Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
D		2108125

Musical Evolution Analysis Based on Social Network and Isomorph Space

Summary

Music has been part of human societies since the beginning of time as an essential component of cultural heritage. As part of an effort to understand the role music has played in the collective human experience,a method is required to quantify musical evolution.

Firstly, we create a directed adjacent matrix with artists as nodes so that the influence network can be quantified. Utilize **characteristic path length**, **clustering coefficient** and **Katz centrality** to measure 'musical influence' of one artist from three perspectives. Then apply Newman's **Method of Optimal Modularity** to divide the influence network into several communities. The parameters reflect the strength of influence the communities division provide the scope of influence.

Secondly, we constructed a **musical characteristic space** via the musical features of each song. Consider a song(artist) as a node in this space, we can use distance to capture the similarity between two songs. To simplify the process, **principal component analysis(PCA)** is applied. When it comes to genres, we develop genre radius, genre distance, genre independence and some other indexes to describe the characteristics of genres.

Moreover, the method is suitable for the circumstance when consider about time. When we analyze the change of genre radius, genre distance and other indexes over time, we can find out when the musical revolution toke place and who the revolutionaries are. Last but not least, this method can also be applied to explore the relationship between social factors and musical evolution.

Keywords: characteristic path length; clustering coefficient; Katz centrality; Method of Optimal Modularity; musical characteristic space; principal component analysis(PCA)

Contents

1 Introduction **3**

2 Our Work **3**

3 Assumption **4**

4 Symbol Description **4**

5 Social Network Analysis **4**

5.1 Musical Inufluence Parameters 4

5.1.1 Characteristic Path Length 4

5.1.2 Clustering Coefficient 6

5.1.3 Katz Centrality 7

5.2 Creating A Subnetwork of Directed Influence Network 7

5.3 Results Analysis 8

6 Measure the Similarity **10**

6.1 Data Processing 10

6.2 Distance Definition 10

6.3 Distance between Artists 11

7 Similarity and Influence between Genres **12**

7.1 Distance between Genres 12

7.2 Similarity between Genres 13

7.3 Measure the Interplay between Genres 13

7.4 Change over Time 13

8 Similarity and Influence between Artists **15**

9 Analysis on Musical Revolution **17**

9.1 Confirm the Revolution Time 17

9.2 Confirm the Revolutionary 18

10 Process of Influence **18**

10.1 Genres Independence 18

10.2 Influence Process 19

11 Other Factors of Revolution **20**

12 Document to ICM Society **21**

Appendices **22**

Appendix A Complete Data **22**

Appendix B Code **23**

1 Introduction

Music has been part of human societies since the beginning of time as an essential component of cultural heritage. There are many factors that can influence artists when they create a new piece of music, including their innate ingenuity, current social or political events, access to new instruments or tools, or other personal experiences. As part of an effort to understand the role music has played in the collective human experience, it is necessary to develop a method to quantify musical evolution.

Throughout history, musicians have been heavily influenced by previous generations of artists. Therefore, musical influence can be used as an indicator to measure the relationship between artists. At the same time, the result of the influence can be reflected in the works of an artist, so the study of the characteristics of an art work can be used as a sample to measure the relationship between artists.

A slew of prior arts shed light on the study of musical influence:Morton et al[4] presented two **audio content-based systems for influence recognition**:a system using a spectral representation and support vector machines and another system that obtains features by using a deep belief network and then logistic regression for classification, Nicholas J. Bryan and Ge Wang[1]: presented **a computational analysis of musical influence networks** and proposed a method of influence rank, unifying song-level networks to higher-level artist and genre net-works via a collapse-and-sum approach. However both of the methods mentioned above take the content of the songs as the starting point to explain the influence between artists. Constructing maps to reflect influences from the similarity of content ignores the opinions of the artists themselves, for example, the people who is identified as influenced by a certain artist may have no intersection with each other.

In this work we've got data from **AllMusic.com** and **Spotifys API** including the research that represents musical influencers and followers, reported by the artists themselves, as well as the opinions of industry experts ,and some musical features. An influence network is constructed and used to study the musical influence. At the same time we build a model to quantify the similarity of each song with the musical features above.

2 Our Work

In this work, we propose an method to measure the influence and similarity of artists. By utilizing this model, we can interpret the complex relationship among the genres, artists and their songs.

- **SNA(social network analysis)**: We create an influence network of all artists, apply parameters to quantify the influence and divide the network into different communities.
- **Distance definition**: We utilize musical features to construct a space(musical characteristic space). Distance between artist is used to measure the similarity between them. Other parameters in this space are also used to reflect a genre's characteristic.
- **Analysis on genres**: We set several parameters to capture the relationship between two genres, and expand the musical characteristic space from the perspective of time.

- **Analysis on artists**: We redefine several parameters to adopt the new circumstance, and utilize the measure of influence to figure out the weight of musical characteristic to determine influence.
- **Analysis on Revolution**: We set a standard to select the revolution period, and seek artists with great influence as revolutionaries.
- **Process of influence**: We use the calculated indicators to figure out the process of influence.
- **Other factors of revolution**: We obtain several points that represents the revolution. According to the events happened in world, try to explain the musical evolution

3 Assumption

- An artist won't change his musical style easily so that the mean of his music is representative enough
- The influence network is relatively stable
- An artist will have influence only after he published his music.

4 Symbol Description

For compactness, we define a series of symbols for some notation concerning the analysis on reviews and star ratings in **Table 1**.

5 Social Network Analysis

influence_data data set can be used to establish a directed network with individual artist as the node. It is specified that the direction of the network is from followers to influencer(see in **Figure 1**). An adjacency matrix $\boldsymbol{A}$ can be constructed.

5.1 Musical Inufluence Parameters

Influence network is a common social network, we determine to use the parameters that describe the characteristics of the network to measure the musical influence. Thus parameters (characteristic path length, clustering coefficient and Katz centrality) that can be calculated from the network are selected. Then, analyze what these three parameters exactly interpret in this social network.

5.1.1 Characteristic Path Length

In a undirected graph, the characteristic path length(CPL) of the network refers to the average path length of all node pairs in the network, and the path length between two nodes L_{ij} which is the amount of edges in the shortest path connecting two nodes.

In our work, we need to make corresponding modifications to fit the circumstance. For a node i, defined the out-edge as a disconnecting edge, a tree $T(i)$ can be generated with node i as the root

Symbol	Definition
$\boldsymbol{A}$	The adjacency matrix of the influence network
a_{ij}	The element of adjacency matrix
$k_{in,i}$	The in-degree of the node i
k_i	The sum of in- and out-degree of node i
$\boldsymbol{T}_K$	Katz influence matrix
α	Decay factor of Katz influence matrix
Q	Modularity in community division
$L^{(g)}(i)$	The characteristic path length of node i in community g
$C^{(g)}(i)$	The clustering coefficient of node i in community g
$\boldsymbol{x}$	PCA musical features vector
$\boldsymbol{y}$	Original musical feature vector
$d(\boldsymbol{x(i)}, \boldsymbol{x(j)})$	Distance between song(artist) i and j
r_h	Radius of genre h
$\widetilde{d}(i,j)$	Distance between genres i, j
v_i	Standard deviation of the artists in the same genre
$\widetilde{P}(i,j)$	Independence between genre i, j
U	The time-musical-feature space
η	The degree of independence between artist and those who he influence
$\mathbb{N}$	The amount of the music published of the same genre

Table 1: Notations

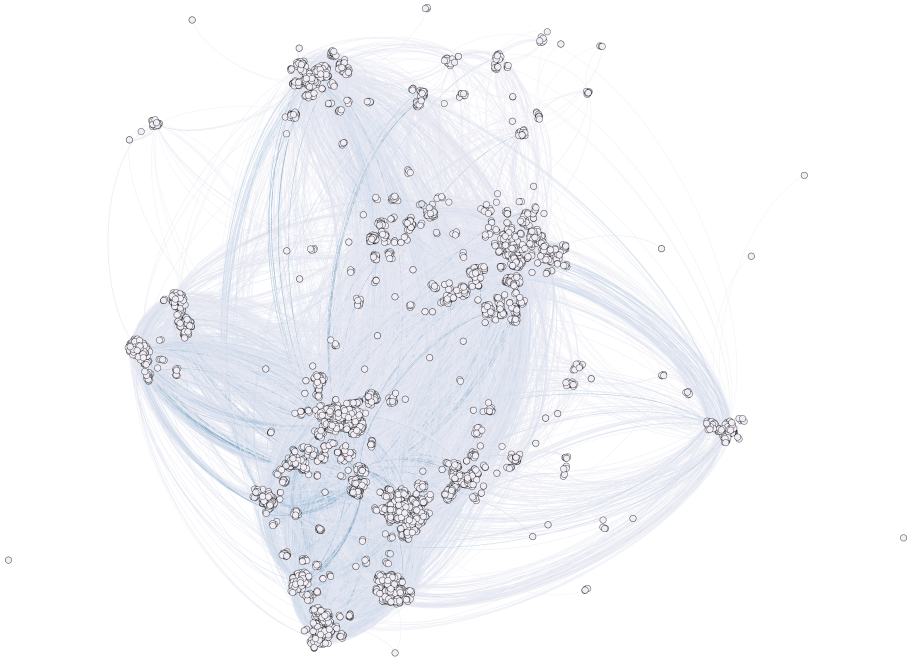

Figure 1: Influence network

node. Sum up all of the $L_{ij}, j \in D(i)$ where $D(i)$ is the set of the leaf nodes belonging to node $i(D \subset T(i))$, the modified CPL is going to be:

$$L(i) = \sum_{j \in D(i)} \frac{L_{ij}}{k_{in,i}}$$

Where $k_{in,i}$ is the in-degree of the node i. From the definition, the modified characteristic path length(MCPL) measure the average depth of root node i.

5.1.2 Clustering Coefficient

A large number of networks show a tendency for link formation between neighboring vertices, i.e., the network topology deviates from uncorrelated random networks in which triangles are sparse. This tendency is called clustering[6]. Assuming that one node i is connected to k nodes with k edges(including in-edges and out-edges), the maximum possible amount of edges, connecting two of those nodes will be $\frac{k(k-1)}{2}$, the portion of the real amount of these edges to the maximum is determined as the clustering coefficient(CC) of this node i.

$$C(i) = \frac{2}{k(k-1)} \sum_{j,k \in B(i)} (a_{kj} + a_{jk})$$

where $B(i)$ is the set of nodes directly connecting with node i, a_{ij} the element of the adjacent matrix $\boldsymbol{A}$.

5.1.3 Katz Centrality

To capture the (direct and indirect)influence of one node(artist), we need a parameter that is able to reflect the indirect connection. As what we build is a social network essentially, it's equivalent to studying inter-personal and inter-group relations. Ordinary indices of 'status' is taken into account. However, most serious investigators of inter-personal and inter-group relations have been dissatisfied with the ordinary indices of 'status', of the popularity contest type[3]. Thus we introduce a new status index: Katz centrality(KC) which provides a mechanism to capture indirect influence between nodes.

The Katz influence matrix $\boldsymbol{T}_K$ is defined as:

$$\boldsymbol{T}_K = (\boldsymbol{I} - \alpha \boldsymbol{A})^{-1} - \boldsymbol{I}$$

where $\boldsymbol{I}$ is an identity matrix, $\boldsymbol{A}$ is the adjacency matrix as before, and α is a decay factor. This can be written as

$$\boldsymbol{T}_K = \alpha \boldsymbol{A} + \alpha^2 \boldsymbol{A}^2 + \cdots + \alpha^k \boldsymbol{A}^k + \cdots$$

According to others' research[3] $\frac{1}{\alpha}$ should be greater than the largest eigenvalue of adjacency matrix $\boldsymbol{A}$ but smaller than twice that eigenvalue(here we have $\frac{1}{\alpha} = 4$).

5.2 Creating A Subnetwork of Directed Influence Network

Many networks of interest in the sciences, including social networks, computer networks, and metabolic and regulatory networks, are found to divide naturally into communities or modules[5]. Similarly, the nodes of the influence network can be classified into different community according to the community detection technology. In our work, we apply Newman's Method of Optimal Modularity to divide the influence network.

The modularity Q is defined via:

$$Q = \frac{1}{4m} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) (s_i s_j + 1) = \frac{1}{4m} \sum_{ij} \left(a_{ij} - \frac{k_i k_j}{2m} \right) s_i s_j$$

where d_i is the degree(sum of in- and out-degree) of node i, $m = \frac{1}{2} \sum_i k_i = \frac{1}{2} \sum_{ij} a_{ij}$, for a particular division of the network into two groups let $s_i = 1$ if node i belongs to group 1 and $s_i = -1$ if it belongs to group 2. The modularity Q can conveniently be written in matric form as:

$$Q = \frac{1}{4m} \boldsymbol{s^T B s} = \frac{1}{4m} \sum_{i=1}^{n} (\boldsymbol{u_i^T s})^2 \beta_i$$

Let $\boldsymbol{s}$ be the column vector of s_i , $\boldsymbol{B}$ is a matrix with elements defined as $b_{ij} = a_{ij} - \frac{k_i k_j}{2m}$, β_i is the eigenvalues of matrix $\boldsymbol{B}$, u_i is the corresponding eigenvector.

We want to maximize the modularity by choosing an appropriate division of the network, or equivalently by choosing appropriate value of the index vector $\boldsymbol{s}$.

When dividing network into more than two communities, the correct approach is to write the additional contribution ΔQ to the modularity upon further dividing a group g of size n_g in two as[5]:

$$\Delta Q = \frac{1}{2m}\left[\frac{1}{2}\sum_{i,j\in g} b_{ij}(s_i s_j + 1) - \sum_{i,j\in g} b_{ij}\right] = \frac{1}{4m}\boldsymbol{s^T B^{(g)} s}$$

$\boldsymbol{B^g}$ is the $n_g \times n_g$ matrix with elements indexed by the labels i, j of vertices within group g, and having value $b_{ij}^{(g)} = b_{ij} - \delta_{ij}\sum_{k\in g} b_{ik}$, where $\delta_{ij}$is the Kronecker δ-symbol, we can now apply the spectral approach to this generalized modularity matrix, just as before, to maximize ΔQ.

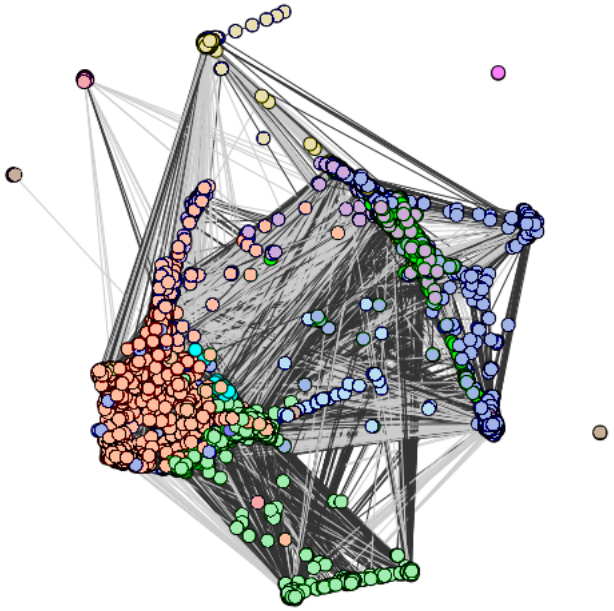

Figure 2: The community division of influence network

We can see in **Figure 2**, nodes with the same color are divided into the same community. The division task stop when the matrix $\boldsymbol{B}^g$ doesn't have positive eigenvalue. Under the circumstance of our influence network, the algorithm divide this network into 29 communities with the modularity $Q = 0.454$.

5.3 Results Analysis

The amount of communities is greater than the amount of musical genres, and most of the communities contain more than one genre. As one community is a subnetwork of the influence network, we can analyze it just as before. For one community, we can calculate the characteristic path length(CPL) $L^{(g)}(i)$, clustering coefficient(CC) $C^{(g)}(i)$ and Katz centrality(KC) $\boldsymbol{T}_K^{(g)}$. These three 'music influence' measures reveal different characteristic in one subnetwork.

- **Characteristic path length**:In the network, the modified characteristic path length refers to how many generations a certain artist's influence can last. CPL measure the 'music influence' from the perspective of longevity.
- **Clustering coefficient**:In a community, this parameter is used to measure the tightness of the community. With great CC, nodes in one community tend to connect with each other more tightly(see in **Figure 3**), with little CC, nodes in one community are more likely to distribute loosely(see in**Figure 4**).

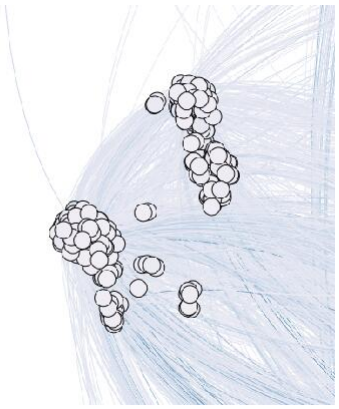

Figure 3: Tight-knit community

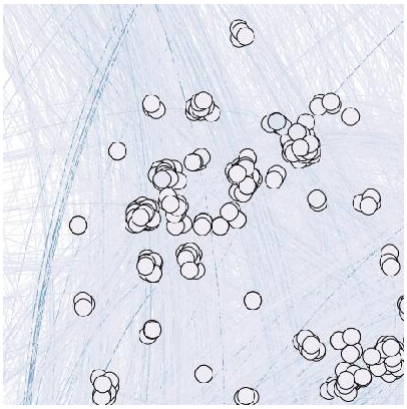

Figure 4: Sparse community

- **Katz centrality**:This parameter provide a mechanism to capture indirect influence between nodes, compute an overall influence rank among each node, and observe the influence of one node to another. With $\boldsymbol{T}_K^{(g)}$ obtained, we can view the column of a node to find who influenced the node, or view the row of the node to find who the node influenced[2]. Summing the components in one columns, from each column we can find the most influential node, while summing the rows to find the most influenced node.

With the parameters above, we can rank the musical influence in three perspective. We select top four artists respectively.(see in **Table 2**)

	Characteristic path length	Clustering coefficient	Kartz centrality
1	Fifth Harmony	Space	Cab Calloway
2	Ariana Grande	Stan Levey	Louis Jordan
3	Demi Lovato	Stan Rogers	T-Bone Walker
4	The Cheetah Girls	Paul Gonsalves	The Mills Brothers

Table 2: Influence ranking from three perspective

6 Measure the Similarity

In order to describe musical similarity we need to establish a indicator: it will be smaller when two samples are more likely the same. It is reasonably thought to use a certain 'distance' to measure how close two samples are. The greater the distance, the less the similarity.

6.1 Data Processing

The establishment of distance between different music depends on its musical characteristics, noted as vector $\boldsymbol{y}$. $\boldsymbol{y}$ is a vector with high dimension, for which data processing is required before the definition of distance.

It is observed that the most of the musical characteristics are continuous values except for the value of *Explicit* and *Mode*, whose value is rather '1' or '0'. And more than 96% value in *Explicit* is 0, so this feature is omitted. Then for the rest 13 characteristics we're going to apply **principal component analysis(PCA)** to reduce the dimension of $\boldsymbol{y}$. In general, almost any data matrix can be simplified by PCA[7]. We set the PCA musical characteristics vector as $\boldsymbol{x}$.

The rest 13 variables are of different scales, which degrades effect of PCA. The data is standardized as follows:

$$Z_i(y_i) = \frac{y_i - \overline{y}_i}{S(y_i)} = \frac{y_i - \overline{y}_i}{\sqrt{\frac{(y_i - \overline{y}_i)^2}{n-1}}}$$

Where $\overline{y}_i$ is the average of the musical characteristics $y_i (i = 1 \cdots 13)$, n is the amount of samples.

After the standardization, PCA is utilized. PCA is a statistical method of data reduction, it is by using a orthogonal transformation that turn the original vector whose components are correlated into a new vector whose components are uncorrelated. In this problem, the PCA extract several characteristics, among which we usually select first few of them. Here, we chose 4 most representative characteristics, x_i represents the musical characteristics extracted by data reduction.

	1	2	3	4
x_1	energy(0.48)	loudness(0.47)	danceability(0.22)	popularity(0.32)
x_2	danceability(0.56)	valence(0.51)	acousticness(0.20)	duration(-0.0.42)
x_3	liveness(0.62)	acousticness(0.60)	speechine(0.13)	popularity(-0.15)
x_4	mode(0.67)	tempo(0.23)	key(-0.52)	loudness(0.66)

Table 3: The composition of the four original musical characteristics that contributed the most

Via PCA, the new musical characteristics are aggregated by several original musical characteristics, some of them are of the practical meaning. (Seein**Table 3**)x_1 is the linear combination of *energy* and *loudness*, and these two characteristics are of the majority of x_1, reflecting x_1 has something to do with *energy* and *loudness*. However, not all x_i can have a clear practical meaning, but they are essential to the distance definition in the following section.

6.2 Distance Definition

For any two songs(music) i and j, parameters $\boldsymbol{x_i}, \boldsymbol{x_j}$ can be calculated. Regard $\boldsymbol{x_i}, \boldsymbol{x_j}$ as elements (points) in the space of musical characteristics $W(i, j \in W)$. The distance between the two

songs(music) is defined via:

$$d(\boldsymbol{x(i)}, \boldsymbol{x(j)}) = \sqrt{\sum_{n=1}^{4} (x_n(i) - x_n(j))^2}$$

The distance can describe how similar two songs(music) are. The closer two point are in the W, the more similar the corresponding songs(music) are. Similarly, the distance between different artists can be computed by their music(if he has several songs, calculate the mean of them) respectively, using the above formula.

6.3 Distance between Artists

To capture the similarity of music(artists) in the same genre, we need to define a center in W. When the music in the same genre all get close to center, every artist seem the same as the others. Denote this center as $\overline{\boldsymbol{x}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x(i)}$, where N is the amount of the artist in the same genre. Therefore similarity of this group can be defined as the radius of the genre in W as:

$$r = \frac{1}{N} \sum_{i=1}^{N} d(\boldsymbol{x(i)}, \overline{\boldsymbol{x}}) = \overline{d(\boldsymbol{x(i)}, \overline{\boldsymbol{x}})}$$

The genre radius r is the average distance of the artists to the $\overline{\boldsymbol{x}}$(see in **Figure 5**, the space of musical characteristics is of much greater dimension, we take three dimension for example)

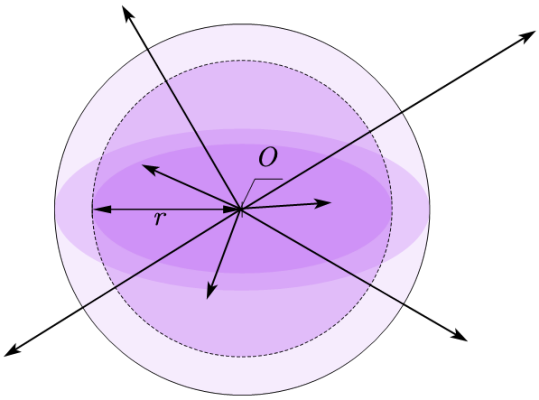

Figure 5: The distance applied in genre

By comparing genre radius, we can know whether the music of the same genre is similar to each other and whether they are close to each other in W. Therefore the indicator can be used to figure out whether artists from the same genre produce more similar songs. Firstly, record the genre information of every artist via the *influence_data* file, with which label the artists with corresponding genres in the *data_by_artist* file. Then calculate the center and radius of each genres via data in the *data_by_artist* file. We take five genres for example(see the appendix for the rest).

random	Pop/Rock	R&B	reggae	Folk	Classical
24.02	1.84	1.75	1.69	1.51	1.52

Table 4: Comparison of the radii of several genres

(See in **Table 4**)*Random* refers to the result obtained by randomly selecting 1000 samples from all artists and treat them as a special genre '*random*'. It can be found that artists of the same genre share a similar musical style, as the genre radius r is much smaller than the genre radius of 'random'.

7 Similarity and Influence between Genres

Genre is a concept of group. To measure the similarity of genre, we use radius, the characteristic of a group of points in W. In this section, we're going to talk about the similarity and influence between the genres. It is reasonable to thought that another characteristic of group will be utilized.

7.1 Distance between Genres

First of all, we consider the center $\boldsymbol{x}$ as the main characteristic of a genre, calculate the distance between two genre canters. Secondly, consider the space a genre occupy from the perspective of mean, and use r as the boundary of such space. Therefore we can consider using the center and radius to measure the distance between genres(see in **Figure 6**). Assuming that r_i, r_j are the radii of genres i, j, $\overline{\boldsymbol{x}}_{\boldsymbol{i}}, \overline{\boldsymbol{x}}_{\boldsymbol{j}}$ are the center of the genres i, j in W, distance between genres i, j is defined via:

$$\widetilde{d}(i,j) = d(\overline{\boldsymbol{x}}_{\boldsymbol{i}}, \overline{\boldsymbol{x}}_{\boldsymbol{j}}) - r_i - r_j$$

Note it that $\widetilde{d}$ and d are of different meanings, the former represents the distance between two groups and the latter reflects the distance between two samples.

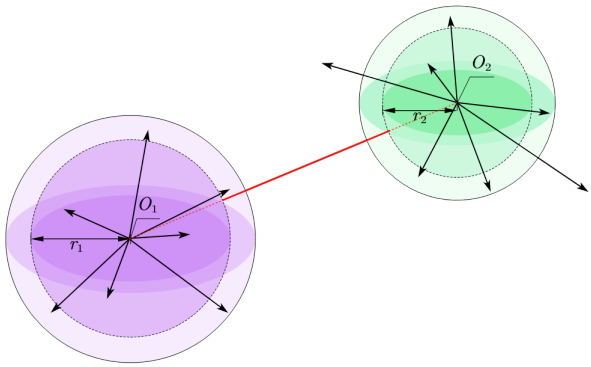

Figure 6: Distance between genres

Similar to the distance between artists mentioned above, the concept of genre distance can be used to describe the similarity between two genres. The application of genre distance is shown as follows.

7.2 Similarity between Genres

To some extent, the length of the solid red line(see in **Figure 6**): the distance between two genres, reflects the degree of difference, and we want to find out which components of $\boldsymbol{x}$ is the most representative. Set $\Delta\boldsymbol{x} = \overline{\boldsymbol{x}}_i - \overline{\boldsymbol{x}}_j$, where we can capture the difference from $\Delta\boldsymbol{x}$'s component.

$$|\Delta x_M| = \max_{1\leqslant n\leqslant 4} |\Delta x_n| \qquad |\Delta x_m| = \min_{1\leqslant n\leqslant 4} |\Delta x_n|$$

After finding the corresponding tags M, m, utilize the correlation matrix to figure out the most similar component and the most different component. According to the former work, we have $\boldsymbol{x}_{4\times 1} = \boldsymbol{R}_{4\times 13}\boldsymbol{y}_{13\times 1}$. Rank the elements of correlation matrix $\boldsymbol{R}$, R_{Mi} and R_{mi} $(i = 1, 2, 3, 4)$ respectively. The greater the $|R_{Mi}|$ is the more likely the y_i is to represent the $\boldsymbol{R_M}$ while the greater the $|R_{mi}|$ is the more likely the y_i is to represent the $\boldsymbol{R_m}$.($\boldsymbol{R_M}, \boldsymbol{R_m}$ are the vectors including R_{Mi}, R_{mi} $(i = 1, 2, 3, 4)$ respectively)

To measure the difference, we can use $\Delta\boldsymbol{x}$ and the genres distance defined above. With the M, m gained above, set:

$$\widetilde{d}(i,j)_M = \widetilde{d}(i,j)\frac{\Delta\boldsymbol{x}^T \cdot \boldsymbol{\xi}_M}{(\Delta\boldsymbol{x}^T\Delta\boldsymbol{x})^{\frac{1}{2}}} \qquad \widetilde{d}(i,j)_m = \widetilde{d}(i,j)\frac{\Delta\boldsymbol{x}^T \cdot \boldsymbol{\xi}_m}{(\Delta\boldsymbol{x}^T\Delta\boldsymbol{x})^{\frac{1}{2}}}$$

Where $\boldsymbol{\xi}_i(4\times 1)$ is the unit vector with the i^{th} variable is 1 other variable is 0.

7.3 Measure the Interplay between Genres

In the radius definition, we defined r as the mean boundary of a genre in W. However, the standard deviation v of group(artists of the same genre) in W can also be used to measure one characteristic of a genre. v is defined as:

$$v = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d(\boldsymbol{x(i)}, \overline{\boldsymbol{x}}) - r)^2}$$

It can be seen from the two figures below(see in **Figure 7 and Figure 8**) that the group with smaller v in W are more concentrated in the spherical shell, while those with larger v are more dispersed in the characteristic space. If one standard deviation is added to the genre radius, a dynamic boundary can be generated. With larger v the dynamic boundary is more different from mean boundary. This parameter can be used to measure the independence of one group.

7.4 Change over Time

In order to introduce time dimension, we first identify the music of the same genre from *full_music_data*, and then divide them according to label 'year'. With the data of each year, there is a musical characteristic space $W^{(t)}$. Therefore we can calculate the genre center $\overline{\boldsymbol{x}}$, genre radius r and standard deviation v, as well as $\overline{d(i,j)}$. for each year. With several values, the time series are established.

At every year, we construct $W^{(t)}$(take 2 dimension for example), thus the time-musical feature space U should be the set of the musical characteristic space$W$(see in **Figure 9 and Figure 10**).

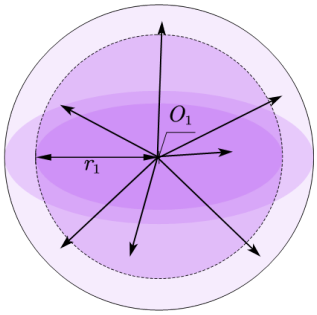

Figure 7: v is relatively small

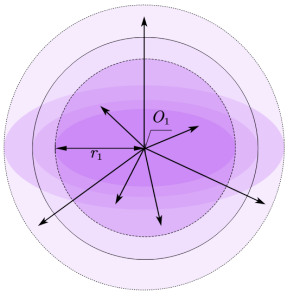

Figure 8: v is relatively great

Select a interval(1955-2020) with sufficient data and choose four genres: *Pop_Rock, Latin, R&B, Country*. Draw 6 genre-radius-related curves between the four genres(see in**Figure 11 and Figure 12**)

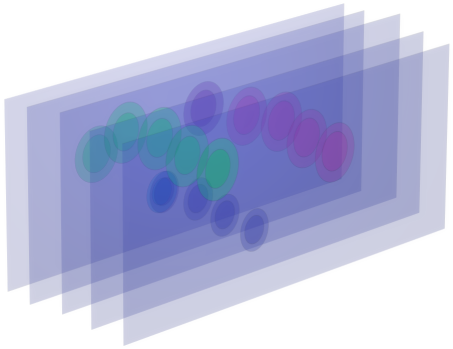

Figure 9: Diagram of U

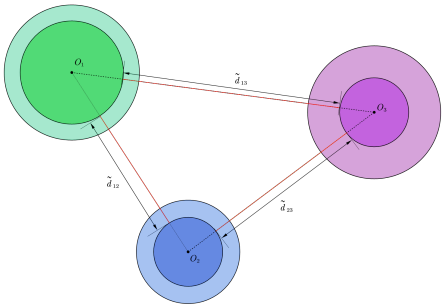

Figure 10: $W^{(t)}$ in one year

(See in **Figure 11**)The distances between *Pop_Rock* and *R&B*, *Pop_Rock* and *Latin*, *Pop_Rock* and *Country* are shown.

The distance between *Pop_Rock* and *Latin* remain stable, meaning these two genres are relatively different. However, in the recent 20 years, distances between *Pop_Rock* and *R&B*, *Pop_Rock* and *Country* reduce. It is said that *R&B,Country* become more and more similar to *Pop_Rock*. (See in **Figure 12**)Distance between *Latin* and *R&B* increases so does the distance between *Latin* and *Country*. Moreover, the distance between *R&B* and *Country* tend to be stable.

Genres radius r and genres standard deviation v also change over time. (See in **Figure 13 ,Figure 14**)No matter the r or v of *Pop_Rock* stay in a high level, which means a large dynamic boundary. Conversely, *Latin* gain a stable mean boundary. Moreover during 10-30, both r and v of four genres atsy in a higher level(compare with themselves). It is thought that during 10-30, the music culture exchanges frequently, each genre redefines a new dynamic boundary and mean

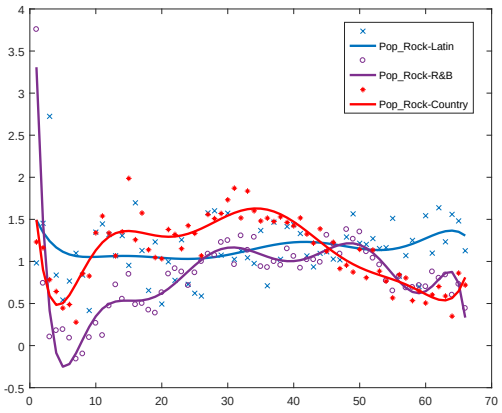

Figure 11: *Pop_Rock* with *Latin*, *R&B* and *Country*

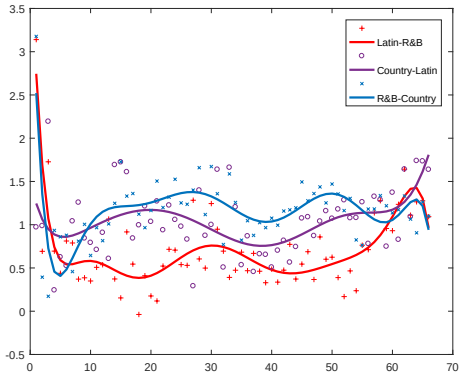

Figure 12: The other three curves

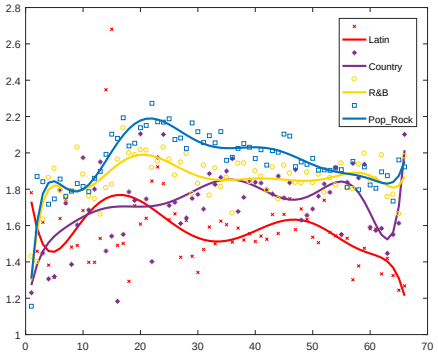

Figure 13: r changes over time

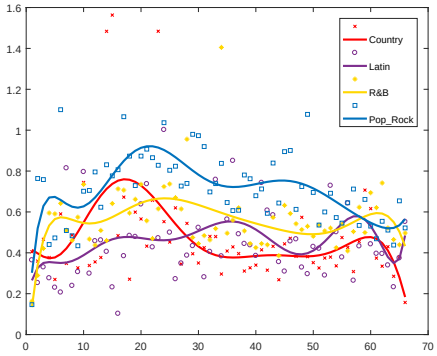

Figure 14: v change over time

boundary.

8 Similarity and Influence between Artists

Intuitively, we think that similarity has something to do with influence. In this section, we will focus on the similarity of the group that is influenced by a certain artist.

The similarity defined by distance above can't apply directly in this circumstance. As genre is a unweighted group, $\overline{\boldsymbol{x}}$ can be utilized as the center of the group and the radius r can be used to measure similarity(mean boundary). However, we are talking about the similarity between one artist i and those who follows i. (See in **Figure 15**, gray sphere refer to those influenced by i, and the center of the green sphere represents the artist i)Even though the artists influenced by i share a similar musical style, it is still possible that artist i has a relatively different music style from those he influences. In such case, the radius r of this group(the gray sphere) may be little but similarity

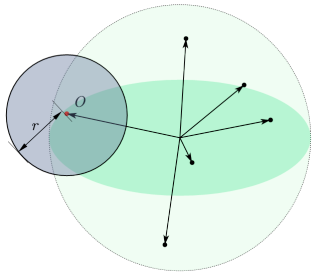

Figure 15: An extreme counterexample

between artist i and those he influence is little.

Thus we redefine the measure considering the effect caused by artist i as:

$$\eta = \frac{N \cdot d(x_i, \overline{x})}{\sum_{n=1}^{N} d(x_n, \overline{x})}$$

Where N is the amount of artists that artist i influence(includes artist i). Apply the definition above to those top artists in the influence rank(section 5) to calculate the η of each top influence artist(see in **Figure 16**).

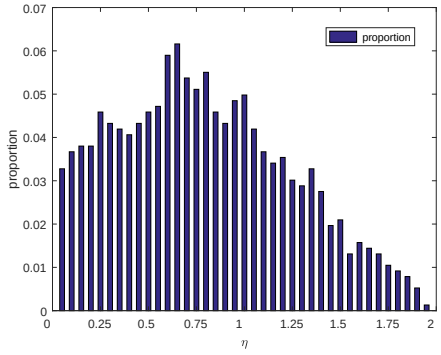

Figure 16: Distribution of η

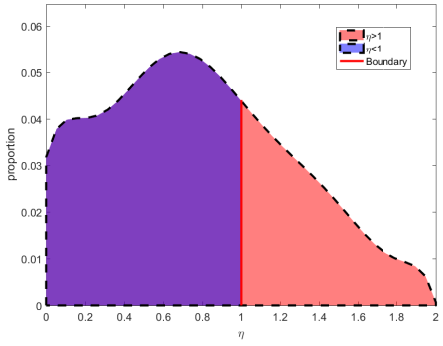

Figure 17: Classification of η

We select the top 1000 artist ranked by Katz centrality. Those artist with little η play a more significant role in affecting their followers' music. (See in **Figure 17**)Most of the influence artists have η less than 1. Therefore the influnecers do affect the music created by their followers.

For detailed information of musical feature in process of influence, we build **Multiple linear regression** model to calculate the contribution of each component of $\boldsymbol{y}$ in the process of influence. (In section 5)We calculate a characteristic to represent the influence:Katz matrix $\boldsymbol{T}_K$. Let $T_K(i) = \sum_j T_{K(j,i)}$ be the influence score of artist i. For artist i:

$$T_K(i) = \lambda_1 y_1 + \lambda_2 y_2 + \cdots + \lambda_{13} y_{13}$$

Note it that standardization of $\boldsymbol{y}$ should be done before regression. Find the $\lambda_{max} = \max\limits_{1 \leqslant n \leqslant 13} \lambda_n$ which is the most contagious musical characteristic. Only when the artists are of the same genre can we apply multiple linear regression. As in different genres a characteristic may play different role in the influence, only in the same genres can this model produce a practical result.

9 Analysis on Musical Revolution

We have already defined the mean boundary and the dynamic boundary with r and v. When the r increases, it is more likely to generate peculiar music in the corresponding genre. While the v increases, it is thought to make the corresponding genre unstable and more easy to absorb new music elements from other genres. However when r and v are both relatively little, the corresponding genre will be stable and refuse to change itself.

9.1 Confirm the Revolution Time

To analyze the musical revolution, we first count the genre radius r, the genre standard deviation v and the amount of songs published of every year of each genre. In this section, we take *Pop_rock* and *country* for example to seek the time of musical revolution(see in **Figure 18, Figure 19**,red curve refers to r, black curve refers to v, blue bar represents the amount of songs published).

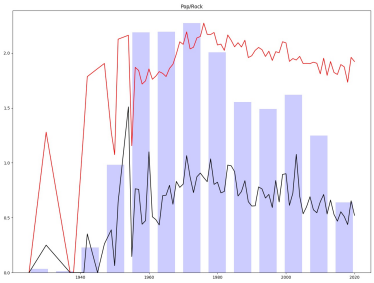

Figure 18: Pop_Rock

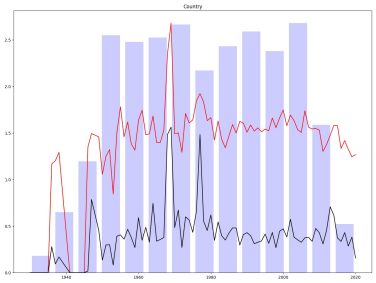

Figure 19: Country

Revolution would make a differnece to the genres that time, for which we can select the revolution period via the change of r and v over time. Meanwhile the amount of the music published $\mathbb{N}$ can reflect the revolution to a certain extent. In all, we are supposed to find a time period to make

$$\Delta\mathbb{N}(t) = \mathbb{N}(t + \Delta t) - \mathbb{N}(t) \qquad \Delta r(t) = r(t + \Delta t) - r(t) \qquad \Delta v(t) = v(t + \Delta t) - v(t)$$

as large as possible. To appropriately reflect the situation of a genre, we choose 5 years as a unit observation period. Calculate the mean of r, v and $\mathbb{N}$ during a 5-year-long period. Then the period t_{max} can be selected by comparing former three indicators, and we deem the corresponding time period as the revolution period.

9.2 Confirm the Revolutionary

There are three parameters to measure 'musical' influence:CPL, CC and KC. Calculate the influence score from three parameters of each artists, and observe the distribution of the 'musical' influence(see in **Figure 20**).

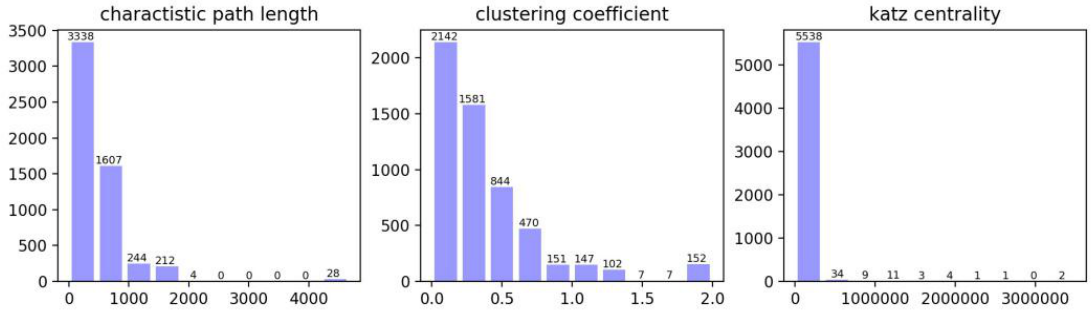

Figure 20: Distribution of three parameters

As social network follow a power-law degree distribution, only few of artist will play a significant role in the network. Katz centrality perform the best among three parameters, so we select it as the measure of influence in a network.

With revolution period tag T_{max} gained, we can count all of the artists who had pubulished music during that time period. Among the goup of artists, we can seek the one who is ranked the first and regard him as the revolutionary.

Take *Pop_rock, Country, R&B* for example(see in **Table 5**):

Pop_Rock	Country	R&B
Little Richard	Roy Acuff	Wynoie Harris
Elvis Presley	Hank Williams	Ray Charles
Bo Diddley	Ernest Tubb	Little Willie John

Table 5: Artists who caused the revolution with great influence

In the 1950s and 1960s, *Pop_Rock, Country, R&B* developed at an incredible speed, lots of representative artists came out, like *Little Richard, Elvis Presley,Bo Diddley*etc. Those artists were the revolutionaries of their genres. The method can be applied to the rest genres.

10 Process of Influence

10.1 Genres Independence

Solely with the genre radius, it unable to know the interaction between two genres in W. However, with the standard deviation it is feasible to confirm whether two genres spheres overlap and whether there is a possible influence between them(see in **Figure 21 and Figure 22**).

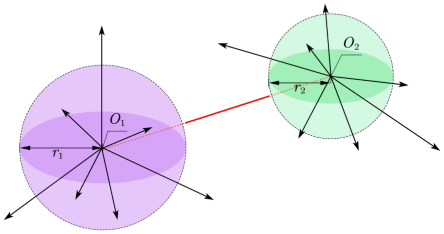

Figure 21: v is relatively small

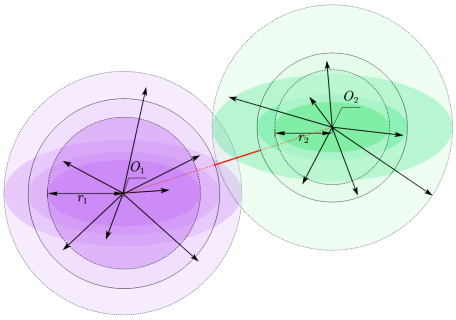

Figure 22: v is relatively great

Thus we need to establish a measure: the larger this parameter is the less possible two genres interact with each other. Therefore the independence $\widetilde{P}(i,j)$ between two genres is defined as:

$$\widetilde{P}(i,j) = d(\overline{\boldsymbol{x}_{\boldsymbol{i}}}, \overline{\boldsymbol{x}_{\boldsymbol{j}}}) - r_i - r_j - 2v_i - 2v_j$$

In this way, even if the distance $d(\overline{\boldsymbol{x}_{\boldsymbol{i}}}, \overline{\boldsymbol{x}_{\boldsymbol{j}}})$ between the two genres remains stable, as long as the v_i, v_j are large enough, the musical exchange and influence still exist between the two genres.

10.2 Influence Process

Even though it is calculated that there is interplay between two genres, it is unable to figure out who the influencer is, who the follower is or other circumstances. However, it is resolvable in our work. The process of influence can be unveiled by musical influence parameters of these two genres, comparing the sum of overall in-degree of each genres and sum of overall out-degree of each genre. With the larger influence parameters and larger in-degree, the genre who influence is confirmed. The process of influence can be reflected by genres independence.Take *Pop_rock, Latin, R&B* for example(see in **Figure 23, Figure 24**).

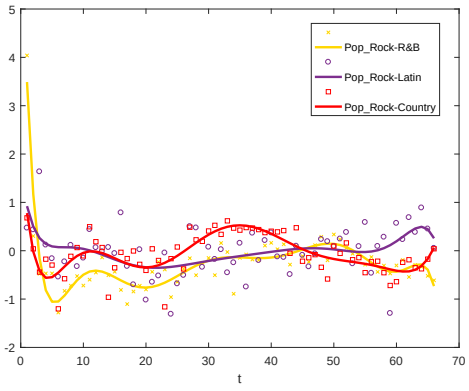

Figure 23: Genres independence-time

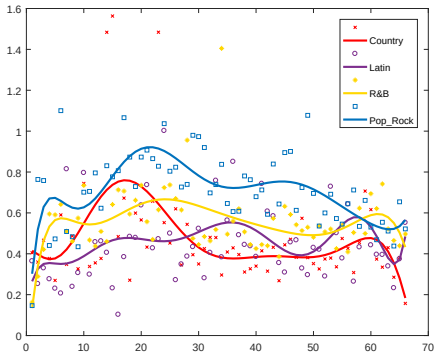

Figure 24: Standard deviation-time

It is observed that the deviation v waves during 20-30, genres independence $\widetilde{P}(i,j)$ waves

during 30-40. Establish the influence process from this perspective. With the development of a genre, its v and r will increase, contributing to the combination with other genres. At that time, the genre with great influence will output its musical characteristics, as a result of which each genre will change their position in the musical characteristic space W. This process is how a genre influence others. Similarly, artist belong to a genre, so the analysis process is the same, only need to replace genre independence with the sum of the distance about artist to each center of genres.

11 Other Factors of Revolution

Music influence can be considered from the perspective of genre radius and genre deviation. In this work our model shows that *Pop_Rock* is the genre with the largest relative influence in the 1950s.

We have discussed how to identify the revolution period and revolutionaries. Here we will discuss the influence of other factors, such as the Internet and the usage of new musical instruments, on genres. Referring to the method of finding radius, standard deviation and quantity change mentioned above. We need to dig deeper into how genre parameters change over time. Fit the time series, discuss the extreme point, inflection point and other points containing special information, so as to find the influence of relevant social factors on music.

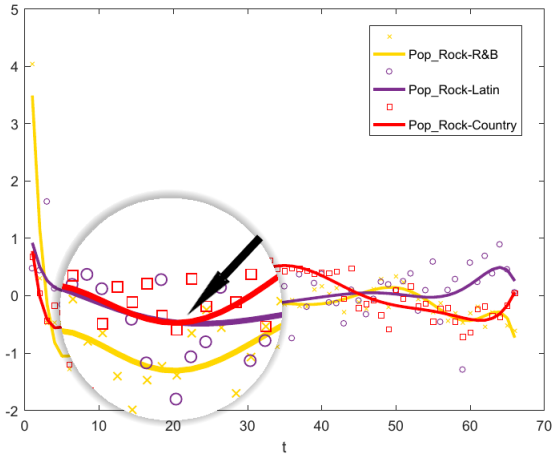

Figure 25: Genres independence curve

For the point shown in **Figure 25**, we wonder if there is something other factors influence the genres. Via the time of these special points, we do some research to seek the other factors that influence music evolution.eg

1. The music of the 1950s reflected the beginning of major social changes in the world, especially in the United States, as reflected in the explosion of the parameters defined above

2. The sounds of the 1960's straddled a large dichotomy between the ultimate commercialism with completely manufactured bands (like The Archies and The Monkees) and revolutionary artistry (Bob Dylan and Jimi Hendrix) with some of the greatest singer-songwriters and instrumentalists emerging on the scene(the extreme point among 10).

3. CDs were the most popular way to listen to one's favorite songs until the middle of the 2000's when computer and internet technology advanced significantly to make MP3 format and MP3 players more viable means of entertainment(v shows the increasing trend).

From the example, we can confirm that there are several other factors influencing the musical revolution.

12 Document to ICM Society

The values of using our approach are as follows:

1. The characteristic path length(CPL), clustering coefficient(CC) and Katz centrality(KC) can be used to describe the influence of artists in the network, and it can be well verified that these musicians with great influence in the network also have a revolutionary promoting effect on the evolution of music. In particular, the results of Katz centrality are very close to reality.

2. PCA is used to data reduction, reduce data dimensions, making it possible to solve the problem even though the dimension of data is much larger.

3. By using the concept of isomorph, the aggregated musical characteristic space is connected with $\mathbb{R}^4$, which makes the description of the problem more vivid. Introduce the concept of distance and then get genre radius and genre deviation. For different elements establish a different distance can be a good description of the relation among genres, artists and songs.

4. The radius and deviation of genres that change over time can skillfully find out its actual meaning and reflect the musical evolution process through its changing trend.

With more and richer data,

1. For the increase of genres, the community detect algorithm can be used to divide the network into different communities. The community can be regarded as the integration of genres, and the change and interaction between communities can also reflect the change of genres.

2. For the increase of artists, the Katz centrality algorithm is no longer applicable as the sudden increase of the time complexity. The characteristic path length can limit the number of steps of the DFS algorithm to reduce the time complexity, but it will not affect the clustering coefficient.

References

[1] Nicholas J Bryan and Gen Wang. Musical influence network analysis and rank of sample-based music. In *ISMIR*, pages 329–334, 2011.

[2] Charles H Hubbell. An input-output approach to clique identification. *Sociometry*, pages 377–399, 1965.

[3] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[4] Brandon G Morton and Youngmoo E Kim. Acoustic features for recognizing musical artist influence. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 1117–1122. IEEE, 2015.

[5] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[6] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998.

[7] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

Appendices

Appendix A Complete Data

Table 6: r and v for each genres

genre	Pop/Rock	Latin	Jazz	R&B	International	Country	Classical
mean	1.84	1.63	1.82	1.75	1.83	1.45	1.52
standard vediation	0.64	0.41	0.61	0.49	0.68	0.42	0.58

genre	New Age	Religious	Folk	Unknown	Avant-Garde	Children's
mean	1.09	1.71	1.50	0.25	1.08	1.12
standard vediation	0.34	0.34	0.57	0.03	0.21	0.22

genre	Electronic	Stage&Screen	Comedy_Spoken	Easy Listening	Vocal	Reggae	Blues
mean	2.00	1.65	1.70	1.18	1.74	1.69	1.53
standard vediation	0.71	0.68	1.19	0.29	0.54	0.46	0.33

Appendix B Code

```
#!/usr/bin/python3
# coding: utf-8

# import required models
import igraph
from igraph import plot
import pandas as pd

# read influence_data data set
data = pd.read_csv('../data/influence_data.csv')

# generate the edges of graph as a list of tuples, and each tuple reprsent an edge
edges = []
for i in list(map(tuple, data[['inf_id', 'flw_id']].values)):
    tp = (i[0], i[1])
    if tp in edges or tp[::-1] in edges:
        continue
    else:
        edges.append(tp)

# create object Graph by edges
g = igraph.Graph(edges)
# community fastgreedy algorithm
g_ifmp = g.community_fastgreedy()
print(g_ifmp.optimal_count)
g2 = g_ifmp.as_clustering(n=16)


# beautify figure style and store image
visual_style = {}
visual_style["vertex_size"] = 10
visual_style["layout"] = g.layout("drl")
visual_style["margin"] = 80
out = plot(g2, **visual_style)
out.save('res.eps')

# acquire parameter
print(g2.q)
f = open('membership1.csv', 'a+')
mbs = g2.membership
for i in range(len(mbs)):
    print(i)
    f.write("%d,%d\n" % (i, mbs[i]))
f.close()
```

```
#!/usr/bin/python3
# coding: utf-8

import pandas as pd
import numpy as np

# read influence_data data set
```

```
data = pd.read_csv('../data/influence_data.csv')

# generate adjacency matrix
n = max(max(data['inf_id'].values) + 1, max(data['flw_id'].values) + 1)
A = np.zeros((n, n))
for i in range(len(data)):
    A[data['flw_id'][i], data['inf_id'][i]] = 1

# ## clustering coefficient
# calculate the clustering coeffiient and store in list clst
clst = []
for i in range(len(A)):
    k = sum(A[i, :]) + sum(A[:, i])
    print(k)
    e = 0
    neighbor = []
    for j in range(len(A)):
        if A[i, j] == 1 or A[j, i] == 1:
            neighbor.append(j)
    # get non-reapting nodes using set
    neighbor = set(neighbor)
    for j in neighbor:
        for l in neighbor:
            e += A[j, l] + A[l, j]
    if k != 0 and k != 1:
        clst.append(2 / k / (k - 1) * e)
    else:
        clst.append(0)

# write out clustering coefficient
f = open('clst1.csv', 'a+')
for i in range(len(clst)):
    f.write("%d,%f\n" % (i, clst[i]))
f.close()

# ## characteristic path length
# store each infulencer 's followers in dict infulencers
infulencers = {}
for i in range(len(A)):
    infulencers[i] = np.where(A[:, i])


# using dfs algorithm to find leaf node
# reached_nodes
def dfs(node, ttl):
    global reached_nodes
    reached_nodes.append(node)
    if ttl == 0:
        return 0
    sum = 0
    if len(infulencers[node][0]) == 0:
        return 0
    for i in infulencers[node][0]:
        if i in reached_nodes:
```

```
            continue
        sum += dfs(i, ttl - 1) + 1
    return sum


length_sum = {}
for i in range(len(A)):
    reached_nodes = []
    length_sum[i] = dfs(i, 100)
    print(i)

# calculate characteristic path length
length_res = {}
for i in length_sum.keys():
    k = sum(A[:, i])
    if k == 0:
        length_res[i] = 0
    else:
        length_res[i] = length_sum[i] / k

# write out
f = open('length1.csv', 'a+')
for i in length_sum.keys():
    f.write("%d,%f\n" % (i, length_sum[i]))
f.close

# ## katz centrality
# calculate eigenvalues
lmds, _ = np.linalg.eig(A)
lmd = np.max(lmds)
print(lmd)
# get the maxium eigenvalues and estimate  decay factor a, a is 1/4
a = 1 / 4
Ik = np.zeros(A.shape)
pre_A = np.identity(len(A))
for i in range(100):
    pre_A = a * pre_A.dot(A)
    Ik += pre_A

# write out
f1 = open('influcing_katz1.csv', 'a+')
f2 = open('influced_katz1.csv', 'a+')
for i in range(len(A)):
    f1.write("%d, %f\n" % (i, sum(Ik[:, i])))
    f2.write("%d, %f\n" % (i, sum(Ik[i, :])))
f1.close()
f2.close()
```