Problem Chosen	2021	Team Control Number
D	MCM/ICM Summary Sheet	**2108783**

The research of musical influence based on the characteristic of a network

Summary

Music has numerous genres and forms. It is a meaningful and interesting work to study the evolution process and influence factors of music. In order to finish the work, we have established mathematical models and algorithms based on quantitative data.

Firstly, we define the elements of Musical Influence Network. A node represents a musician, who owns a music characteristic vector which contains musician s features. A link in the network is an edge from a follower to an influencer, which describes influence of the influencer on the follower. Then we use Modified PageRank Algorithm to calculate importance degree of a musician. Using the calculation results, we select some nodes in the network to build and visualize a subnetwork. Through the analysis of the subnetwork, we find the fact that minority of the musician has a great influence on others and there exists close connection between different clusters in network.

Secondly, we establish the Music Similarity Calculation System based on weighted Entropy Weight Method and Weighted Euclidean Distance. Through analysis of the meaning of original data, we select 11 features which describe the trait of music including danceability, energy, valence. Then we use the Entropy Weight Method to determine the weight of all 11 features and finally we use the reciprocal of the Weighted Euclidean distance as the similarity between two songs or two musicians.

Thirdly, applying our Musical Influence Network and Music Similarity Calculation System, we analyze the music influence between and within different music genres over time. For influence between different genres, popular music, like Pop/Rock and R&B, has an extremely strong influence on other genres. As for Internal influence, We find musicians of those most popular genres (like Jazz and Pop/Rock) tend to communicate within the genre, while those niche music tend to learn from the popular ones. Through calculating similarity, we do find musicians within genre are more similarity than musician between genres. To research the relationship and difference between genres, we use the Hierarchical Clustering Analysis to cluster the genres into 4 groups, then compare difference of every characteristic, and find speechness and instrumentalness are the key features to distinguish a genre.

Finally, we analyze the change of whole music world over years, and we find 3 music revolutions which really have happened and change the music world. And we update our PageRank Algorithm to Dynamic PageRank Algorithm by taking time and genre into consideration. In the result of New model, ranking of dynamic influencers does improve. We also study the music genres.

Keywords: Modified PageRank Algorithm, Music Similarity Calculation System, Entropy Weight Method, Weighted Euclidean Distance, Dynamic PageRank Algorithm.

Contents

1 Introduction

1.1 Problem Background

Music is an art form that has existed since the embryonic stage of human civilization. It has a long history and rich cultural connotations. It is a treasure of human civilization. Since the 20th century, popular music has become the mainstream of music, numerous of excellent works have emerged, and the genres of music have also increased a lot, greatly enriching peoples spiritual world.

We want to understand the evolution of music and the factors that influence the evolution of music. However, many of the existing analysises are subjective, explained by music lore, and lack rigorous tests based on quantitative data and statistics. So we plan to build a model to quantify the evolution of music and the influence factors.

1.2 Problem Restatement

- Establish a directed network of musical influence between influencers and followers, and describe musical influences by developing parameters and exploring subsets.
- Establish a measure of music similarity, and use parameters to describe the similarity between artists within and between music genres.
- Compare the similarities within and between music genres; discover the characteristics of different music genres, changes in the types of music, and identify whether music genres are related to each other.
- Identify whether influencers truly have an impact on followers, and whether there are some musical characteristics that are more contagious.
- Look for the characteristics of the musical revolution in the data and determine which artists can represent the revolutionaries.
- Analyze the evolution and influence process of a music genre; establishing indicators;find dynamic influencers, and explain changes in music genres and artists
- Explain the cultural influence of music; find out the social, political, and technological changes reflected in the network.

1.3 Our Work

In order to measure musical influence, we have completed the following tasks:

- We use a directed network to describe the influence of music and establish a PageRank Algorithm to weigh the influence between influencers and followers. Then we create a subnetwork in order to display our model and calculate the influence between genres at the same time.
- We create a calculation system to calculate music similarity based on original data by using Entropy Weight Method and Weighted Euclidean Distance. In this model, we use similarities to find out the similarity and difference between

and within genres. Also, through Hierarchical Clustering we find out the most distinguished factors to classify genre and the relations between them as well.

- Based on the two main models, we establish a theory Ideal Admirer to explore the influence of influencers on followers. We use similarities to study revolutions and evolutions in music and improve PR model to find dynamic influencers. At last, we connect our findings with cultural influence and other realities.

2 Assumption and Symbol Explanation

2.1 Basic Assumption

- The data are representative. The data is very large and the data sources are reliable, which means the data we received can reflect the real situation.
- The followers only have the influencers shown in data sheet. Followers' influencers are stated by themselves, so these influencers must have a great influence on them while other influencers with small influence can be ignored.
- Assume that the influence of influencer can see through followers music indicators.

2.2 Symbols and Definition

Definitions of symbols employed in this paper are listed in Table1.

Table 1: Notations

Symbols	Description
I	Influencers in the directed network
F	Followers in the directed network
PR	PageRank Value
N	Number of followers/ connected nodes
L	Out-degree
σ	The ratio of the influence
IF	Influence of the influencer on the follower
$IFSP$	The proportion of the influence between genres
ω_i	The weight of different music features
Similarity	The measure of similarity
$V_i^{(j)}$	i_{th} feature vector
D_j	The measure of difference between features
PCS	The percentage of similarity change
θ_{time}	Time dynamic index
θ_{genre}	Genre dynamic index
α	The coefficient of Time dynamic index
β	The coefficient of Genre dynamic index

3 Data Preprocessing

Before we present our mathematical model, for the sake of data accuracy and reliability, we have checked the data provided, and performed the following data preprocessing on the data:

- In the *influence_data* data set, through data filtering and matching, we found that the same *influencer_id* or *follower_id* corresponds to multiple *influencer_name* or *follower_name*. We find that the IDs with this situation are 957340, 893383, and 303506, as shown in the table2, the name in red is correct

Table 2: Errors in Data

influencer_id / follower_id	*influencer_name / follower_name*
957340	Helloween/ Nightwish
893383	The Muffs/ The Kills
303506	Usger/ Day26

- We find that artists with the same influencer_name or follower_name have different influencer_id or follower_id.In order not to affect the subsequent mathematical model we develop, we add serial numbers at the end of these *influencer_name* or follower_name.
- To facilitate subsequent data processing, we set IDs for the newly added genres in the *full_music_data* data set. The specific IDs is shown in the table3:

Table 3: Genre ID

genre	genre_id
Avant-Garde	1
Blues	2
Children's	3
Classical	4
Comedy/Spoken	5
Country	6
Easy Listening	7
Electronic	8
Folk	9
International	10
Jazz	11
Latin	12
New Age	13
Pop/Rock	14
R&B	15
Reggae	16
Religious	17
Stage & Screen	18
Unknown	19
Vocal	20

- When we develop measures of music similarity, because the dimensions of Characteristics of music and Types of vocals are different, it is not conducive for us to develop measures of music similarity, so we use Z-score Normalization to normalize the data of Characteristics of music and Type of vocals data.

  Z-score Normalization normalize the mean and standard deviation of the original data. The processed data conforms to the standardized normal distribution, the following is the normalization function:

$$x_{normalization} = \frac{x - \mu}{\sigma} \tag{1}$$

4 The Directed Networks of Musical Influence

In our musical influence directed network:

- Node: A musician, who could be an influencer or a follower. Each node has a characteristic vector V which describes scores of the musician in danceability, energy, valence, tempo, loudness, acousticness, instrumentalness, liveness, speechiness, mode, year.
- Link: An edge from follower to influencer, which represent the relationship of influence between two musician.
- NodeWeight: an indicator which measures the influence of a musician. In our later model, we will use PageRank Algorithm to calculate the value, so in this artcle, we call value of NodeWeight of a Node with PR.
- LinkWeight: an indicator which measures the vitality degree of influence of the influencer on the follower. In our later Dynamic PageRank model, we will define how this value is calculated.

4.1 The Modified PageRank Algorithm

In order to use the *influence_data* data set to create directed networks of musical influence and develop parameters that capture music influence in this network, we develop a Musical Influence Evaluation Model based on modified PageRank Algorithm.

PageRank (PR) is an algorithm used by Google Search to rank web pages in their search engine results. PageRank is an effective way of measuring the importance of website pages by calculating the PR value. The PR value of a page is obtained through iteration algorithm based on the importance of all the pages linked to it (linked pages). A page with more links will have a higher PR value. The larger the PR value is, the more linked pages of the page, and the greater the importance of the page denotes. Because the influence of music is a directed network from influencer to followers, similar to the characteristics of web pages. So we develop the Musical Influence Evaluation Model based on the PageRank Algorithm.

Because the number of outbound links can be 0, which means those webs do not link to any other web pages (isolated web pages). Therefore, the PageRank formula needs to be revised, so the damping factor q is added to the formula, and q generally takes the value $q = 0.85$.

Its practical significance is the probability that, at any time, the user will continue to browse the next page after reaching a certain page. $1-q=0.15$ is the probability that the user stops clicking and jumps to a new URL randomly. From this we get the formula of PageRank:

$$\mathrm{PR}\ (I)=\frac{1-q}{N}+q\sum_{F_i}\frac{\mathrm{PR}\ (F_i)}{L\,(F_i)} \tag{2}$$

where F_i represent all the nodes that point to node I in a network. q is a constant. In this case, F_i represents followers, I represents influencers, $\mathrm{PR}\ (F_i)$ denotes The PageRank value of follower, and $\mathrm{PR}\ (I)$ denotes The PageRank value of influencer.

However, the relationship between influencer and followers is different from the relationship between web pages. When browsing the web, there is a high probability that the user will continue to brows the next web page, so q generally takes the value $q=0.85$, but in the musical influence, an artist can become the influencer of other artists, but the probability of the followers of the influencer become the influencers of other people is not that high, so we define the value of q as 0.3 .

Through the iterative calculation of the *influence_data* data set, we get the PR value. Due to the large amount of data, we select the top 20 artists for display. The data is shown in Table4:

Table 4: The Rank of musical influence by the PR value

Ranking	*artist _ name*	$\mathrm{PR}\ (I)$	Ranking	*artist _ name*	$\mathrm{PR}\ (I)$
1	The Beatles	0.004545	11	Chuck Berry	0.001452
2	Bob Dylan	0.002687	12	Jimi Hendrix	0.001451
3	The Rolling Stones	0.001948	13	Elvis Presley	0.001435
4	Hank Williams	0.001770	14	James Brown	0.001434
5	Louis Jordan	0.001754	15	Woody Guthrie	0.001418
6	Billie Holiday	0.001737	16	Sam Cooke	0.001370
7	Lester Young	0.001526	17	Miles Davis	0.001356
8	Cab Calloway	0.001506	18	The Kinks	0.001278
9	David Bowie	0.001489	19	Led Zeppelin	0.001261
10	Roy Acuff	0.001475	20	Muddy Waters	0.001259

4.2 Visualization of the Directed Network

By using the Modified PageRank Algorithm, we get the PR value of the artists, which reflects their influence. In order to make the directed network we build more intuitive, we make the visualization of the directed network.

In our case, artists are viewed as nodes and the influence as links among them, the links are composed of directed line segments. The thickness of the directed line segment represents the influence of the influencer on his\her follower, and the size of the node represents the influence of the artist. Both of the influence is measured by the corresponding PR value. Since a follower usually has multiple influencers, and the influence of the influencers is different, so the influence on the follower should be different. Therefore we use the percentage of the PR value of an influencer in the sum of PR values of all influencers to measure the influence of this influencer on followers. We can calculate the influence $\sigma(k)$ of influencer I_k to follower F by the formula:

$$\sigma(k) = \frac{\mathrm{PR}\ (I_k)}{\sum_{i=1}^{n} \mathrm{PR}\ (I_i)}(k = 1, 2 \cdots, n) \tag{3}$$

$$IF\,(I_k) = \sigma_k \times PR(F) \tag{4}$$

Using the above method, we use networkx package in Python to calculate importance degree measured by IF of all influences in this musical influence network.

In our network, there are 5603 Nodes and nearly 42000 links. In our visualization process, considering the observability, we choose three artists with greater influence (Bob Dylan, The Beatle, and Cab Calloway) to create a subnetwork. Then We add 10 random followers of each of them to the subnetwork and the same for newcomers. Repeat the process for three times and we create this subnetwork.

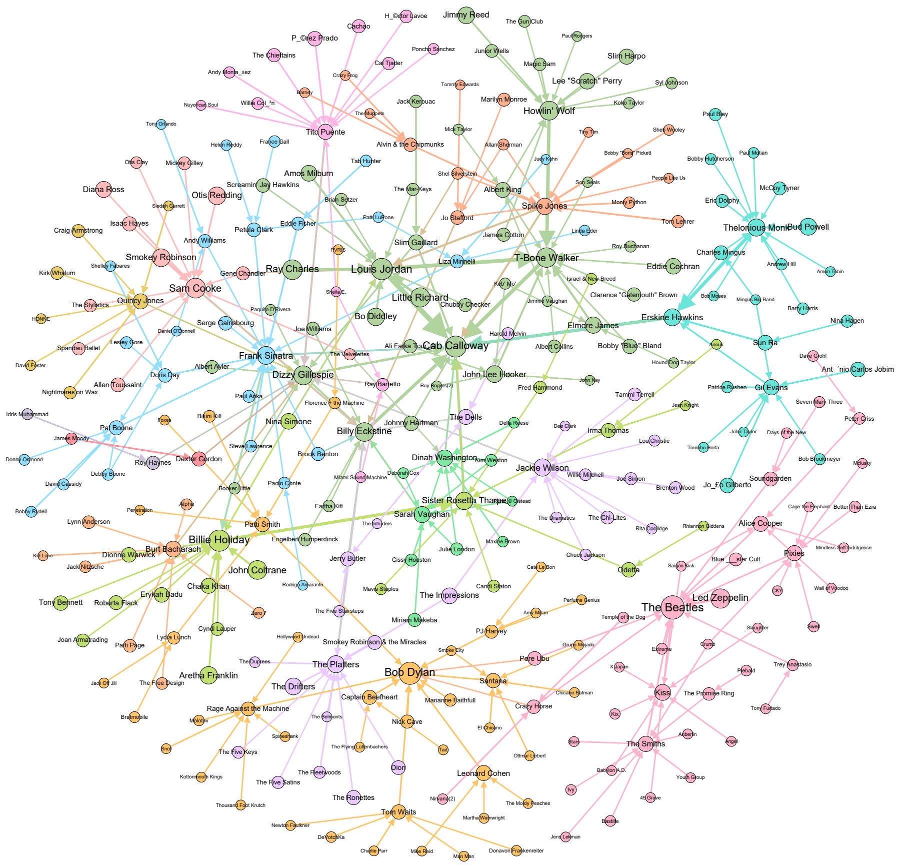

Figure 1: The Directed Network

In the network, a musician who owns bigger Node is more important in music history. A Link wider represents bigger influence. So, its easy to see that there are some artists Node is significantly bigger than others, That is the minority in the network has a great influence on others. From the perspective of clustering, there appear several clusters which have been colored, but the network is still closely connected and communication between clusters if frequent which may represent the communication between the different music genres.

4.3 Influence Between and Within Genres

In order to compare the similiarities and influences between and within genres. We can improve the $IF(I_k)$ function established above, calculating $IFSP$ to represent the proportion of the influence within genre in all the influence, then $1 - IFSP$ can represent the proportion of the influence between genres in all the influence.

Figure2 is the colum diagram drawn based on the calculated $IFSP$:

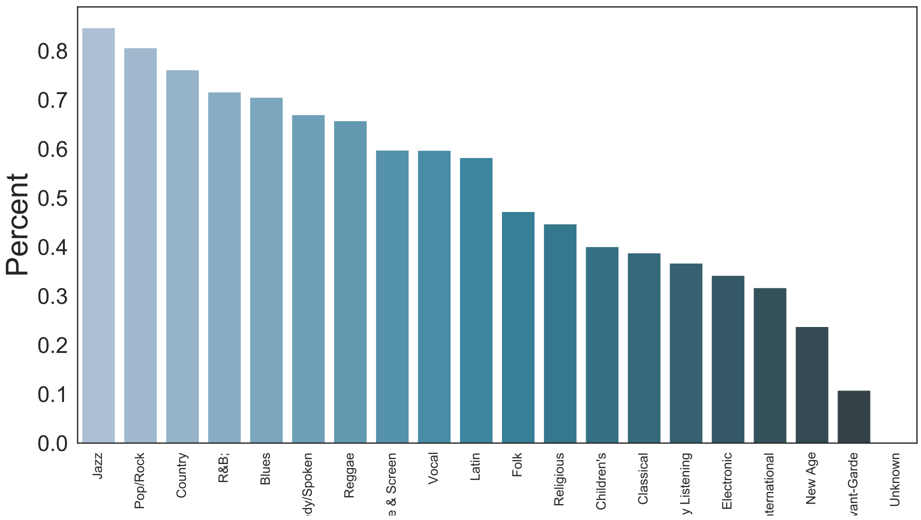

Figure 2: $IFSP$ of Different Music Genres

From figure2, we can find that most mainstream music genres, such as Jazz, Pop/Rock, Country, etc. are affected by more than 70% of their own music genres. While musics such as Electronic, New Age, Avant-Grade, etc., which are relatively niche music genres is less than 40% affected by their own music genre.

We think that mainstream music genres have numerous listeners and long history, which have formed their own unique and styles, so it is mainly influenced by its own music genre; and relatively niche music genres generally have not formed its own stable style. So it will be greatly affected by other genres.

5 Measures of Music Similarity

5.1 Feature Observation

In order to develop a measure of music similarity, we draw frequency histograms for the Characteristics of music and Type of vocals of different music genres. From the frequency histograms, we find that in some features, there are significant differences between music genres; at the same time, in some features, the differences between music geners are very small. For example: in acousticness and liveness, there are huge differences between music genres, while there are small differences in speechness and loudness. The corresponding four frequency histograms are shown in Figure3:

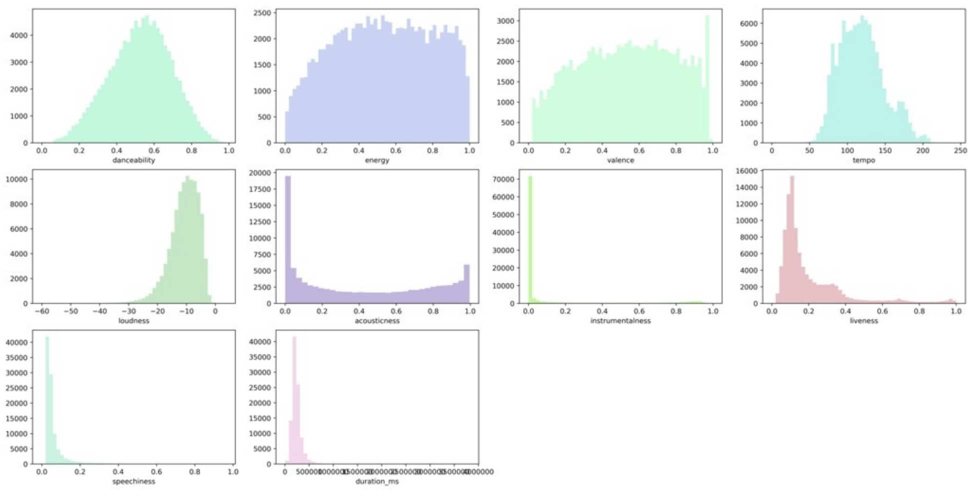

Figure 3: Music Features

Besides that, there are 3 discrete indicators: mode, explicit, key. Among 13 different indicators, we decide to abandon two of them explicit and duration_ms because they cant indicate more useful information about the characteristic of music.

5.2 The Entropy Weight Method (EWM)

Therefore, when develop measures of music similarity, different Characteristics of music and Types of vocal have different importance, so we have to weight different music features. Since we plan to use Eucledian Distance as measures of music similarity, we refer to the OneHot method which allows the representation of categorical data to be more expressive for machines, so that Eucledian Distance can be better used as measures of music similarity.

In information theory, entropy is a measure of uncertainty. The greater the uncertainty is, the greater the entropy are and the greater the amount of information contained. According to the characteristics of entropy, the randomness and disorder degree of an event can be judged by calculating the entropy value, and the degree of dispersion of an index can also be judged by the entropy value. The greater the degree of dispersion of the index is, the greater the influence or waight of the index on the comprehensive evaluation is.

When develop measures of music similarity, different Characteristics of music and Types of vocal have different importance, so we have to weight different music features. Therefore we plan to use Eucledian Distance as measures of music similar- ity to calculate the weights of different in dicators. The principle is shown in formula:

$$E_j = -\frac{1}{\ln n} \sum_{i=1}^{n} p_{ij} \ln p_{ij} \tag{5}$$

$$\omega_i = \frac{1 - E_i}{k - \sum E_i} (i = 1, 2, \ldots, k) \tag{6}$$

p_{ij} denotes frequency, E denotes Information Entropy, $omega$ denotes weights.

The calculated weights and feature_id are shown in the table5:

Table 5: Weights of Music Features

feature_id	feature	weight
1	danceability	0.010440536
2	energy	0.717223433
3	valence	0.042845599
4	tempo	0.004524814
5	loudness	0.010937639
6	acousticness	0.094040133
7	instrumentalness	0.050451853
8	liveness	0.026861306
9	speechiness	0.042674688
10	mode	0.469387891
11	key	2.384430156

5.3 Weighted Euclidean Distance

We have already used Z-score Normalization to normalize the data of Characteristics of music and Types of vocals data, so we can use the data normalized to calculate the Euclidean Distance between different music. We view the value of a set of characteristics of I, F, V is a high dimensional vector so that the Euclidean Distance of two vectors indicates the difference of their music.

Due to different features of discrete variables , which cant be easily added or subtracted, we refer to the OneHot method which allows the representation of categorical data to be more expressive for machines.

Let $\omega_1, \omega_2, \cdots, \omega_{11}$ denote the weight of different music features, $n = 1, 2, \cdots, 11$ denote the feature_id, and X, Y can denotes different songs or different artists or different music genres, so the similarity between X and Y can be measured by Euclidean Distance $d(X, Y)$:

$$d(X, Y) = \sqrt{\omega_1 (x_1 - y_1)^2 + \omega_2 (x_2 - y_2)^2 + \cdots + \omega_{11} (x_{11} - y_{11})^2} \tag{7}$$

5.4 Similarity Comparison Between Artists

Because the smaller the Euclidean Distance is, the higher the similarity between X and Y is. In order to make the data easier to observe and analysis, we perform MMS processing on Euclidean Distance, and we create a new function Similarity(X, Y) as the measure of similarity. The functional formula is as follows:

$$\text{Similarity}(X, Y) = \frac{1}{d(X, Y)} \tag{8}$$

The bigger the Similarity(X, Y) is, the higher the similarity between X and Y is.

In order to use this measure to compare the similarity of artissts within and between genres. We select the data of Pop/Rock(genre_id=14) as Pop/Rock genre has the most samples and is the most representative genre. We calculate the geometric center value of all the data in the Pop/Rock genre, and calculate the similarity between each artist in the Pop/Rock genre and the geometric center value of Pop/Rock genre.

At the same time, we calculate the geometric center value of all the data of all music genres, and then calculate the similarity between each artist in the popular music genre and the geometric center value of all music genres.

We have drawn the frequency histograms of the two sets of similarity:

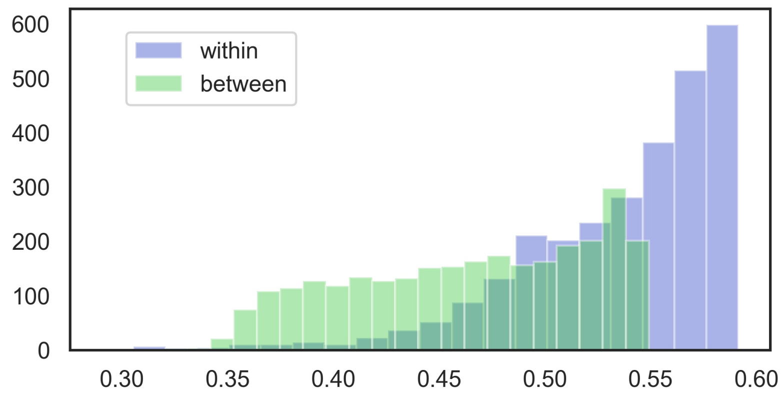

Figure 4: Frequency Histograms of Similarity

From Figure4, we can see that the similarity of artists within genre is obviously bigger than the similarity between genres. The mean value of the similarity between each artist in the popular music genre and the geometric center value of all music genre is 0.5365278088979288,The mean value of the similarity between each artist in the popular music genre and the geometric center value of all music genres is 0.4652363433349869, The results are consistent with the above conclusion.

5.5 Similarities Between and Within Genres

We calculate the means of the Characteristics of music and Types of vocals corresponding to each music genre as the vec-tor representing the overall features of this music genre.

Then, we calculate the similarities between artists in each genres as well as the similarities between genre based on mean values. We visualize the calculation results by Heat Map, and the results are shown in the figure5.

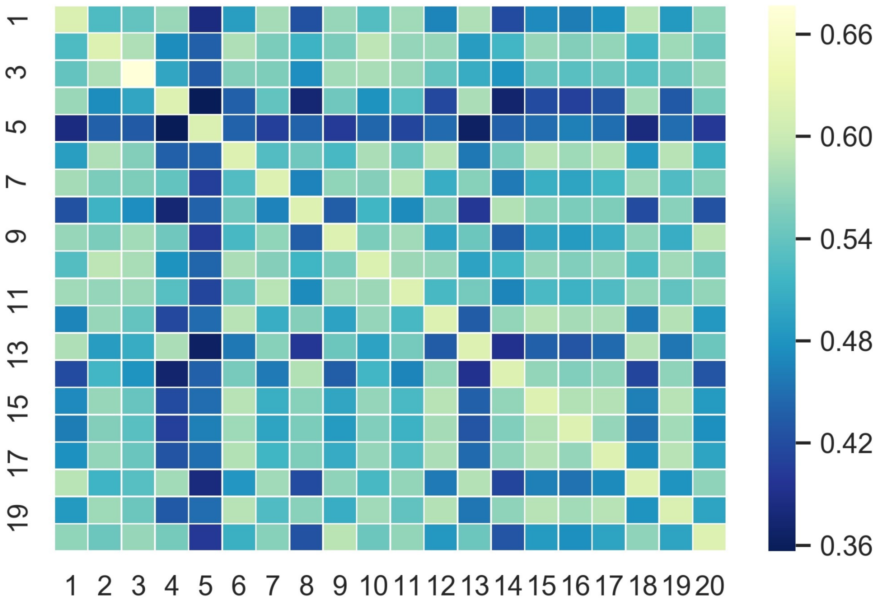

Figure 5: Heat Map of Similarities

From figure5, we can clearly and intuitively see the similarities within genres is much bigger than the similarities between genres, which is in accord with our calculation results.

6 Analysis on Classifications and Relations Among Music Genres

In order to further explore the differences and relationships between different music genres, we decide to use Hierarchical Clustering Analysis to gather similar music genres and separate the significantly different genres.

6.1 Cluster Model Establishment

Cluster analysis is often used to reasonably merge and classify the research objects according to given similarity indexes. Cluster analysis is expected to capture the nature structure of data for better understanding data and facilitating further research.

For its simplicity and flexibility, we select Hierarchical Clustering Analysis from various clustering algorithms. Hierarchical Clustering Analysis is a kind of distance-based cluster analysis method. By calculating the distance between sample points or brands and selecting the closest two points or groups to aggregate into a new group until the termination condition is reached, Hierarchical Clustering Analysis brings similar samples into groups.

According to principles of Hierarchical Clustering Analysis, we need to give (1) definition of distance between samples; (2) definition of distance of different groups. To be consistent with musical similarity measure we have established in section 5, the distance between samples X and Y is defined as $d(X, Y)$

To search the marked characteristics that distinguish music genres, we need to maximize the difference between different groups, so we define the distance between two groups of sample by Complete Linkage, which defines the distance between two brands as the maximum of distances between sample points of two brands as: $\mathrm{Max}\left\{V_i^{(j)}\right\}$

By using sklearn package in Python, we classify all music genres into 4 groups. The results are shown in Table6.

Table 6: Groups

Classifications of Groups
R&B, Pop/Rock, Electronic, Latin, Country,Religious, Reggae
Comedy/Spoken
Jazz, Easy Listening, Vocal, Folk, International, Blues, Children's
Stage&Screen, Avant-Grade, New Age, Classicle

In order to make the clustering results more intuitive, we used Principal Component Analysis (PCA) to reduce the dimensionality of the data to display the clustering results in the Figure6:

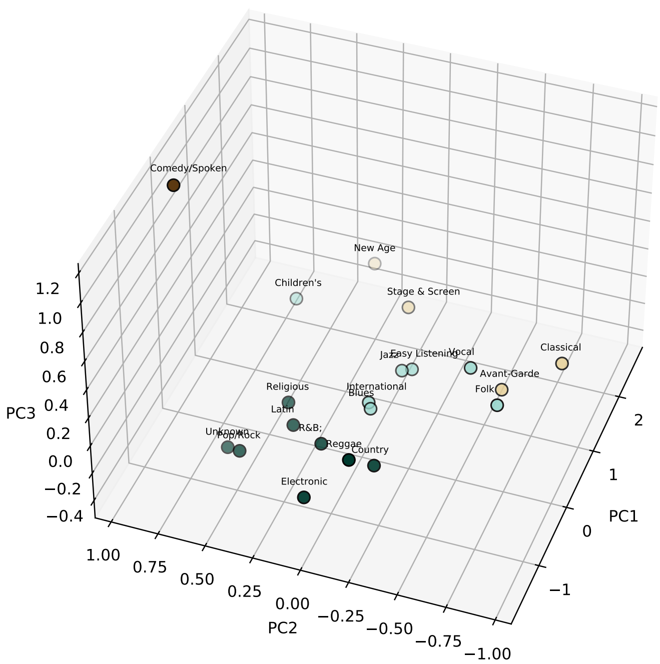

Figure 6: Clustering Results

6.2 Relationship Between the Genres

Comedy/Spoken is grouped individually, which is consistent with our common cognition that Comedy/Spoken is a so unique form that we rarely classify as a kind of music. New Age is a new music genre that combines the features of Pop and Classical, so its being grouped with Classical reflects the effectiveness of the cluster algorithm to some extent. The factor that R&B Pop/Rock Reggae are also grouped together also verified our observations on the heatmap. In general, the classification results are good. So we believe there exists a closer relationship between the members of the same groups.

6.3 What Distinguish a Genre?

In Hierarchical Clustering Analysis, we choose the *Complete Linkage* to calculate the distance between two brands to maximize the differences between groups. To search characteristics that distinguish genres, we replace the four groups with geometric centers of each groups, then calculate the difference of each characteristic in four group as:

$$D_j = \frac{\mathrm{Max}\left(V_i^{(j)}\right) - M_{in}\left(V_i^{(j)}\right)}{\mathrm{Min}\left(V_i^{(j)}\right)}(i = 1, 2 \cdots, n) \tag{9}$$

Table 7: D_j Values

Features	D_j
acousticness	1.414725
danceability	0.594540
energy	1.390823
instrumentalness	16.288928
key	0.232365
kind	1.500000
liveness	2.525010
loudness	-0.497463
mode	0.106799
popularity	0.515284
speechiness	9.101242
tempo	0.125223
valence	1.252960

By observing the result, we find that in instrumentalness and speechiness two characteristics gap between those four groups is significant. We believe the instrumentalness and speechness are the key features to classify the genres.

7 Comprehensive Application

In this section, we apply the modes established above to further explore the relationships between influencers and followers, as well as analyze the evolution and revolution of music.

7.1 Exploration on the Influences of Influencers on Followers

To discuss whether identified influencers can influence respective artists in practical terms or not, we decide to weigh the changes of followers style of works compared to their influencers.

When considering the influence between people, we find that the best to measure whether a person has an influence on others is whether others will learn and imitate from this person. In this way, we use the percentage change of similarity between the earliest and the latest music style of followers and music style of influencers as the measurement index PCS. The higher value of PCS indicates higher effect the influencer has on his follower:

$$PCS = \frac{\text{Similarity }(I, F)_{\text{latest}} - \text{Similarity }(I, F)_{\text{earliest}}}{\text{Similarity }(I, F)_{\text{earliest}}} \tag{10}$$

However, not every follower has only one person to admire, and the degree of his worship is basically different as well. In order to make the results more precise, we put forward a theory called Ideal Admirer. We assume that the influence of the i-th influencer on a person accounts for w_i of his all influencers, so that we can illustrate can portray his ideal admirer to everyone:

$$I_{\text{ideal}} = \sum \sigma_i I_i \tag{11}$$

In Section Two, we have calculated w_i for each influencer of each follower through PR model. Applying I_ideal as I in PCS, we calculate all the change rate of similarity of followers. The distribution of PCS is as follows:

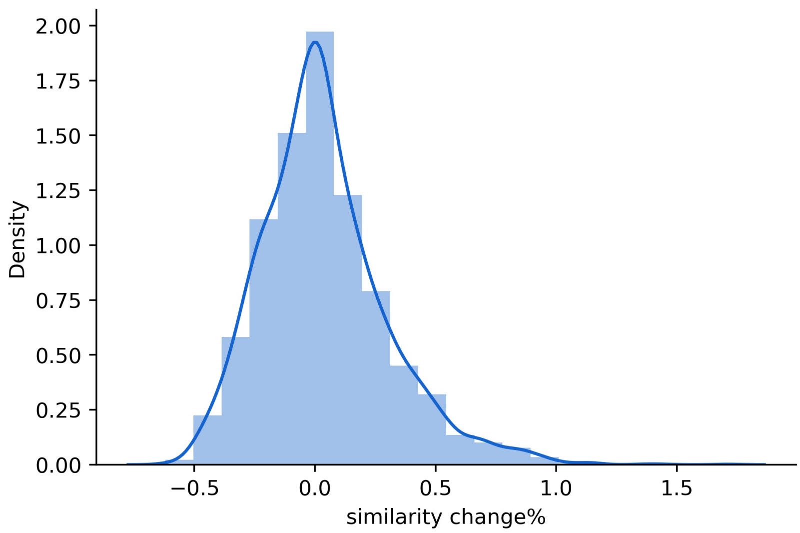

Figure 7: The Distribution of PCS

The area of the part of a distribution map greater than 0 takes up 52.4% of total area. We can see that the numbers of people being more similar with their followers is nearly equal to that of people who not, which shows that the influencers dont necessarily influence the creation of music. This result is also corresponded to our understanding that, though a lot of artists follow their idol, there are still some artists change their style to find more appropriate way to continue composition after imitating their idol at first.

After that, we want to find the factors that can influence these followers from influencers. Therefore, we calculate all the percent changes of 9 indicators of influencers and get the mean of each of them (as mode and key are discrete, we dont consider about them):

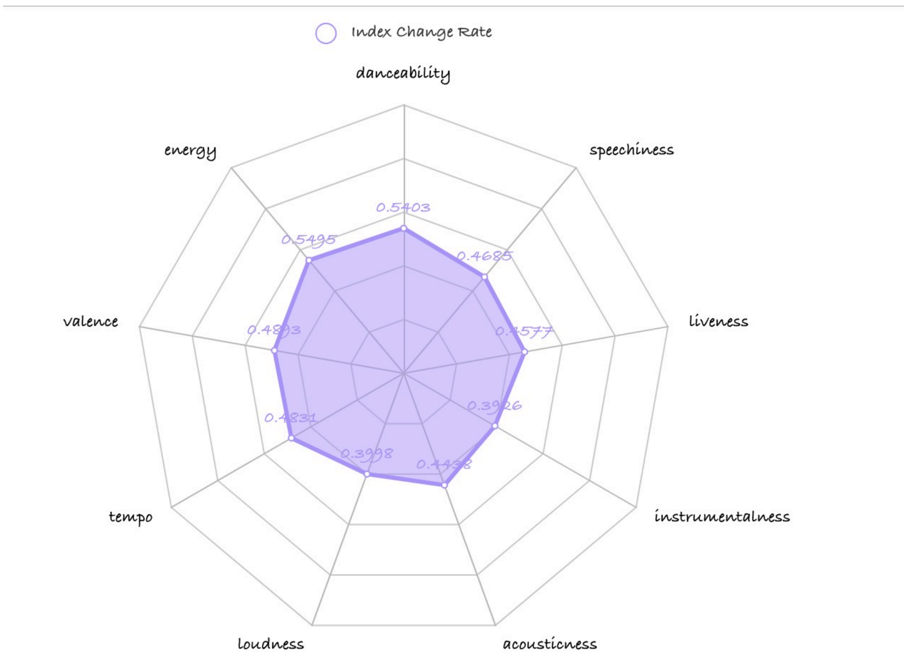

Figure 8: Changes of 9 Indicators

Through the Picture8, it can be clearly observed that danceability, energy and valence are changed most in personal music style. For this reason, we consider these three indicators are the most contagious.

7.2 Revolutions in Musical Evolution

We think Music Revolution refers to a huge change in music trends. We think that the huge changes in music trends can be expressed by the changes in music similarity between years, which can represented by $d(X, Y)$ over a period of time. So we can calculate the means of music features for each year, and calculated the $d(X, Y)$ value with the means of music features in the last year. In order to facilitate our observation, we have drawn a line chart to describe the changes in music similarity between years:

From figure ,we can find that there are huge changes during 1950s, 1960s and 1980.

In history, in 1950s, American music became more diverse. Rock music developed from Blues and other music genres was popular all over the USA, and jazz music took a back seat.

1960s is often claimed as 'British Invasion'. On 26 December 1963, The Beatles released *I want to hold your hand* in the USA. They were swiftly followed by dozens of British acts who, over the next few years.

In the 1980s, Rhythm and Blues were re-used in the form of R&B, and this usage has continued to the present. When Prince and dance-based pop stars like Michael Jackson and Madonna became popular, contemporary R&B was born.

Since The Beatles was the most representative band in the 1960s, we calculate the changes in music similarity between music features of years and The Beatles. The results are shown in Table8:

Table 8: Similarity Between The Beatles and All the Music1e

Years	Similarity values
1960	0.5215879291500649
1961	0.5322281256861955
1962	0.5546338974725465
1963	0.5577132020409185
1964	0.5851589185763689
1965	0.6183273505694477
1966	0.6376039057352495
1967	0.6116343256825704
1968	0.6228598886325913

As the results show, in the 1960s, the similarities between all music and The Beatles show an overall upward trend, indicating that the 1960s is a period of the music revolution.

Due to the large amount of data, we select the most representative features *energy* and *valence* as the example. The following are the figures:

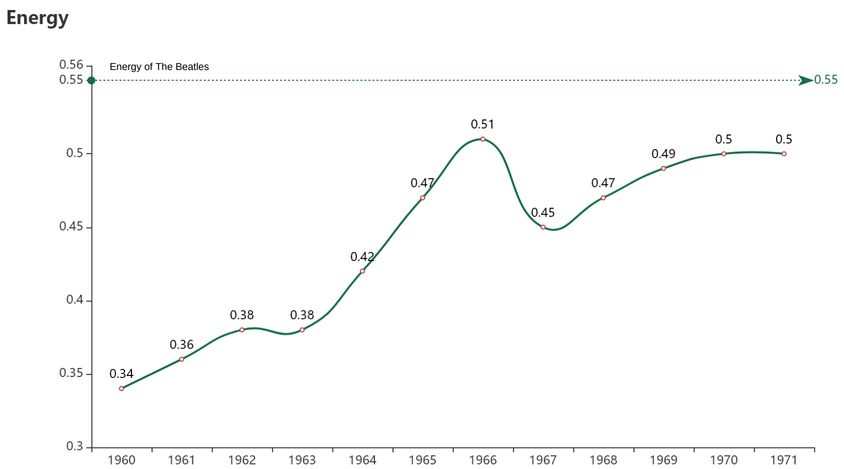

Figure 9: Energy

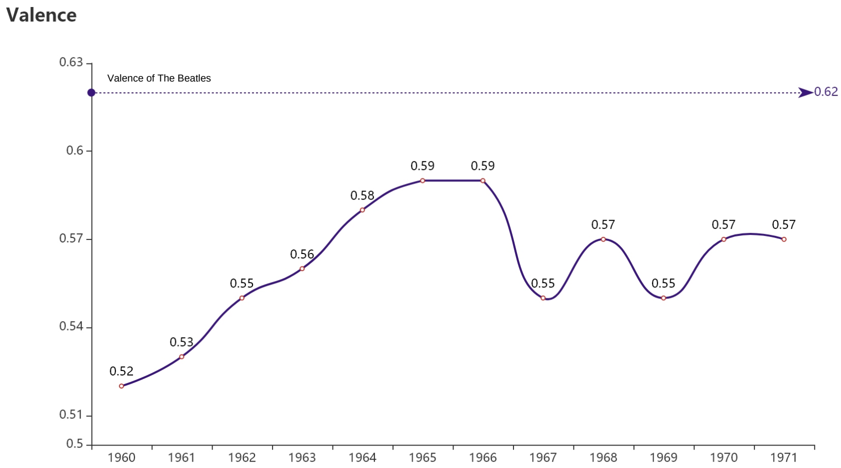

Figure 10: Valence

From Figure9 and figure10, we can draw the same conclusion that 1960s is a period of the music revolution.

7.3 Who are the Dynamic Ifluencers?

We have mentioned that we use the PR values to measure the influence of an artist. However in the PageRank Algorithm, the weight of each link in the directed network is the same. But a influencer's influence on followers is different, so we have to add corresponding weights to the links in the directed network. In order to establish indicators for measuring dynamic influencer, through reading related papers, we think that if an artist can influence artists who are far from his age, or influence artists who are very different from his music genre. Then we can define him/her to be a dynamic influencer.

Therefore, we modify the PageRank Algorithm. We establish the time dynamic index θ_{time} and the genre dynamic index θ_{genre}. The specific formulas are as follows:

$$\theta_{time} = \alpha \frac{|t_1 - t_2|}{10} \tag{12}$$

$$\theta_{genre} = \beta d(X, Y) \tag{13}$$

Then, we establish the overall dynamic index W_{ij}:

$$\mathrm{W}_{ij} = \begin{cases} 0 & \text{, If there is no link between nodes } I \text{ and } F_i \\ \theta_{time} + \theta_{genre} & \end{cases} \tag{14}$$

We put W_{ij} into the formula of PR value:

$$\mathrm{DPR}\ (I) = \frac{1-q}{N} + q \sum_{F_i} \mathrm{PR}\ (F_i)\, \mathrm{W}_{ij} \tag{15}$$

Then we recalculated the PR value and re-ranked the artists. We also add the PR rank to the table as comparison and list the changes of rank. The new rank is shown in Table9:

Table 9: Rank of Dynamic Artists

Ranking	By PR(I)	By DPR (I)	The value of
1	The Beatles	The Beatles(+0)	0.004424506259051719
2	Bob Dylan	Bob Dylan(+0)	0.0024876077030674005
3	The Rolling Stones	Hank Williams(+1)	0.002374811935053274
4	Hank William	Billie Holiday(+2)	0.0022441155708884377
5	Louis Jordan	Louis Jordan(+0)	0.0019481415446670977
6	Billie Holiday	The Rolling Stones(-3)	0.0018989216875588126
7	Lester Young	Woody Guthrie(+8)	0.0018331238587048623
8	Cab Calloway	Lester Young(-1)	0.0016601348708563494
9	David Bowie	Cab Calloway(-1)	0.001650122870042569
10	Roy Acuff	Roy Acuff(-2)	0.001585506090716556
11	Chuck Berry	David Bowie(-2)	0.001585505741603026
12	Jimi Hendrix	Miles Davis(+5)	0.001553637011591875
13	Elvis Presley	James Brown(+1)	0.001508399626543891
14	James Brown	Jimi Hendrix(-2)	0.0014464106435114526
15	Woody Guthrie	Marvin Gaye(+10)	0.0014260307594694262

7.4 The Changes of Music Genres Over Time

In order to study the changes of music genres over time, we have counted the number of artists of various music genres in different years, and draw the box plots based on the statistical results:

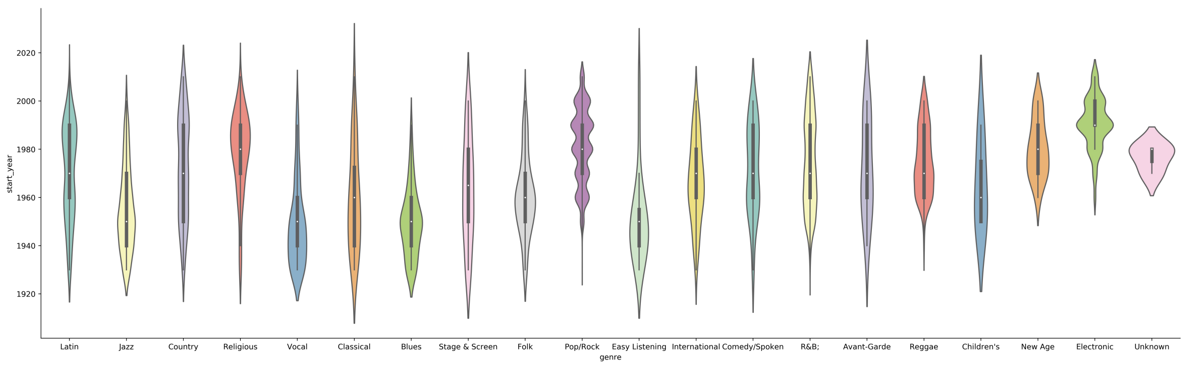

Figure 11: Artists Time Distribution

Through the Figure11, we can find the change of popular music genres in different periods.

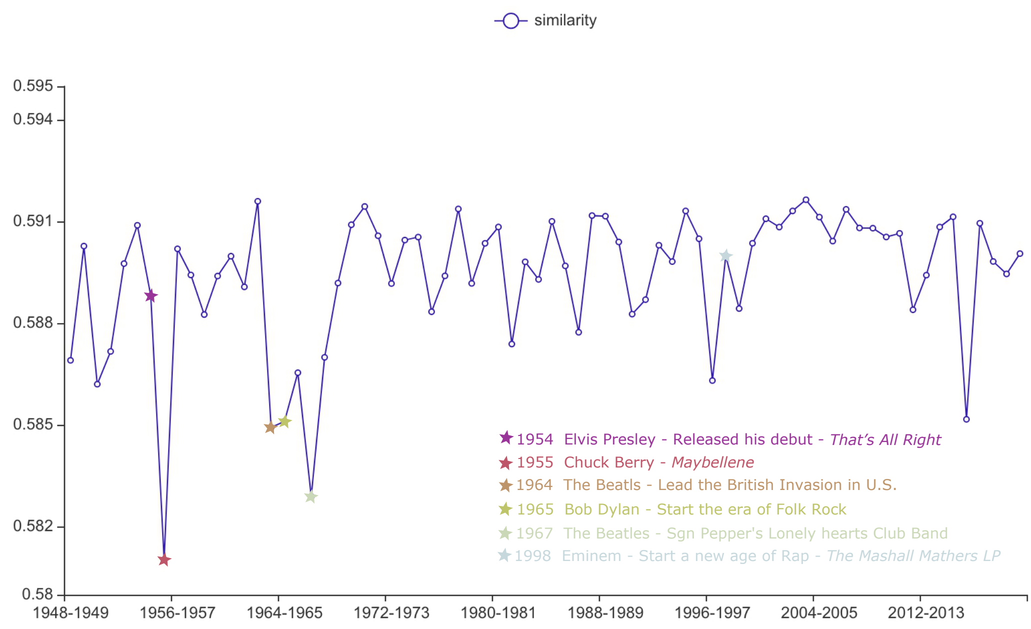

Figure 12: Revolution Events

We use time series to calculate the similarity between years, and connect it with events (Figure12). Through this charts, we can clearly see that during 1954-1956 and 1964-1967, similarity had significant changes and maintained a low position. In 1954 and 1955, Elvis Presley and Chuck Berry ushered in the era of rock and roll. In 1964, Beatles' concert in the United States led the British invasion, while Bob Dylan created a new era of folk rock. The cultural integration of music makes people constantly innovate in this field, and the low similarity between years also shows that artists are constantly trying new styles in music. Since the 1990s, with the continuous development of science and technology, the forms of music creation have gradually diversified. People's thinking is more open and diversified, which makes some minority music gradually rise. While at the same time, due to the development of network and the influence of mainstream media, music always has a main line of development. This has led to fluctuations after the 1990s

8 Sensitivity Analysis

In our music influence network analysis, Dynamic PageRank Algorithm is an esstential part which determines influence of every musician. In this algorithm, there are 3 parameters, namely damping factor q, in θ_{time} in θ_{genre}. Through multiple attempts, we determine the values of the three are $0.33, 2, 5$. To test the robustness of our model, we change the value of these parameters in turn when others are fixed. The results of the test are:

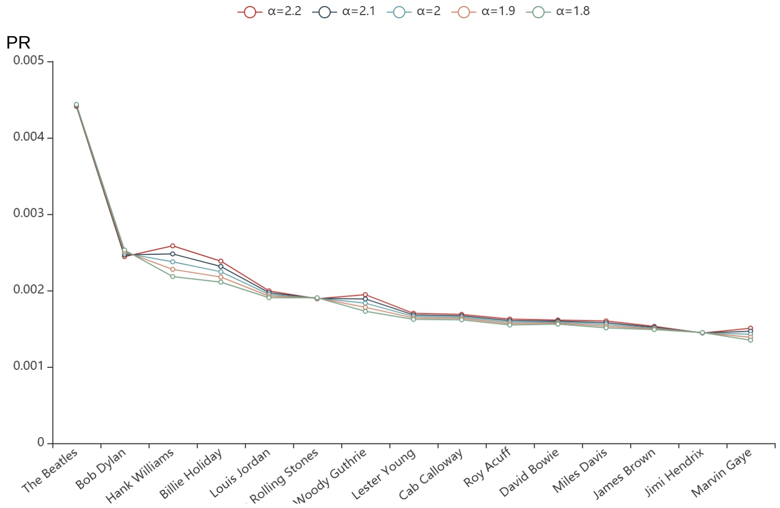

Figure 13: α

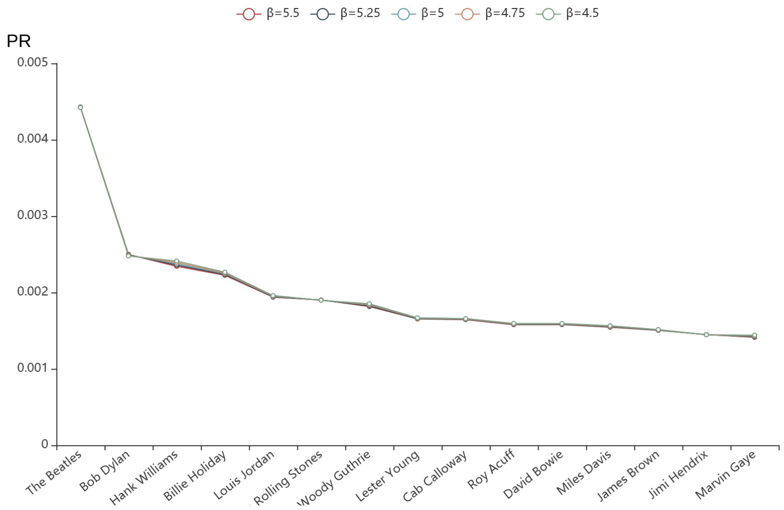

Figure 14: β

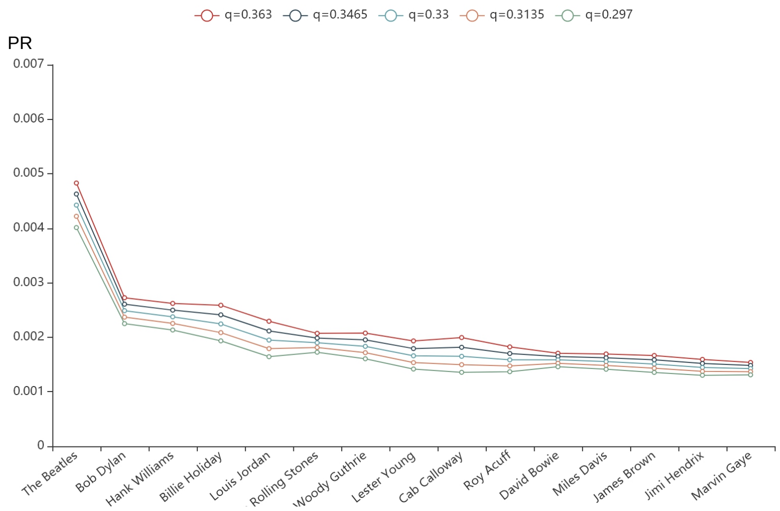

Figure 15: q

According to the results, we find model is quite stable when only is changing, the line The line hardly moves. Data also shows that when changes per 1%, IFSP averagely changes 0.0067% . When or q is changing, though from the lines IFSP of someone changes obviously, but the results of calculation show that for the one whose IFSP changes most dramatically when q changes per 1%IFSP averagely changes per 0.87% and when changes per 1%,IFSP averagely changes per 2%. So, in general, our model is relatively stable.

9 The Doccument for The ICM Society

Theme:The value of our approach in understanding the influence of music through network; Method optimization with increasing data volume.

Dear ICM Society

Modern time is an era of constant changes and innovations, and cultural changes seem to be ceaseless. This is especially obvious in popular music. Although there have been a lot of academic studies and books study the origin and evolution of music, most of the statements are anecdotal and not based on scientific evidence, so the results of the research are often not convincing. In order to better understand the impact of music, we need rigorous hypotheses and mathematical analysis based on statistics and quantitative data.

Our mathematical model is based on statistics to developed a complex directed network. Through the directed network, we can have a clearer and more intuitive understanding towards the influence of influencers on followers. In order to quantify the influence, we use The Modified PageRank Algorithm to calculate the PR value to measure the influence of different artists, and we build the function NR to measure their influence on other artists. Based on the calculation results, we successfully visualized the influence between influencers and followers. We have used mathematical models to quantify the influence of music, rather than based on subjective descriptions, which is an effective and scientific study of the influence of music.

At the same time, based on the directed network, we can measure the similarity between music genres and artists. Through the analysis and research of similarities, we can effectively sort out the evolution of music, which cant be achieved by subjective analysis. What's more, the above methods can be extended to many other research fields, such as: other art types, history, literature, etc. It has strong applicability.

As you have said, the two problem data sets are limited to only some general music genres like Pop/Rock, Country, Jazz, etc. and to those artists common to both data set. But some music genres contain many sub-categories, For example, Country can be divided into dozens of categories such as Classic country, Christian country music, and Bluegrass, etc. Although many songs belong to the same music genre, there are significant differences between them, which brings great difficulties when we analyze the influence and similarity of music, making it difficult for us to summarize a more detailed and clear directed network, and can only find relatively vague relations. If we are provided with more or richer data, our network will become more clear and detailed, which allows us to have a more accurate understanding of the features and the evolution process of different music genres, and enhance the effectiveness of the analysis results.

We think that the further research on music can fall on the history of music evolution and the establishment of music classification standards. Through the directed network, we can establish a time series of changes in music features. We can take historic events into consideration to explore the underlying reasons behind the changes in music features. Because music has a plenty of forms, it is a very difficult job to classify music genres, but through the directed network, we can calculate the similarities within music genres numerically, achieving effective classification of music, which can facilitate future in-depth research.

Sicerely yours,
Your friends

References

[1] Mauch, M., Maccallum, R. M., Levy, M. & Leroi, A. M. (2015). The evolution of popular music: USA 19602010. Royal society open science, 2 (5), s. 150081. doi:10.1098/rsos.150081

[2] Albán, M., Choksi, V., & Tsai, S.B. (2015). CS 224 W Final Report : Influence Networks in Popular Music.

[3] A. Berenzweig, B. Logan, D. P. W. Ellis and B. Whitman, "A Large-Scale Evaluation of Acoustic and Subjective Music-Similarity Measures," in Computer Music Journal, vol. 28, no. 2, pp. 63-76, June 2004, doi: 10.1162/014892604323112257.

[4] Meng, A., Ahrendt, P., Larsen, J., & Hansen, L. K. (2007). Temporal feature integration for music genre classification. IEEE Transactions on Audio, Speech, and Language Processing, 15(5), 1654-1664.

[5] Aucouturier, J. J., & Pachet, F. (2002, October). Music similarity measures: What's the use?. In ISMIR (pp. 13-17).

[6] Serrà, J., Corral, Á., Boguñá, M., Haro, M., & Arcos, J. L. (2012). Measuring the evolution of contemporary western popular music. Scientific reports, 2(1), 1-6.

Appendices

Codes for The Entropy Weight Method

```
def cal_weight(data):
# continuous variable
data1=data.iloc[:,:9]
data1 = (data1-data1.mean())/(data1.std())
m,n=data1.shape
data1=data1.values
k=1/np.log(m)
yij=data1.sum(axis=0)
pij=data1/yij
test=pij*np.log(pij)
test=np.nan_to_num(test)
```

```
ej=-k*(test.sum(axis=0))
wi=(1-ej)/np.sum(1-ej)
#discrete variable
data2=data.iloc[:,9:]
index=['mode','key']
for ind in index:
w=0
res=data2.groupby(ind).count()
a=[x[0]/len(data2) for x in res.values]
for x in a:
w=w-np.log(x)*x
wi=np.append(wi,w)
return wi
```

Codes for Calculating Similarity

```
#calculate similarities
def cal_dis1(a1,a2,b1,b2): #a1,a2 means the continuous and discretepart of a
d1=np.square(a1-b1)
d2=0
i=0
while i<len(d1):
d2=d2+d1[i]*weight[i]
i=i+1
for x,y in zip(a2,b2): #one hot
if x!=y:
d2=d2+1*weight[i]
i=i+1
d2=np.sqrt(d2)

return 1/d2
```