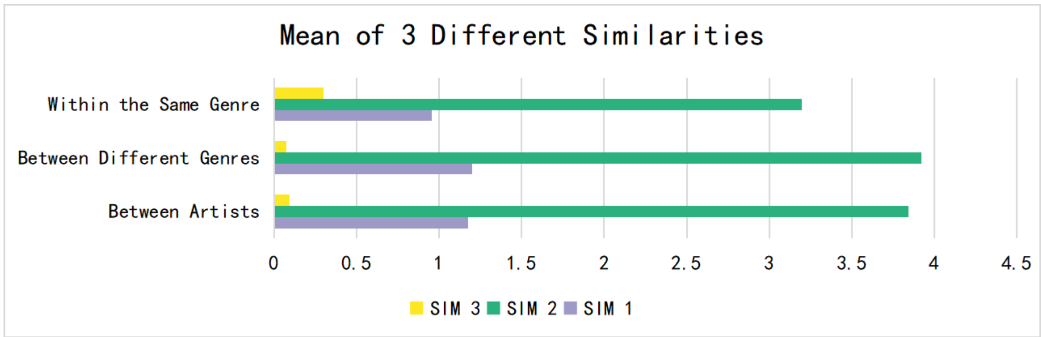

Figure 10: Visualization of similarity within and between genres

From **Fig.10**, we can clearly conclude that artists within one genre do appear similar compared to artists between genres. So the common assumption to be examined holds here.

3.3 Task 3: Discussion about genres: similarities, influences, changes and relations

We have developed a model of similarity (measure) and network of influence, and use an average to evaluate similarities within and between genres. Nevertheless, using a mean value to characterize similarity is somewhat rough and neglect some details that worthy of attention. For this part, we will further discuss the properties of genres, including their similarities, influence network, changes over time, and relations between certain genres.

3.3.1 Model preparation

(1) Evaluation of similarities and influences between genres:

We shall use an average similarity measure for characterization for examination of similarities between genres roughly, which can be calculated using **Formula (16)**:

$$SIM(A,B) = \frac{\sum_{i \in U_A, j \in U_B} SIM(i,j)}{n_{A,B}}.$$

But the method mentioned above is rough and neglects details. We apply an **agglomerative clustering algorithm** for further analysis in this problem. Agglomerative clustering starts with N clusters, which include only **one** object at first. We combine two clusters with the closest similarity each time, until we find a cluster that contains all data. Similarity between clusters can be measured by means of **average link clustering** defined as follows:

$$\overline{d(\mathcal{A},\mathcal{B})} = \frac{1}{n_G n_H} \sum_{i \in \mathcal{A}} \sum_{j \in \mathcal{B}} d_{ij}. \tag{18}$$

We apply **Scikit-learn** for this problem, which is an open-source package of machine learning for Python[1].

As for the evaluation of influences, we further develop the network we used in **Task 1** and build a double-layer network model to address the problem. Similar to the degree centrality we exploit as the measures of influence, we define two parameters named R and R' as follows, which represent the number of Influencer-Follower pairs within and between genres respectively:

$$\begin{cases} R(N_i)_{[A]} = \sum_{j=1}^{g} x_{ij} (i \neq j, N_i \; and \; N_j \; are \; both \; in \; genres \; A), \\ R'(N_i)_{[A,B]} = \sum_{j=1}^{g} x_{ij} (i \neq j, N_i \in A \; and \; N_j \in B), \end{cases} \tag{19}$$

where $R'_{[A,B]} = \sum_{N_i \in A} R'(N_i)_{[A,B]}$ indicates the influences genre A exerts on genre B. Like what we have done in **Task 1**, we will visualize our network and come up with a conclusion.

3.3.2 Model Establishment and Results

(1) Evaluation of similarity and relation between genres:

We first visualize the results of our rough algorithm, which uses average as an approximation. We graph our results in **Fig.11** using cosine similarity function ranging from $[0,1]$, where 1 means equivalence.

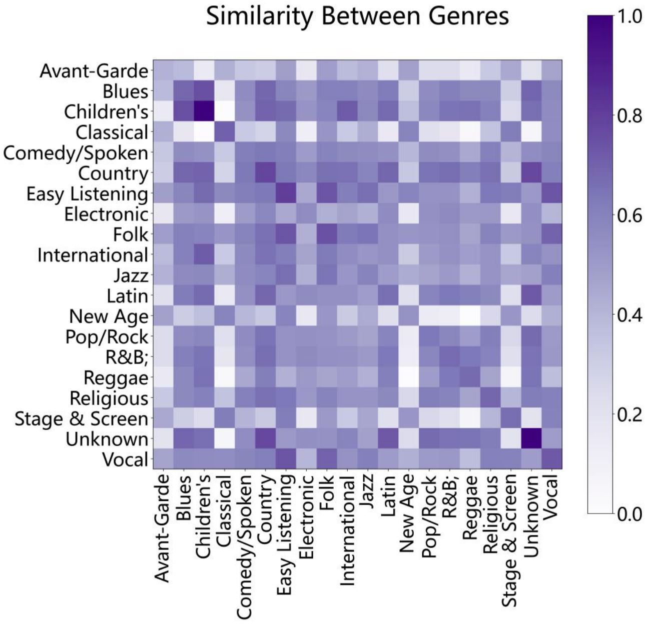

Figure 11: Similarity between genres, using cosine similarity function as a measure

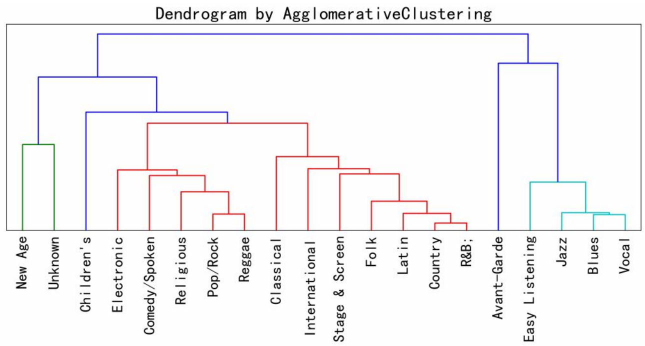

Figure 12: Dendrogram of genres

We can also graph the dendrogram by clustering algorithm, the result of which is shown in **Fig.12**.

We can category 19 genres mentioned in the data into 3 groups, with 3 special minorities (**including New Age, Avant-Garde and Children's**). We do not assign artists categorized as **Unknown**. The three groups are divided as follows, each sharing :

Group A: Electronic, Comedy/Spoken, Religious, Pop/Rock, Reggae.

Group B: Classical, International, Stage & Screen, Folk, Latin, Country, R&B.

Group C: Easy Listening, Jazz, Blues, Vocal.

Since we have a great variety of data and several indexes to analyze, it is essential for us to categorize all indexes reflecting characteristics. We have already obtained the correlation between variables using R-type clustering in **Task 1**, and weeded out 3 indexes reflecting characteristics. Also we notice that key, mode, and duration_ms may depend on the artists' active era, so we will visualize these 3 indexes separately. The visualization of data is presented in **Fig.13** as follows, where we finish a comparison of all indexes reflecting characteristics. All indexes we apply here have been standardized using **Formula**

(1).

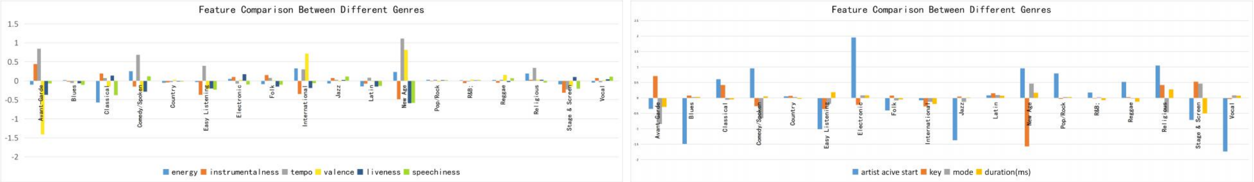

Figure 13: Visualization of data: comparison of indexes of characteristics

We can easily tell the differences between genres due to some of their outstanding characteristics. We set genre Country, which is close to **the center of feature set** as our reference frame. Below are two simple case studies.

I. Classical music. Energy and speechiness of classical music tend to be below average, while its instrumentalness is relatively high. Complicated techniques of musical expression, underlying thoughts and calm tone of classical music contribute to this quantitative expression of characteristics.

II. Avant-Garde. It has the highest instrumentalness index among all genres. Accordingly, it possesses the highest value of tempo and the lowest value of valence. This has something to do with its techniques which is so different from traditions.

Generally speaking, there always exist some characteristics that make genres stand out and be distinguished. Using a quantitative model and data visualization, we can observe their features intuitively.

(2) Evaluation of time-changing process of genres Here we conduct a case study about genre Pop/Rock and R&B, pointing out their relations and their time-changing process.

3.4 Task 5: Identifying music revolutions and revolutionary artists

Identifying music revolutions has been a tough problem for research into history of music. For our quantitative research, we cannot find a proper scalar function that can measure revolution clearly. Instead, we extract data from *data_by_year.csv* and learn about the changes occurring with regard to all indexes measuring characteristics. As for the identification of revolutionary artists, we establish a measure of revolution named **REV** considering a variety of aspects.

3.4.1 Model Preparation

Our goal is to discover a measure of revolutionary artists. Based on models we develop in **Task 1** and **Task 2**, we introduce **REV** to measure how revolutionary artists can be. We consider following 4 factors: Influence (IF, measured by degree centrality of nodes in our network of influence), similarity between influencers and followers ($\overline{SIM}$, taking an average of all SIM value between one particular influencer and all his followers), popularity of musician (Pop, extracted from the data set directly) and similarity between influencers and the year he was in (SIM_{trend}). For the simplicity of our model, REV is defined by the following multiplication:

$$REV = IF \cdot \overline{SIM} \cdot Pop \cdot SIM_{trend}. \tag{20}$$

In order to standardize all indexes and keep them positive, we introduce the following standardization, which is different from **Formula (1)**:

$$N = \frac{X - \bar{X}}{X_{max} - X_{min}}, \tag{21}$$

where N is the standardized data, X is the original data, $\bar{X}$, X_{max} and X_{min} are the average, maximum and minimum of the original data group.

3.4.2 Model Establishment and Solutions

(1) Identification of Revolutions in the history of music:

We can extract the data from *data_by_year.csv* and visualize the change of characteristics with time, which is shown in **Fig.14**. The horizontal axis represents year (numbered 1 to 100, starting from 1921), and the vertical axis represents the value of indexes reflecting characteristics.

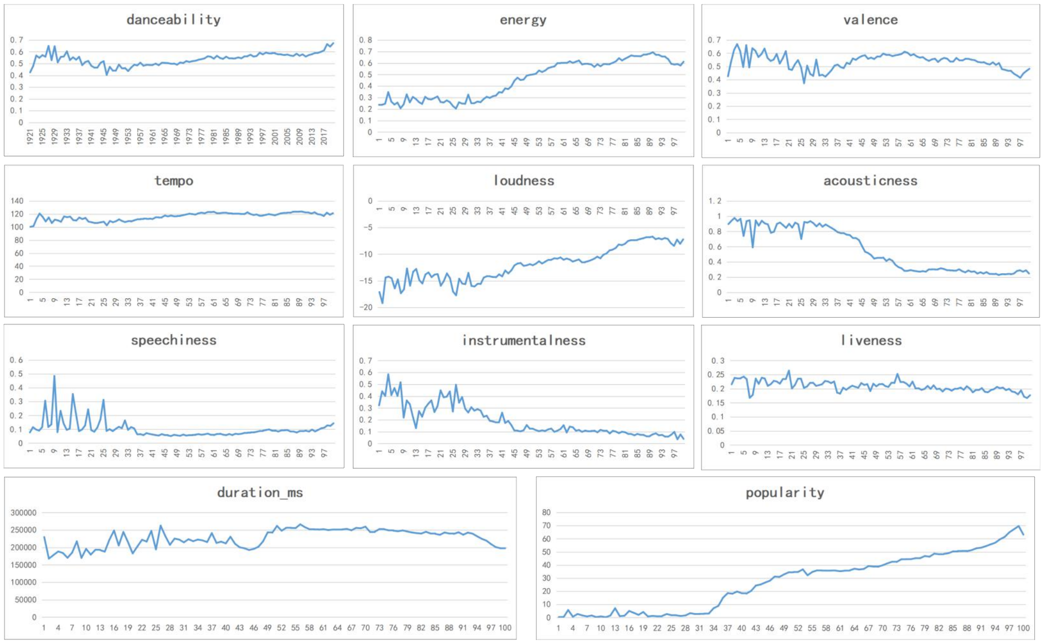

Figure 14: Changes of all indexes reflecting characteristics with year

By deeply looking into these changes, we can identify the possible revolutions and come up with factors that lead to their happening. Energy and loudness rises drastically in 1950s, and they experience relatively mild increase during the 1990s. The change of acousticness and instrumentalness is consistent with what we discover about energy and loudness. Also, popularity increases drastically during the 1950s, while liveness is steady throughout the years. All these indicators suggest that there were two major revolutions taking place, in the 1950s and 1990s respectively.

(2) Identification of Revolutionary Artists: We go on to explore revolutionary artists in our network, using the REV value we defined.

5 Reference

[1] Linton, C. Freeman. Centrality in social networks conceptual clarification[J]. Social Networks, 1978.
[2] HASTIE, T., TIBSHIRANI, R., & FRIEDMAN, J. H. (2001). *The elements of statistical learning: data mining, inference, and prediction*. New York, Springer.
[3] Pedregosa et al. *Scikit-learn: Machine Learning in Python*, JMLR 12, pp. 2825-2830, 2011.