Problem Chosen
D

2021
MCM/ICM
Summary Sheet

Team Control Number
2118761

Deep Dive into Music: Analysis on Influence Network and Musical Evolution

Summary

Music takes an important part in peoples daily life, and is evolving over time. Just as individuals in a social network can influence each other, the artworks of a music artist will also be influenced by other artists. Our team was commissioned by the ICM to conduct research on the influence of artists, the degree of similarity within and between artists and genres, and the evolution of musical characteristics.

First, inspired by topological sorting and PageRank, we proposed a measure called **InfluenceRank**, and took several influence subnetworks to validate the rationality of this measure.

Then, we embedded music and artists into the vector space, and used the cosine of the angle between them as the similarity measure. Based on this, we found that the music and artists **in the same genre are more similar compared with those between different genres.** Next we generalized cosine similarity and devised mean square error to measure the similarity of a genre from complementary perspectives. Meanwhile, we **modified the InfluenceRank** to measure the influence between and within genres. We also obtained the genres distinguishing features through **univariate feature selection** and quantified the relevance between genres by calculating **correlation**. Afterwards, we depicted artists from music level to better capture the temporal information, and designed a **relative distance difference approach** to model actual influence among artists, by which the Push influencers and Pull influencers were identified. We also examined the **change distribution of every music characteristics** during two time periods, and found that **loudness** and **energy** are contagious.

In order to manifest the change of evolution speed, we computed the **first difference** of features time series. We then quantified such revolutionary change by **Augmented Dickey Fuller test** and found **acousticness** and **speechiness** are the two major signs of musical revolution. Besides, revolutionary artists were identified by summarizing the number of music works he released in revolutionary decades. In terms of musical evolution, we first designed **representations with temporal information** for genres and built up time series for each genre accordingly. After that we calculated correlation between the overall musical evolution and the genre-specific one to reveal influence process, where we took **Jazz** as a representative to illustrate how genres were involved in musical revolution. Additionally, we chose the aggregated time-aware similarity between the influencer and his followers as the indicator for revealing the dynamic influencer. Results showed that **The Beatles** had the dominant influence **from 1960 to 2010**. Finally, we observed the features time series and verified that external events making impact on the musical evolution. For example, Internet and CD increased the popularity of music, and acousticness of music is decreasing with the development of musical technology, as more electrical amplification is applied.

Keywords: Network Analysis; Time Series; Musical Evolution;

Contents

1 Introduction

1.1 Background

Music is a vital component of human society, and has been changing in many aspects such as genres. Such evolution is relevant to a variety of factors, which includes the influence among artists and the impact of external events. Therefore, it's a meaningful task to quantify the musical evolution and understand the collaborative and temporal influence among artists and musics.

1.2 Problem Restatement

Our team is invited to design an effective model to demonstrate the evolution of music, specifically, we are required to accomplish the following tasks:

1. Construct a network of musical influence according to *influence_data* and explore critical parameters that descibe *influence*, then validate the effect of these parameters on a subnetwork.
2. Design a metric measuring the *similarity* between music, then compare the artists similarity within the same genre and across different genres according to the metric.
3. Explore the *similarity* and *influence* between and within genres, find the *salient feature* of a genre and the *correlation* between genres, meanwhile, identify the *temporal change* of genres.
4. Figure out whether the influencer *actually* influences the followers as shown in *influence_data*, and find *'contagious' music characterstics* if there is any.
5. Identify *characterstics* that signify revolutions in musical evolution, and discover revolutional *artists* in the network.
6. Analyse the musical evolution process in terms of that occured over time in one genre, describe the dynamic influencers and the *change* of this genre.
7. Explain how the cultual influence as well as external events such as social, political or technological changes are represented in the network.
8. Describe how our work can be extended if more data is provided and recommend further study of music and its effect on culture.

1.3 Our Work

In response to the requests, we have done the following:

Firstly, using the data in *influence_data*, we construct the **influence network**. Then a novel metric named **InfluenceRank** is proposed to capture artists' music influence in the network. Next, we design several measures to reveal the similarity of musics. Using these measures, further experiments are conducted to illustrate the similarity between artists within and across genres. Additionally, InfluenceRank and similarity measures are generalized and used for representating genres. Insights of genres are made as required and examples are provided for illustration. Afterwards, we utilize similarity information within the influence networks to inspect the confidence of 'influence', and identify contagious characters. Moreover, we use time series analysis to determine revolutionary features and artists. Influence process is also studied by designing a **time-aware** representation. Finally we discuss the impact of external events comprehensively.

2 Assumptions and Justifications

- The music is characteized only on the given factors in the data, regardless of other information.
- The influencer artists nearer to the 'source' of influencers is considered more important, meanwhile, the influencer artists connected to more followers are more important, so as the influencer artists whose followers are more similarr to him/her is considered more important.

3 Notations

Table 1: Notations.

Notation	Definition
G	An influence graph.
$\mathsf{PR}(\cdot)$	PageRank value of a node.
$\mathsf{IR}(\cdot)$	InfluenceRank of a node or a genre.
$\mathbf{s}$	A song's vector representation.
$\mathbf{a}$	An artist's vector representation.
$\mathsf{sim}(\cdot,\cdot)$	Two vectors' similarity.
$\mathsf{SIM}(\cdot)$ or $\mathsf{SIM}(\cdot,\cdot)$	Similarity within / between genres.

4 The Influence Network

In this section, we create a directed network $G = (V, E)$ to express the musical influence using the *influence_data*, where each node is an artist, and the directed edge from node x to node y indicates that x influences y. The network is visualized as Figure 1.

4.1 Music Influence and InfluenceRank

Two complementary parameters are proposed to capture the music influence in the network.

First of all, the most valuable information contained in the influence network is the connection between nodes i.e. the influence relationship among artists, which reveals the order of influence. Therefore, we propose to utilize topological ordering to capture the essence of musical influence. A **topological sequence** of a network $G = (V, E)$ is a linear order $<_T$ on its node set V. If $(x, y) \in E$, then $x <_T y$. Apparently, a network G can be topological sorted if and only if it is acyclic. Algorithm 1 is an algorithm computing a network's topological sequence from [1].

In the influence network, if a node x comes after node y, it means that the artist represented by node x is directly or indirectly influenced by that of y in the occasion of music influence. More

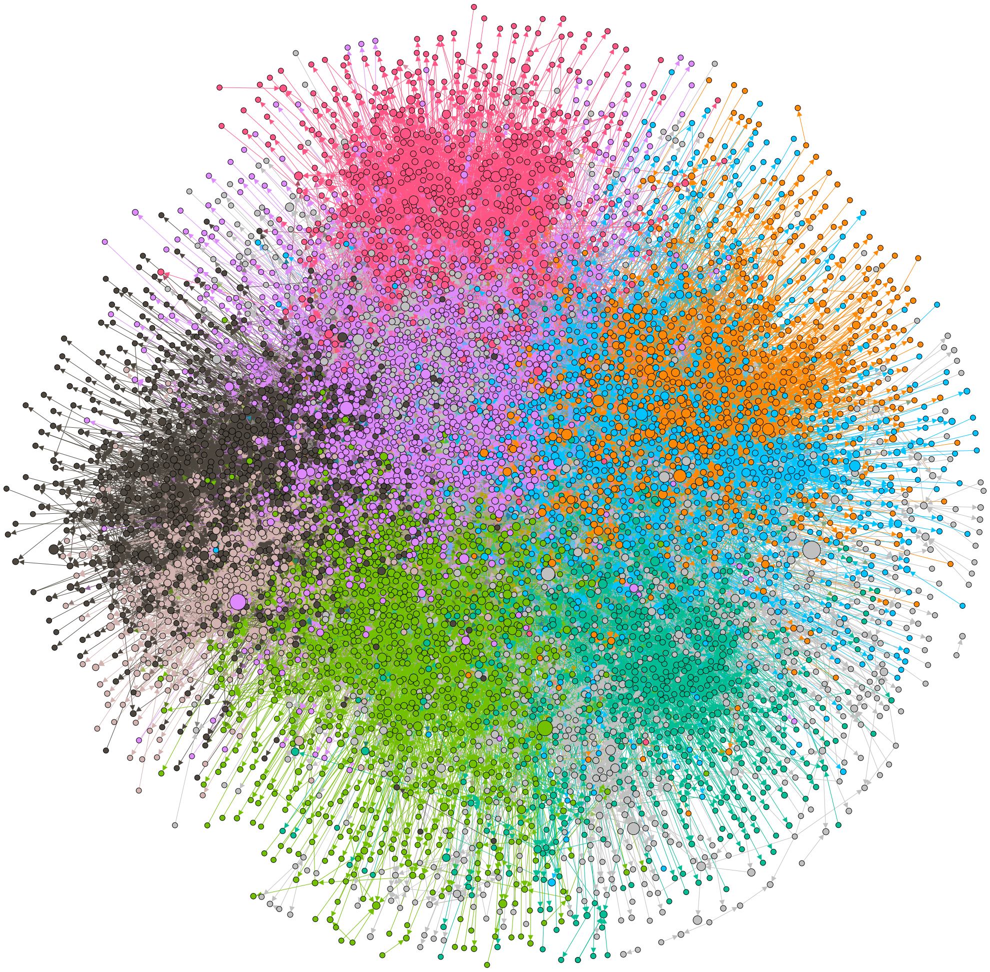


Figure 1: Visualization of the influence network.

Algorithm 1 TOPOLOGICAL-SORT

Input: Network $G = (V, E)$.

Output: A list of nodes.

```
1: Call DFS(G) to compute finishing times v.f for each vertex v as each vertex is finished, insert it onto the front of a list.
2: As each vertex is finished, insert it onto the front of a list.
3: return The list of nodes.
```

precisely, given a subnetwork of the influence network, a topological sequence in the subnetwork captures the sequential influencing order, whose head indicates the source influencer that contributes most to the development of the subnetwork. In the meanwhile, the topological sequence plays a big role in reflecting the importance of each artist in the subnetwork, as the significant artist will be followed by many artists, thus stands in the front of the sequence.

However, in the given data we find several records show that artists of earlier active start are influenced by artists of later active start, which results in a cycle in the whole network and fails the topological ordering. As a consequence, the topological sequence cannot be used in subnetworks which contain such reversed relationships. Therefore, a novel ranking method named **InfluenceRank** is derived from **PageRank** and is devised as an alternative to explore such sequential information in subnetworks with cycle.

PageRank [2] is a widely-used technique for ranking pages on the Internet, which can be viewed as ranking nodes in a large directed graph. Suppose that $\mathbf{M} = (M_{ij})_{n\times n}$ is the adjacency matrix of directed graph $G = (V, E)$, with $M_{ij} = 1$ only if there is an edge from node i to node j. Define the out-degree of each node as:

$$d_i = \sum_{j=1}^{n} M_{ij}. \tag{1}$$

Then the iteration formula of the PageRank is represented as follows:

$$p_i^{(k+1)} = \frac{1-\alpha}{n} + \alpha \sum_{j=1}^{n} \frac{M_{ij}}{d_j} p_j^{(k)} \tag{2}$$

$$\mathsf{PR}(i) = \lim_{k\to\infty} p_i^{(k)} \tag{3}$$

where $p_i^{(k)}$ denotes the PageRank value of the node i after k iterations ($p_i^{(0)} = \frac{1}{n}$), α is the damping parameter in $[0, 1]$.

The PageRank value for a node is higher in case that the in-degree of the node is bigger. Therefore, in music influence network, the node with lower PageRank is attached with more importance. Considering the value of PageRank may vary extensively because the whole network is too large, we transform the value of PageRank to its 'rank', namely **InfluenceRank**. The InfluenceRank of artist i is defined as:

$$\mathsf{IR}(i) = |\{n|\mathsf{PR}(n) \leq \mathsf{PR}(i)\}|, \tag{4}$$

where n denotes nodes of artists that has lower PageRank value than that of artist i, and $|\cdot|$ means a set's cardinality. Through this way, artists with smaller PageRank i.e. more important are ranked higher according to InfluenceRank, which is similar in topological ordering.

4.2 Subnetwork Validation

We use *influence_data* and choose artists active in the 2000s to build a influence network, then selecte three small subnetworks by picking weakly connected components in the network to demonstrate the topological sequence and InfluenceRank. The results are summarized in Table 2.

First, an acyclic (weakly) connected component of the whole influence network (A) is constructed. In this subnetwork, a topological order is ['724720', '158540', '1580437', '428329']. As for the InfluenceRank, the rank of these four nodes are $2, 2, 3, 4$ respectively, indicating artists with ID '724720' and '158540' are the source of influence who contribute most to the music influence in this subnetwork, and the rest artists are directly or indirectly influenced by them. The result of InfluenceRank is corresponding to the topological order, which means that the position of nodes in the topological order of this subnetwork can be represented by InfluenceRank.

Table 2: The PageRank value and InfluenceRank of nodes in subnetworks.

Subnetwork	Node ID	PageRank Value	InfluenceRank
A	158540	0.089220	2
	724720	0.175438	2
	1580437	0.250000	3
	428329	0.399123	4
B	863733	0.089220	2
	159340	0.089220	2
	355998	0.150822	4
	551115	0.150822	4
	688515	0.217419	5
	2517403	0.302497	6
C	894962	0.120883	2
	843259	0.120883	2
	1000302	0.172257	3
	662416	0.223632	4
	2527840	0.362346	5

Second, a (weakly) connected component of artists with a cycle (B) is extracted. The InfluenceRank sequence is [‘863733’, ‘159340’, ‘355998’, ‘551115’, ‘688515’, ‘2517403’] with ‘863733’ and ‘159340’ both ranking 2 as well as ‘355998’ and ‘551115’ both ranking 4, which matches the pseudo-topological order. The result reveals that in this subnetwork, artists with ID ‘863733’ and ‘159340’ both occupy an important position in the network of influence.

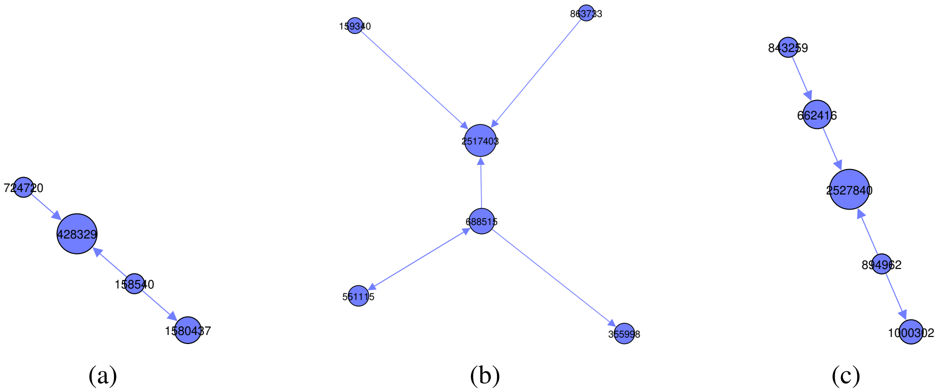
Finally, in the subnetwork composed by 5 nodes without any circle in it (C), a topological order of it is [‘843259’, ‘662416’, ‘894962’, ‘2527840’, ‘1000302’]. By comparing the InfluenceRank sequence shown in Table 2, we can find that the topological sort basically conforms to the order of InfluenceRank. Artists ‘843259’ and ‘662416’ are close to the source of relationship network of influence and are somehow more important to the evolution of music.

Through the comparison of topological order and InfluenceRank in these subnetworks, we can draw a conclusion that topological order and InfluenceRank has a positive correlation. Thus, We can use InfluenceRank instead of topological order as the parameter of influence. We can get the rank of artists in the network by applying InfluenceRank to the whole network. The result can be seen in Figure 2.

Artists with higher rank are more important in the influence network. Through this ranking, we can analyze the status and importance of different artists in the influence network to better understand the relation between music works.

5 Insight of Genres

In this section, we carefully design several metrics for music similarity and influence from different perspective, then further study the them between and within genres.



(a) (b) (c)

Figure 2: Three subnetworks.

5.1 Music Similarity

Artists' works are influenced by many factors, and the musical works of different artists may be similar in certain features. We represent the characteristics of each work in vector form in order to measure the similarity between musical works. Since there are songs of the same name by different composers in different years, we index each song.

Importantly, the value of each field in the musical features varies extensively from one another, which will impair the similarity measure. Therefore, normalization of features is necessary. We choose Z-Normalization [3] to regulate the input features. Given a raw musical features of song $\tilde{\mathbf{s}}_i$ as $(f_{i1}, f_{i2}, \cdots, f_{im})^{\mathrm{T}}$, the z-normalized vector can be calculated by:

$$\mathsf{ZNorm}(\tilde{\mathbf{s}}_i) = (\tilde{\mathbf{s}}_i - \boldsymbol{\mu})/\boldsymbol{\sigma}, \tag{5}$$

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{s}}_i, \tag{6}$$

$$\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \cdots, \sigma_m) \tag{7}$$

where $\boldsymbol{\mu}$ is the samples' mean value and $\boldsymbol{\sigma}$ is composed by standard deviations of all features respectively. Here, 'division'$(/)$ means divide two vectors' components one by one. Then for the song $\tilde{\mathbf{s}}_i$, it is represented as a normalized vector $\mathbf{s}_i \in \mathbb{R}^m$:

$$\mathbf{s}_i = \mathsf{ZNorm}(\tilde{\mathbf{s}}_i). \tag{8}$$

When evaluating the similarity of individuals, there are mainly two methods: distance and similarity. We choose **cosine similarity** as a measure of musical similarity. Compared to the distance measure, cosine similarity focuses on the direction difference between two vectors, rather than in distance or its length. Cosine similarity evaluates how similar two vectors are by calculating the cosine of the angle between them:

$$\mathsf{sim}(\mathbf{s}_i, \mathbf{s}_j) = \frac{\mathbf{s}_i^{\mathrm{T}}\mathbf{s}_j}{\|\mathbf{s}_i\|\|\mathbf{s}_j\|}, \tag{9}$$

and $\mathsf{sim}(\mathbf{s}_i, \mathbf{s}_j)$ is in range $[-1, 1]$, the bigger the value is, the smaller the angle is. And when $\mathsf{sim}(\mathbf{s}_i, \mathbf{s}_j)$ is close to 1, it means that the two musical works represented by the vectors are very similar. In this way, the similarity between musical works i and j can be represented by $\mathsf{sim}(\mathbf{s}_i, \mathbf{s}_j)$.

Similarly, given an artist $\mathbf{a}_i \in \mathbb{R}^n$, he or she can also be represented by the normalized features in the given data. Then the similarity between any pair of artist $\mathbf{a}_i$ and $\mathbf{a}_j$ can also be calculated by the cosine similarity between them.

5.2 Genre Similarities

5.2.1 Similarity Within Genres

As the similarity between artists can be losslessly captured in the cosine similarity of artist vector, it is quite intuitive to summarize the cosine similarity between each pair of artist within one genre to represent the overall similarity within the genre G_i, which can be fomulated as

$$\mathsf{SIM}(G_i) = \frac{\sum\limits_{\mathbf{u},\mathbf{v}\in G_i,\mathbf{u}\neq\mathbf{v}} \mathsf{sim}(\mathbf{u}, \mathbf{v})}{\binom{|V_i|}{2}}, \tag{10}$$

where $\mathsf{SIM}(G_i)$ is the **overall cosine similarity** of genre G_i, $\mathsf{sim}(\mathbf{u}, \mathbf{v})$ denotes the cosine similarity between artist $\mathbf{u}$ and artist $\mathbf{v}$. Partitioning artists by their main genre and calculating each genre's internal cosine similarity, the top-3 and last-3 most similar genres are listed in Table 3.

Table 3: Genres' overall cosine similarity.

Genre	Internal Cosine Similarity
Classical	0.6122
New Age	0.5532
Stage&Screen	0.5332
International	0.1094
R&B	0.1047
Pop/Rock	0.0793

Additionally, cosine similarity lacks the information of distance, which may impair the similarity score. Therefore, a promoted method which is called **Mean Square Error (MSE) similarity** is proposed to better reveal the similarity within genres from the distance perspective. To be specific, suppose there are $|G_i|$ artist in genre G_i, the mean square loss of a genre can be calculated as

$$\mathbf{c}_i = \frac{\sum\limits_{\mathbf{u}\in G_i} \mathbf{u}}{|G_i|} \tag{11}$$

$$\widetilde{\mathsf{SIM}}(G_i) = \frac{\sum\limits_{\mathbf{u}\in G_i} \|\mathbf{u} - \mathbf{c}_i\|^2}{|G_i|} \tag{12}$$

where $\mathbf{c}_i$ is the genre's 'artist center', $\widetilde{\mathsf{SIM}}(G_i)$ is the MSE similarity of genre G_i and $\|\cdot\|$ is the 2-norm Euclidean distance. The smaller the MSE is, the larger the similarity within genre is.

Partitioning artists by their main genre and calculating each genre's internal MSE similarity, the genres of top and last three least MSE similarity are showed as Table 4.

Table 4: Genres' overall MSE similarity.

Genre	Internal MSE Similarity
Unknown	4.7614
Children's	6.4390
Country	6.9746
New Age	18.3273
Comedy/Spoken	42.1325
Avant-Garde	45.1742

5.2.2 Similarity Between Genres

To compare similarity between genres, we must project the genres into the vector space in advance. Considering we have calculated the representation of each artist, the representation of the genre can be easily derived:

$$\mathbf{r}_i = \frac{\sum\limits_{\mathbf{u}\in G_i} \mathbf{u}}{|G_i|}, \tag{13}$$

where $\mathbf{r}_i$ is the representation of genre G_i. Afterwards, the similarity between any pair of genres G_i and G_j $(i \neq j)$ can be captured by cosine similarity between them:

$$\mathsf{SIM}(G_i, G_j) = \frac{\mathbf{r}_i^{\mathrm{T}}\mathbf{r}_j}{\|\mathbf{r}_i\|\|\mathbf{r}_j\|} \tag{14}$$

Partitioning artists by their main genre and calculating cosine similarity between them, we show several pairs of genres and their similarity in Table 5.

Table 5: Genre pairs' cosine similarity.

Genre #1	Genre #2	Cosine Similarity
New Age	Avant-Garde	0.9543
Vocal	Folk	0.9442
Classical	Stage&Screen	0.9283
Pop/Rock	Blues	-0.8023
Pop/Rock	Jazz	-0.8091
Pop/Rock	International	-0.9342

5.3 Genre Influences

Recall the two measure we proposed in section 4.1 to capture influence, i.e. topological ordering and InfluenceRank, we leverage both of them to illustrate the influences between and within genres.

5.3.1 Influence Within Genres

The most informative property of influence is the order of it, as it indicates who influences who. The InfluenceRank can well preserve such property. Thus, we explain influence within a genre by computing InfluenceRank of each node in the genre. For example, patitioning artists into genres, Table 6 shows the artists' InfluenceRank with genre 'Children's'.

Table 6: Within genre influence.

Node ID	InfluenceRank Within Genre
744969	2
808626	2
784597	4
888942	4

5.3.2 Influence Between Genres

In case of cross-genre comparison, it's no longer practical to calculate rank over the subnetwork of one genre. Instead, we compute each artist's InfluenceRank over the whole graph, add each genres' InfluenceRank together and re-rank them. To be precise, for a genre G_i, the InfluenceRank of it is the rank of summation of all of its nodes:

$$\mathsf{IR}(G_i) = |\{G_j| \sum_{u \in G_j} \mathsf{IR}(v) \leq \sum_{v \in G_i} \mathsf{IR}(u)\}|, \tag{15}$$

where $\mathsf{IR}(u)$ is the influence rank of node u. In this way, the influence is quantitied between genres.

5.4 Genre Character

As is requested, we explore distinctive features to identify a genre by using the technique of univariate feature selection. The result is presented in Table 7.

Table 7: Top 6 and last 3 distinctive features.

Features	Decisive Score
acousticness	189.5771
speechiness	179.9418
energy	141.6220
loudness	117.2777
instrumentalness	116.1389
popularity	87.2263
tempo	19.2836
mode	17.4847
key	2.4393

5.5 Genre Temporal Changing

The evolution of a genre can be reflected by the features of artists of it, so we propose to represent a genre by its artist feature vectors' center i.e. $C(G_i)$. To specify the temporal change of a genre, we divide the artists into different parts by periods of time Given influence network $G = (V, E)$, $V = \{a_1, a_2, \cdots\}$, a genre graph or genre for short is the subgraph where $G_i^{(T)} = (V_i^{(T)}, E_i^{(T)}) \subseteq G$ of a time period $T = [t_1, t_2)$, which can be fomulated as:

$$\mathbf{g}_i^{(T)} = C(G_i^{(T)}) = \frac{\sum_{\mathbf{u}\in V_i^{(T)}} \mathbf{u}}{|V_i|}, \tag{16}$$

where $V_i^{(T)}$ denotes artists that are active in the time period T. By comparing $C(G_i^{(T)})$ with different time period, the temporal change in genre can be revealed.

Using *influence_data* and *data_by_artist*, we devide the data into two periods $T_1 = [1920, 1961)$ and $T_2 = [1961, 2001)$ to illustrate the changing of genres. The results are in the Table 8.

Table 8: Changes of distinctive features.(only with top 5 distinctive features above)

Features	acousticness	speechiness	energy	loudness	instrumentalness
Avant-Garde	-1.009748	-0.200891	0.565449	2.475912	-0.281669
Blues	-0.601848	-0.399728	-0.012385	-0.065000	-0.363530
Children's	-0.853162	-0.857436	-0.529655	-0.832145	-0.025424
Classical	-0.357499	-0.226452	0.401016	1.306106	-0.440818
Comedy/Spoken	-0.772846	-0.032223	1.003460	1.007212	0.080651
Country	-0.912016	-0.047066	0.712743	0.738122	-0.193659
Easy Listening	-0.064749	-0.199962	-0.108447	0.748177	-0.963983
Electronic	-1.444784	0.355707	0.827598	1.481325	-1.570024
Folk	-0.270197	-0.000119	0.219097	0.315613	-0.120269
International	-0.735212	-0.341934	0.289281	0.197233	-0.308599
Jazz	-0.703647	-0.068097	0.444506	0.225616	0.158202
Latin	-0.889380	-0.059358	0.648450	0.641586	-0.436314
New Age	0.556570	-0.020463	-0.302088	-0.291400	-0.036549
Pop/Rock	-0.674007	0.134573	0.691185	0.635392	0.085057
R&B	-0.836967	0.336438	0.281768	0.362426	-0.034110
Reggae	-0.555315	0.924070	0.434076	0.424242	-0.329292
Religious	-0.737690	-0.256635	0.332073	0.727615	-0.267354
Stage&Screen	-0.525560	-0.336959	0.210590	-0.049381	1.148454
Unknown	NaN	NaN	NaN	NaN	NaN
Vocal	-0.228743	-0.193346	0.090863	-0.074616	-0.087005

5.6 Genre Correlation

There is a mature theory in statistics that can be applied directly to calculate the correlation of different genres, correlation coefficient [5]. Since the correlation between genres is hard to tell when focusing on a single time period, we propose to calculate correlation coefficient between genre representations of different time to get the correlation matrix. In practice, we first divide the data into two disjoint parts: $T_1 = [1920, 1961)$ and $T_2 = [1961, 2001)$, when similar number of artists are active in either, and calculate the representation of each genre with given T_1 and T_2, forming $C(G_i^{(T_1)})$ and $C(G_i^{(T_2)})$, then correlation coefficient is computed between the two vectors, the visualized result is presented in Figure 3.

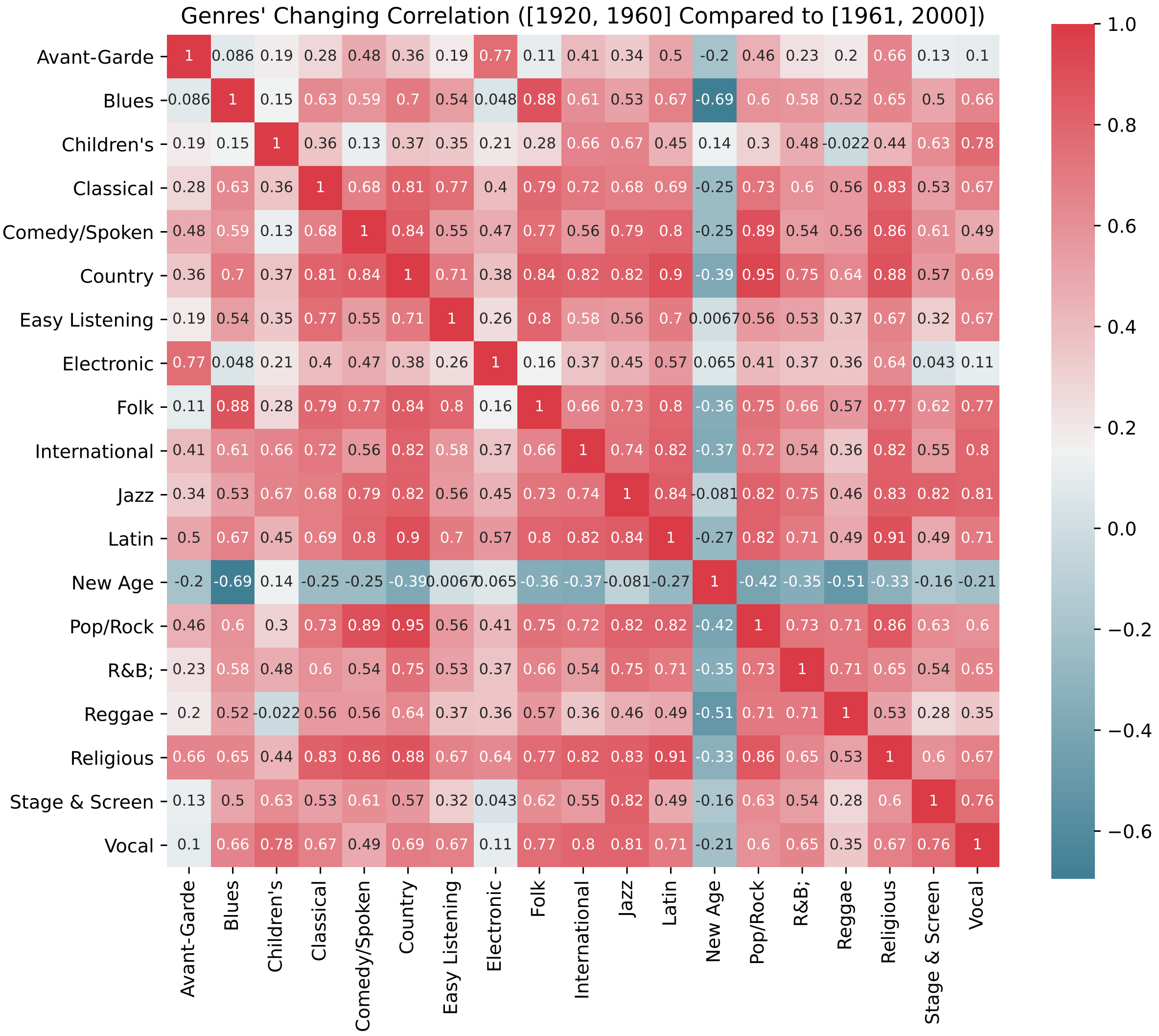


Figure 3: Genres' Changing Correlation.

In Figure 3, the coordinates (rows and columns) corresponding to the unit of brighter color are more related. According to Figure3, the *Latin* and *Jazz* are highly related (the correlation coefficient is 0.84), which makes sense because they are both dancing music.

6 Insights of Musical Evolution and Revolution

6.1 Influence Validation

After quantifying both influences and similarities, we notice that there are implicit correlations between them. In this subsection, we are devoted to verify whether the influencers actually influence their followers to some extent.

On the one hand, we focus on the representation of followers. Although we have represented artists in $\mathbb{R}^n$, it is insufficient to use a single vector because such representation loses the information of time. Therefore, resembling the genre representation above, we come up with a novel method to project follower artists with their released musics. From the granularity of musics, we are able to identify the temporal change within a artist, thereby identify if he is influenced by the influencer. To be specific, for an follower artist, the first and last songs he released are $\mathrm{s}_i^{(f)}$ and $\mathrm{s}_i^{(l)}$ respectively.

On the other hand, we focus on the representation of influencers. We propose to unify all of the songs $\mathbf{a}_i$ released (denoted as $S_i = \{\mathbf{s}_{i,1}, \cdots, \mathbf{s}_{i,N_i}\}$) to embed him / her into $\mathbb{R}^m$ from the granularity of musics. Formally,

$$\bar{\mathbf{a}}_i = \mathsf{style}(\mathbf{a}_i) = \frac{\sum_{j=1}^{N_i} \mathbf{s}_{i,j}}{|S_i|}. \tag{17}$$

Finally, given a influencer who is represented by $\bar{\mathbf{a}}_i$, a follower who is represented by $\mathrm{s}_j^{(f)}$ and $\mathrm{s}_j^{(l)}$, the influence, or the relative distance difference between them is

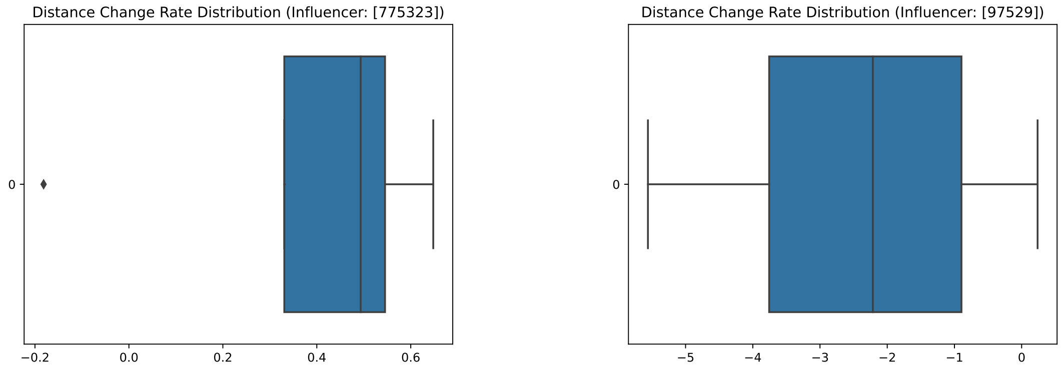
$$\mathsf{influence}(\mathbf{a}_i, \mathbf{a}_j) = \frac{\|\mathrm{s}_j^{(f)} - \bar{\mathbf{a}}_i\| - \|\mathrm{s}_j^{(l)} - \bar{\mathbf{a}}_i\|}{\|\mathrm{s}_j^{(f)} - \bar{\mathbf{a}}_i\|}, \tag{18}$$

When $\mathsf{influence}(\mathbf{a}_i, \mathbf{a}_j)$ is larger than 0, it impies the distance is reducing, while the negtive $\mathsf{influence}(\mathbf{a}_i, \mathbf{a}_j)$ suggests the distance is expanding. The above two kings of influencer are generalized as 'pull' influencer and 'push' influencer, both of which play a role in actually influencing the followers. To verify, we aggregated the $\mathsf{influence}$ from a specific influencer to all of his followers into a distribution. For space concern, we visualize two examples in Figure 4. In both figures, the 'influence' is **real** since the mean value of sampled influencer is not equal to 0.

Furthermore, to identify contagious features, a similar strategy as section 5 is adopted. Given the normalized musical feature vector of year i as $\mathbf{y}_i = (f''_{i1}, f''_{i2}, \cdots, f''_{ik})$, the change range of l-th feature $f''_{\cdot l}$ between year i and year j can be fomulated as:

$$\Delta_{i,j}^{l} = \frac{f''_{il} - f''_{jl}}{f''_{il}}. \tag{19}$$

The distribution of the change range of features is visualized in Figure 5. We compare the change range among every features then identify the top 5 features as 'contagious' ones, which are ***loudness*** and ***energy***. Our findings are the same as the experts[6] in musical analysis field.



(a)

(b)

Figure 4: ‘Pull’ influencer and ‘push’ influencer.

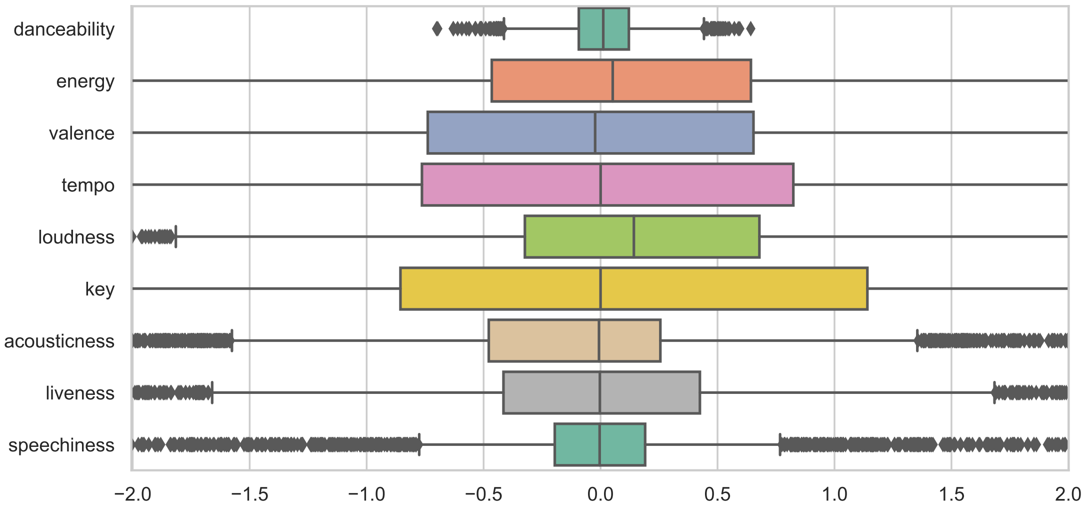


Figure 5: Change of music characteristics.

6.2 Musical Revolution

The temporal change of musics (revolutions included) can be reflected in the mainsteam change of each its features. Therefore, we present each feature of music as a function of year in a unified figure 6 according to the *music_by_year* data.

As the figure shows, some characters are steady while the others change drastically. However, revolution refers to the ‘Major Leap’ rather than steadily increase, which suggests that only observing the increasing or decreasing trend in the given figure is not inadequate. Therefore, we use first-differencing to inspect the change is ‘steady’ or not. Formally, given the unnormalized musical

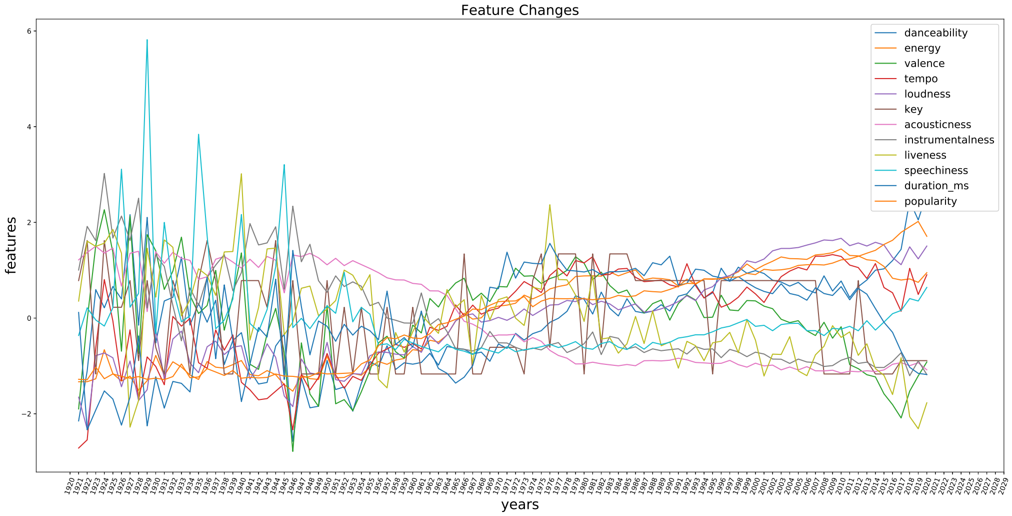


Figure 6: Features' time series.

feature vector of year i as $\mathbf{y}_i = (f''_{i1}, f''_{i2}, \cdots, f''_{ik})$, the first difference of $\mathbf{y}_i$ is:

$$\partial \mathbf{y}_i = \mathbf{y}_i - \mathbf{y}_{i-1}. \tag{20}$$

By first differencing, the speed of change is revealed explicitly. Then we can perform time series analysis to determine whether the changing speed of each feature is stationary. Only when the series is not stationary i.e. it is closely related to time, the corresponding musical features is changing with acceleration during that period of time. Besides, we quantify the stationary of a time series using Augmented Dickey Fuller test [4] and keep the derived MacKinnons approximate p-value. The higher the p-value is, the more unstationary the time series is. Specifically, to identify the revolutionary musical features, we calculate the p-value of each time series where the unnormalized value of the feature is the function of year from 1921 to 2020. The result is quite surprising, only two features i.e. *acousticness* of $p = 0.55$ and *speechiness* of $p = 0.357$ are unstationary, while p values of others are all close to 0.

Moreover, the 'revolution age' i.e. the decade of 'major leap' in musical features can be accurately identified by similar approach. Because the identified two features are considered to signify musical revolution while others not, we compute the p-value in each decade for *acousticness* and *speechiness* to mine temporal bound of revolution. In order to select the most revolutionary decades in history, each decade is attched to the revolution degree of weighted sum of its overall p-value:

$$\mathsf{Revolution}(\overline{\mathbf{y}}) = \overline{p}Q \in \mathbb{R}^{10}, \tag{21}$$

where $\overline{p}$ is the overall p-value of each feature as mentioned above and $Q \in \mathbb{R}^{m \times 10}$ is the aligned p-value of each feature in each decade. Each of the element in $\mathsf{Revolution}(\overline{\mathbf{y}})$ represents the revolution degree of the corresponding decade. We find $\mathbf{1960} \sim \mathbf{1970}$ is the most revolutionary decade.

6.3 Revolutionary Artists

As we have identified the most revolutionary decade in history, it's intuitive to infer that the revolutionary influencers are expected to release more musics than ordinary ones. Therefore, we quantify the revolutionary of an artist by counting the piece of work he released during the evolution era as determined above. The more musics he produced, the more revolutionary he is. Given an artist a_i and the musics he released during revolutionary decade are $S_i = \{s_{i1}, s_{i2}, \cdots\}$, the revolutionary he represents can be quantified by:

$$\text{Revolutionary}(a_i) = |S_i| \tag{22}$$

We calculate the revolutionary of each artists and list the top 5 revolutionary artists in Table 9;

Table 9: Top 5 revolutionary artists.

Name	Number of Artworks in revolutionary decade
The Beach Boys	429
Frank Sinatra	388
Bob Dylan	308
Johnny Cash	283
The Beatles	275

6.4 Evolution of Music

In this section, we explored the changes in the influence of music over time by analyzing the changing trend of musical features. And We find the factors that can identify dynamic influencers, as well as explain the changes of artists and genres.

6.4.1 Influence Process in Genre Jazz

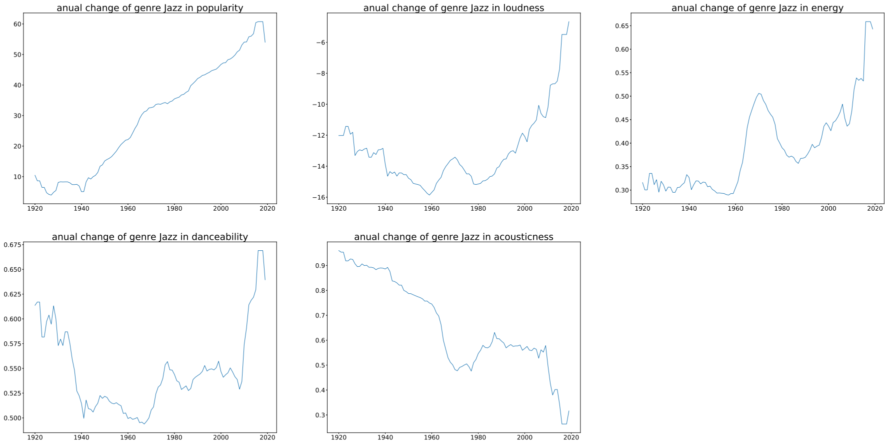


Figure 7: Evolution in genre Jazz.

In the development of music, changes in features are related to the music created by the artists, which in turn has an impact on some genres. The characteristics of the music of a genre at a certain time can be seen as the result of the influence on the artists belonging to this genre. Therefore, we can analyze the influence process in the genre by exploring the trend of feature changes, and illustrate the effect of the influence by comparing the similarity between the change in genre and the overall music evolution.

We choose genre Jazz as the research object and analyses its influence process. We get the set SG_y consists of songs that are created by artists of genre Jazz in year y using data in $influence\_data$ and $full\_music\_data$. For the song si in the set SG_y, its musical features is $\tilde{\mathrm{s}}_i$ (represented as $(f_{i1}, f_{i2}, \cdots, f_{im})^{\mathrm{T}}$). The genre's features in the year y are represented by the mean value si:

$$GF_y = \frac{\sum_{\tilde{\mathbf{s}}_i \in SG_y} \tilde{\mathbf{s}}_i}{|SG_y|} = (fg_{y1}, ..., fg_{ym})^T \tag{23}$$

For each feature of the genre Jazz, we use line charts to show the evolution over time. Part of the results are shown in Figure 7.

From Figure 7, we can get information that *acousticness* of Jazz shows a decreasing. This may because with the maturity of technology, as people tend to add more technology enhancements or electrical amplification, which makes acousticness decreases. Also, we can find Jazz's rise in popularity, which means being played more frequently. As for loudness, there was a general trend of decline from 1920 to 1960, but it increases after 1980s. The energy of Jazz gradually increased over time and increased suddenly in the 1970s, which may be related to the rise of Jazz-Rock at that time. Since then, jazz has developed many branches and the songs have become faster and louder, leading to higher loudness values in the jazz genre. Jazz had a high value of danceability between 1920 and 1930, indicating that it was very danceable, but remained low level after that.It grew rapidly after entering the 2000s, this may have something to do with the fact that jazz began to blend all kinds of music during this period and became more rhythmic.

Then we calculate the correlation between Jazz and all music data in each attribute, which is shown in Table 10 and the results observed from the line chart can be verified. The trend of Jazz popularity is highly consistent with overall development trend of music, where the change of *acousticness* is similar to the revolution of music. The changes in *loudness*, *energy* and *danceability* are also consistent with the overall trend, indicating that Jazz music has been greatly influenced by these features. However, Jazz and the whole music development differs *tempo* and *liveness*.

6.4.2 Indicators of Dynamic Influencer

The influencers in the influence network change over time. The major influencers of each period had significant influence on the contemporary artists, which means the music produced by the followers would have features similar to those of the influencers. In order to reveal the dynamic influencers, we calculate the influencer-follower similarity pairs in a certain period, and select the influencer whose works have the greatest similarity with his followers'. In this part, we also use cosine similarity to calculate the degree of similarity. We take the sum of the similarity as indicators that reveal the dynamic influencers. After calculating the total similarity between each influencer's work and that of his followers, we list the most influential artists of each decade in Table 11. Since there were few artists in the data for the 1920s, we only find two influencers.

Table 10: Similarities between Jazz and the Revolution in music.

Features	Correlation
danceability	0.655331
energy	0.737065
valence	0.276654
tempo	-0.232799
key	0.272979
acousticness	0.887396
instrumentalness	0.435427
liveness	-0.015111
speechiness	0.292435
duration_ms	0.538429
popularity	0.990745

Table 11: Influencers of the Decade.

decades	Top3 Influencers
1920s	Ella Fitzgerald, Billie Holiday
1930s	Frank Sinatra,Billie Holiday, Cab Calloway
1940s	Billie Holiday, Johnny Cash, Ella Fitzgerald
1950s	Charlie Parker, Dizzy Gillespie, Billie Holiday
1960s	The Beatles, Bob Dylan, The Rolling Stones
1970s	The Beatles, Bob Dylan, The Rolling Stones
1980s	The Beatles, Bob Dylan, The Rolling Stones
1990s	The Beatles, Bob Dylan, The Rolling Stones
2000s	The Beatles, Bob Dylan, The Rolling Stones
2010s	The Beatles, Stevie Wonder, Michael Jackson

From the result we could see that in 1920s and 1930s, the main genre of the music is Vocal. In 1930s, Cab Calloway, who is a Jazz musician, had influenced many people. In the 1940s, in addition to Vocal genre, composers of Country genre also occupied an important position in the music influence network. Johnny Cash was a major influencer, whose works highly similar to his followers, indicating that Country music was influencing that time. In the 1950s, the influence of music mainly came from Jazz, but the influence of Vocal sill couldn't be underestimated. In 1960 and for the next half century, Pop/Rock was the absolute dominant influence in music. The Beatles, Bob Dylan, and The Rolling Stones had influenced generations of artists. When it comes to 21 century, Stevie Wonder and Michael Jackson's music lead to the boom in *R&B*. Lots of artists of this time were influenced by them and followed them in music creation, while the influence of The Beatles continues, who still remained a major influence in The 2010s.

6.5 Identify Culture Influence

Cultural factors are reflected through the mutual influence of artists. Music is created by human beings, and artists interact with each other, engaging in political or social issues. Therefore, the analysis of the influence between artists is able to analyze culture influence of music.

Our work reveals the cultural influence of music in time or circumstances by measuring the effect of influencers on music works created by followers. Through our work, we find that the features of the works tend to be consistent with those of influencers to a certain extent, indicating that the music that has been created earlier has more or less influenced the later artists.

Technological change is closely related to the dissemination of musical works. For example, the advent of the Internet has allowed music to reach more people, which makes the influence of music much larger and wider. In the music influence network, when a genres influence range expands significantly, or an artist has more followers, it may be attributed to social, political, or technological change. In addition to changing musical characteristics through changing influence, these factors also directly affect musical characteristics. Our analysis found that music popularity increased significantly in 1950s and 1990s.This is related to the advent of radio, vinyl records and multi-track recorders in 1950-1960 which made music more widely spread. And the birth of MP3 technology in 1900s made it possible for music to spread over the Internet. Political and social changes have similar effect. In addition, we also see a decreasing trend of musical acousticness, indicating the wide application of technology enhancements or electrical amplification in music production as technology progresses. Thus, when new genre of music emerges, or a spencial change happens in features of a certain genre, we can identify that some changes have taken place during this period.

7 Model's Strengths and Weaknesses

7.1 Strengths

- **Our factorization approaches are accurate and practical.** We propose different ways to project musics, artists and genres to vector space by manipulating the given data. Temporal information is also included in specific representations. Meanwhile, it's easy to extend our representation methods if more data is available.

- **Our model innovates in designing influence rank and least square similarity to measure musical influences and similarities.** Based on topological ordering and PageRank, we well-design several metrics to accurately capture the musical influence and a variety of similarities. We then present least square loss and other similarity measure to exploit similarity information from different perspectives. Additionally, computational complexity is also taken into consideration. And rich examples validate the effectiveness of our approach.

- **Our model is reasonable and justifiable.** Results of all experiments are self-contained and explained comprehensively.

7.2 Weaknesses

- Some fine-grained information is sacrificed because of limited computation resources.

- The analysis over the data itself is poor. We will consider more to select informative features in factorization if more data is provided.

8 Document to ICM Society

Music is important part of human society and has evaluated over time. There are many sources of inspiration that stimulates the artists to create music, and sometimes they can be influenced by other artists. It is important to understand music influence, which can help to understand evolutionary and revolutionary trends of artists and genres. However, the influence of music is difficult to describe, and it is hard to determine the role that different artists play in the influence network, as well as the results of that influence from the data of influencers and followers. Our method with low complexity can be applied to many problems related influence network and evaluate the influence between artists.

As artists often influence each other, it is difficult to find the absolute origin of influence. Our approach cleverly solves the problem, by proposing a "Influence Rank" for each artist in the influential network, which can also be used in the network with loops. And that rank can show his or her position in the music influence network.

To measure the role influencers play in the music created by their followers, it is useful to measure the similarities between the works of artists. We propose a method to measure the similarity of musical compositions by applying cosine similarity. We measured the similarities between songs, as well as the similarity of musical works by composers between and within genres. The similarity calculated by our method can be used to analyze changes in artists, musical works and genres to help us insight into music influence networks. Through calculated similarity, we analyze the similarities and differences between the different genres, as well as their interaction over time. Also, we combine similarity with Influence Rank to assess the role of influencers in the network and identify the most contagious features to gain insight into the process of artists influence on others music work.

Our analysis of the trend of music genre change, combined with the influence process, explains the change of artists and genres over time. Through our approach, we can understand the factors that lead to the emergence of a new genre to some extent, leading us to have a deeper understanding of the development history of music. With more or richer data, we will take more musical characteristics into count and classify music genres according to these characteristics. Artists of unknown genres can be classified according to the style of their musical works. With sufficient data, it is possible to analyze the style of an artist's work at different times to understand the influence of music on him. Artists create music to express their feelings or opinions, often reflecting their thinking about society, culture and other aspects. So one direction of future work is to focus on whether the meaning of lyrics and music themes have an impact on culture and what kind of impact they have. And different countries and peoples have different musical styles, but musical characteristics change over time. Future research can focus on whether the influence of music makes them tend to be integrated or become more distinctive, and how the culture of countries and nations has changed, so as to further explore the influence of music on culture.

References

[1] Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). Introduction to algorithms. MIT press.

[2] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab.

[3] E. Kreyszig (1979). Advanced Engineering Mathematics (Fourth ed.). Wiley. p. 880, eq. 5. ISBN 0-471-02140-7.

[4] MacKinnon, J.G. 1994. Approximate asymptotic distribution functions for unit-root and cointegration tests. Journal of Business and Economic Statistics 12, 167-76.

[5] Croxton, Frederick Emory; Cowden, Dudley Johnstone; Klein, Sidney (1968) Applied General Statistics, Pitman. ISBN 9780273403159 (page 625)

[6] Serrà, J., Corral, Á., Boguñá, M. et al. Measuring the Evolution of Contemporary Western Popular Music. Sci Rep 2, 521 (2012). https://doi.org/10.1038/srep00521

Appendices

Appendix A Code(s) Used in Our Model

Python source:

```
import csv
import networkx as nx
from tabulate import tabulate
import numpy as np
import pandas as pd
# ------------------------------Problem 1------------------------------
def build_influence_graph(influence_file, period = range(1920, 2021)):
    csv_file = open(influence_file)
    csv_file.readline()
    csv_reader = csv.reader(csv_file)
    list_inf_fol = [ (row[0], row[4])
        for row in csv_reader
        if int(row[3]) in period
        and int(row[7]) in period]
    G = nx.DiGraph()
    G.add_edges_from(list_inf_fol)
    return G
def pagerank(graph, return_table = True):
    pr_dict = dict(sorted(nx.pagerank(graph).items(), key = lambda x: x[1]))
    if return_table:
        pr_table = tabulate(pr_dict.items(), headers = [’id’, ’PageRank’],
           tablefmt = ’psql’)
        return pr_table
    else:
        return pr_dict
def influence_rank(graph, return_table = True, method = ’max’):
    pr_dict = pagerank(graph, return_table = False)
    df_pr = pd.DataFrame.from_dict(pr_dict, orient = ’index’)
    df_ir = df_pr.rank(method = method)
    if return_table:
        return tabulate(df_ir, headers = [’id’, ’InfluenceRank (PageRank Order
           )’], tablefmt = ’psql’)
    else:
        return df_ir.iloc[:, 0].to_dict()
# ------------------------------Problem 2------------------------------
def fetch_music_characteristics(file):
    df_music = pd.read_csv(file)
    df_music_droped = df_music.drop(
        columns = [’artist_names’, ’artists_id’, ’year’,
            ’release_date’, ’song_title (censored)’])
    df_music_norm = (df_music_droped - df_music_droped.mean()) /
       df_music_droped.std()
    return df_music, df_music_norm
def numpy_cosine_similarity(a, b):
    sim = a.dot(b) / (np.linalg.norm(a) * np.linalg.norm(b))
    return sim
# ------------------------------Problem 3------------------------------
def within_genre_similarity(genre_artists_dict, genre, method = ’cosine’):
    # Get artist vector
    artists_vec_dict = genre_artists_dict[genre]
    similarity_sum = 0
    if method == ’cosine’:
        artists_vec_list = list(artists_vec_dict.items())
        for i in range(len(artists_vec_list)):
            for j in range(i + 1, len(artists_vec_dict)):
                similarity_sum += numpy_cosine_similarity(
                    artists_vec_list[i][1], artists_vec_list[j][1])
        n = len(artists_vec_dict)
        n = n * (n - 1) / 2
        return similarity_sum / n
    elif method == ’dist’:
        df_vec = pd.DataFrame.from_dict(artists_vec_dict, orient = ’index’)
        center = df_vec.mean()
        for i in range(len(df_vec)):
            similarity_sum += ((df_vec.iloc[i] - center).dot(df_vec.iloc[i] -
               center))
        n = len(artists_vec_dict)
        return similarity_sum / n
def between_genre_similarity(genre_artists_dict, genre1, genre2):
    artists_vec_dict1 = genre_artists_dict[genre1]
    artists_vec_dict2 = genre_artists_dict[genre2]
    df_vec1 = pd.DataFrame.from_dict(artists_vec_dict1, orient = ’index’)
    df_vec2 = pd.DataFrame.from_dict(artists_vec_dict2, orient = ’index’)
    center1 = df_vec1.mean()
    center2 = df_vec2.mean()
    return numpy_cosine_similarity(center1.to_numpy(), center2.to_numpy())
```

```
def within_genre_influence(genre_subgraph):
    return influence_rank(genre_subgraph)
def between_genre_influence(graph, id_genre_dict):
    ir_dict = influence_rank(graph, return_table = False)
    genre_sumrank_dict = dict()
    for artist_id, rank in ir_dict.items():
        if id_genre_dict[artist_id] not in genre_sumrank_dict:
            genre_sumrank_dict[id_genre_dict[artist_id]] = [0, 0]
        genre_sumrank_dict[id_genre_dict[artist_id]][0] += rank
        genre_sumrank_dict[id_genre_dict[artist_id]][1] += 1
    for key in genre_sumrank_dict.keys():
        genre_sumrank_dict[key] = genre_sumrank_dict[key][0] /
           genre_sumrank_dict[key][1]
    df_genre_influence = pd.DataFrame.from_dict(genre_sumrank_dict, orient = ’
       index’)
    df_genre_ir = df_genre_influence.rank(method = ’max’)
    df_genre_ir = df_genre_ir.sort_values(df_genre_ir.columns[0])
    return tabulate(df_genre_ir, headers = [’id’, ’InfluenceRank’], tablefmt =
        ’psql’)
# ------------------------------Problem 4------------------------------
def artists_mean(id_work_dist):
    id_mean_dict = dict()
    for key in id_work_dist.keys():
        id_mean_dict[key] = id_work_dist[key].mean()
    return id_mean_dict
def artists_change(id_work_dist):
    id_change_dict = dict()
    for key in id_work_dist.keys():
        df = id_work_dist[key]
        df[’release_date’] = pd.to_datetime(df[’release_date’].apply(
           transformDate), format = ’%Y-%m-%d’)
        df = df.sort_values(by = ’release_date’, ascending = True)
        if (df.iloc[-1, -1] > df.iloc[0, -1]):
            id_change_dict[key] = (df.iloc[0, :-1], df.iloc[-1, :-1])
    return id_change_dict
# ------------------------------Problem 5------------------------------
def get_genre_repr_year(period, full_music_data, ger2id, art2ger):
    genre_repr = {i:[] for i in range(len(ger2id))}
    for i,row in tqdm(full_music_data.iterrows()):
        artists = row[’artists_id’][1:-1].split(’,’)
        date = datetime.strptime(row[’release_date’],"%Y-%m-%d")
        if date < period[1] and date >= period[0]:
            for a in artists:
                try:
                    genre = art2ger[int(a)]
                    genre_repr[ger2id[genre]].append(row.iloc[2:-3])
                except:
                    pass
    for i in range(len(ger2id)):
        if genre_repr[i] == []:
            genre_repr[i] = [pd.Series([0] * 14,index = full_music_data.
               columns[2:-3])]
    genre_repr = {k : pd.concat([j.to_frame().transpose() for j in v],axis=0)
       for k,v in genre_repr.items()}
    genre_repr = {k: v.mean() for k,v in genre_repr.items()}
    return genre_repr
# ------------------------------Problem 6-------------------------------
def get_artist_repr_year(period, full_music_data, id2art):
    artist_repr = {k:[] for k in id2art.keys()}
    for i,row in tqdm(full_music_data.iterrows()):
        artists = row['artists_id'][1:-1].split(',')
        date = datetime.strptime(row['release_date'],"%Y-%m-%d")
        if date < period[1] and date >= period[0]:
            for a in artists:
                try:
                    artist_repr[int(a)].append(row.iloc[2:-3])
                except:
                    pass
    artist_work_count = {k: len(v) for k,v in artist_repr.items()}
    for k in id2art.keys():
        if artist_repr[k] == []:
            artist_repr[k] = [pd.Series([0] * 13,index = full_music_data.
               columns[2:-3])]
    artist_repr = {k : pd.concat([j.to_frame().transpose() for j in v],axis=0)
         for k,v in artist_repr.items()}
    artist_repr = {k: v.mean() for k,v in artist_repr.items()}
    return artist_repr, artist_work_count
```