Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
D		2110013

Music Influence, Similarity and Evolution Analysis: From the Perspective of Network and Statistics

Summary

When we take a dip in the long history, it is not a surprise to see that music has been evolving all the time. We try to build a series of models based on social network analysis and statistics to measure the influence-follow action between artists, genre music similarity and music evolution process.

First, referring to the popular social influence analysis model, we build a music influence network. According to the relationship between influencers and followers, we define 3 features for each artist, including *influence strength*, *influence width* and *social affinity*. On this basis, we give static influence of nodes and music influence on edges by considering cascade propagation. In this way, we take artist as node, influence-follow action as edge, and music influence as edge weight to construct the directed network. Then we use 99 popular artists to visualize the subnetwork, and preliminarily infer the correlation between genres. For example, R&B may have two branches, one of which is related to Jazz and Reggae, while the other to Vocal and Latin. In addition, we propose a unified scheme based on greedy algorithm to find the seed set in influence network, which can be used to find revolutionaries.

Next, we put forward the measures of music similarity, which is defined as the standardized Euclidean distance or its transformation of music characteristic vector. Using t-test, we find that music of the same genre are significantly closer than that of different genres. Moreover, after hierarchical cluster analysis, we find that genres clustered as single cluster do have some characteristics can distinguish them. For example, *danceability* of Reggae is significantly larger than that of other genres. In addition, statistical tests on correlation of different genres from the perspective of similarity reveal that some genres have similar or opposite trend over time. For example, Jazz has a similar development trend as Vocal and Blues, while has an opposite trend to R&B.

Third, we explore the relationship between music similarity and influence behavior in the same genre. The results from t-test indicate influencers actually affect the music created by the followers significantly. Then we use logistic regression to infer which characteristics play an important role. Considering the limitations of linear model, we carry out experiments using XGBoost, which is a popular and high performance machine learning algorithm. The results show that after considering combination of features, *instrumentalness*, *valence* and *speechiness* are the more "contagious" characteristics when The Beatles is the influencer.

Last, we put emphasis on the process of musical evolution. In order to analyze the changes over decades in a rigorous way, we use the change points detecting technology called *bcp* to signify two revolutions, one of which occurred in 1950s and another in 2010s. In this part, we also combine the influence network together with similarity measure to build a dynamic influence model, which helps to identify the revolutionaries in each revolution. We can get the answers that Jazz and Pop/Rock artists dominated the first one, while emergence of new genres led to the second one. Moreover, we chose R&B genre to have a thorough understanding of musical evolution. Statistical conclusions matches the real history of R&B genre in predicting the appearance of a new kind. In addition, by analyzing musical characteristics, we connect the music industry with the surrounding society, such as culture and technology. Within documentary, we reveal the mutual influence effects between music and other fields, such as how Black culture in 1940s gave birth to a new kind of music R&B which in turns played a significant role in promoting Black culture.

The four main parts of our work are interrelated and progressive, forming a complete system. We are confident that our work can make contributions to the related research of music similarity and evolution.

Key Words: Musical Evolution; Static Influence; Dynamic Influence; Social Network Analysis; Greedy Algorithm; T-test; Hierarchical Clustering; Logistic Regression; XGBoost; Change Point Detection

Contents

1 Introduction

1.1 Background

Music, which is almost everywhere in human society, has played an important role in our life. It is not only a media of emotion, but also a witness to the history. When we take a dip in the long history, it is not a surprise to see that music has been evolving all the time. A transformative change of music may be owing to some creative artists, or may be caused by the improvement of technique, political revolution or even the birth of some philosophical trends. We may wonder, what is the relationship between these different genres of music? Do the influencers actually affect the music created by the followers? How does music change or be changed by the world and how do they influence each other? These questions have puzzled human beings for a long time.

1.2 Our Work

Section 2 describes the data sets, important variables, and key assumptions. In Section 3 where we answer Task 1, we build a network based on static and direct influence and propose an algorithm to find seed set as revolutionaries. In Section 4 where we answer Task 2 and Task 3, we develop measures of music similarity, and analyse trend of music over time. In Section 5 where we answer Task 4, we use Logistic regression and XGBoost to infer characteristics importance in influence behaviors. In Section 6 where we answer Task 5, Task 6 and Task 7, we analyze the process of musical evolution in a statistical way, meanwhile providing an effective method to identify revolutionaries by measuring dynamic influence. Section 7 gives brief conclusions of our work.

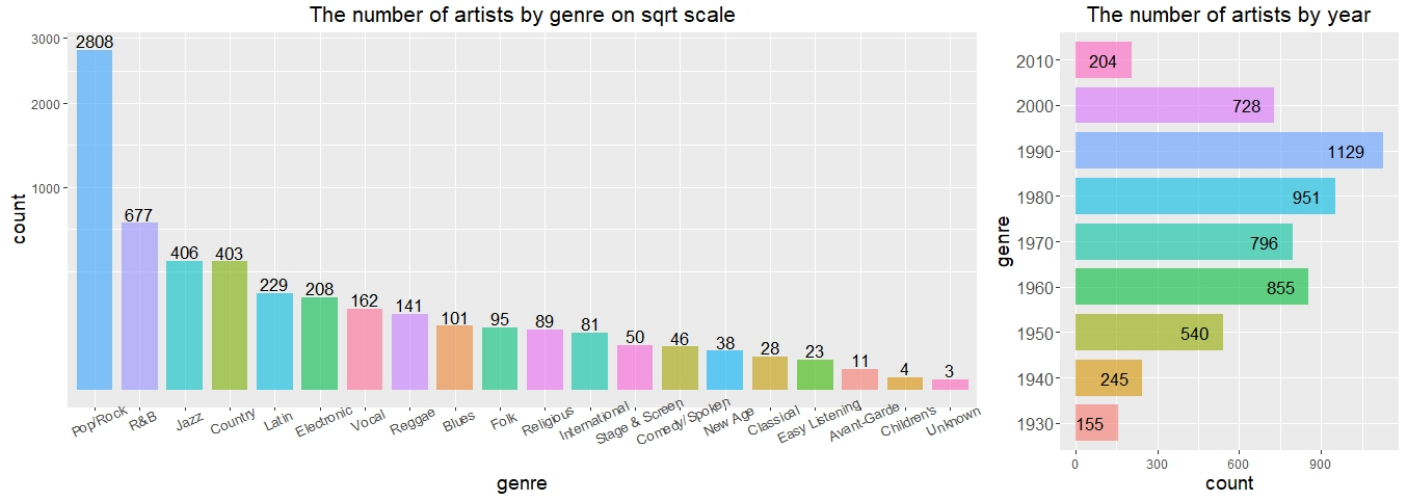
2 Preparations

2.1 Data Sets

Wwe have four datasets available. *influence_data* provides the corresponding information of influencers and followers for 5854 artists in the last 90 years. *full_music_data* provides 16 variable entries, including musical features such as *danceability*, *tempo*, *loudness*, and *key* for each of 98340 songs. *data_by_artist* and *data_by_year* are created from *full_music_data* by taking mean values by artist and across years respectively.

Based on *influence_data* data set, we show the number of artists by genre and decade in Figure 1. There are 3774 influencers, 5046 followers and 5603 artists in this data set. Here we consider both influencers and followers.

Figure 1: The number of artists by genre and decade.



According to Figure 1, the distribution of genres is unbalanced. We define the genre with more than 50 artists as the main genre, or important genre. In the later analysis of similarity and influence between and within genres, only main genres are considered. This is to prevent the existence of outliers in minority genres from leading to the decrease of stability.

2.2 Variable Descriptions

Many mathematical formulas are involved in our work, so we list important variables in Table 1 for query.

Table 1: Important variables used in this work.

Variable	Descriptions
S	The seed set in network G
$\mathcal{A}_u^I$	Set of influencers of node u
$\mathcal{A}_u^F$	Set of followers of node u
N_u	Set of neighbors of node u in G
t_u	The step that node u is activated. Here $1 \leq t_u \leq 9, t_u \in \mathbb{Z}_+$
KL_u^{year}	Kullback-Leibler divergence of influence from node u by year
KL_u^{genre}	Kullback-Leibler divergence of influence from node u by genre
Strength_u	Influence strength of node u
Width_u	Influence width of node u
Affinity_u	Social affinity of node u
R_u	The rank of number of followers of node u within genre
Popu_u^t	Popularity of node u in year t
α_u	Static influence of node u
$\beta_{v,u}$	Direct influence of node v on node u
$\Gamma_{v,u}$	Music influence of node v on node u
$\Gamma_{S,u}$	Total music influence of seed set S on node u
$\rho_{v,u}$	Music similarity between node / artist u and node / artist v
$\rho_{u,S}$	Music similarity between node / artist u and set / genre S
$\gamma_{u,S}^T$	Dynamic influence of node / artist u on set / genre S during decade T.

2.3 Key Assumptions

Before we start the work, we list the key assumptions in the subsequent modeling and analysis. These assumptions are based on literature research and life perception.

- We represent the influence network among artists as a directed graph $G = (V, E)$, where V is the set of nodes and E is the set of edges. Each node in G represents an artist and each directed edge represents a influence-follow action. If an edge $e \in E$ points from node v to node u, it means v is the influencer of u and u is the follower of v.

- The music influence between two artists consists of the direct influence and its cascade propagation. The direct influence is mainly determined by the static influence of influencer, taking into account the social activity of follower and the time attenuation effect, similar to the work [5].
- Similar to the method of social influence analysis, we assume network G is generated by a seed set S. The seed set S generates G by the way of influence spread. The nodes in S can be regarded as the revolutionaries of genres.
- To measure the similarity of different genres, we assume that all these data come from an infinite population, which is reasonable since the sample size is large enough. We further assume that each variable follows a normal distribution, whose mean equals the sample mean and variance equals the sample variance.
- When it comes to the persistence of artists' influence on the musical industy, there's a basic influence which doesn't change with other characteristics but the artist himself. In this artical, we choose the static influence as this unchanged influence.

3 Influence Network

In this section, we use *influence_data* data set to create a directed network of musical influence, which answers Task 1 in the whole problem.

3.1 Measures of Music Influence

In this subsection, we will give the music influence measure. The key assumptions of influence network are listed in subsection 2.3. The static influence of artist is determined by node features, which includes influence strength, influence width and social affinity.

- **Influence Strength**. $\mathcal{A}_u^F$ represents the set of followers of node u, so $|\mathcal{A}_u^F|$ equals to the number of followers. When an artist has more followers, we think he has a higher influence strength. In addition, we need to consider the effect of decades. If there are a large number of active and influential artists in the decade when one begins his music career, it will be more difficult to achieve a higher influence. So we define the influence strength as follows.

$$\text{Strength}_u = \sqrt{|\mathcal{A}_u^F|} \cdot \log(|\{v \in V : t_v = t_u\}|) \tag{1}$$

- **Influence Width**. An artist's influence can also be measured by the influence range. If an artist's followers are evenly distributed in the decade, we think his influence is more lasting. At the same time, if his followers come from more diverse genres, it shows his influence is more broad. First, we calculate the distribution of followers from different decades and genres. If the followers come from more than four decades, we put together the followers from the fourth decade or later. As for the distribution of genres, we only consider the number of followers from same genre and different genres, two types. We denote the distribution obtained above as P_u^{year} and P_u^{genre}. In order to measure the uniformity of these two distributions, we consider the KullbackLeibler divergence between them and uniform distribution, denoted as U. KullbackLeibler divergence is also called relative entropy. It is an asymmetric measure of the difference between two probability distributions. The KullbackLeibler divergence of two continuous distributions P and Q is defined as follows. The definition for discrete distributions is similar.

$$KL(P||Q) = \int P(x) \log \frac{P(x)}{Q(x)} \mathrm{d}x$$

Now we can calculate the KullbackLeibler divergence of influence from node u by year and genre as follows.

$$KL_u^{\text{year}} = KL(P_u^{\text{year}}||U) \cdot \mathbb{1}_{[0,1]}, \quad KL_u^{\text{genre}} = KL(P_u^{\text{genre}}||U) \cdot \mathbb{1}_{[0,1]}$$

We add the indicator function from 0 to 1 here to avoid the appearance of very large values and interfere with the result. Finally, we define the influence width as follows.

$$\text{Width}_u = \text{Logistic}(1 - KL_u^{\text{year}}) \cdot \text{Logistic}(1 - KL_u^{\text{genre}}) \tag{2}$$

Here the definition of logistic function is

$$\text{Logistic}(x) = \frac{e^x}{1 + e^x}$$

- **Social Affinity**. Social affinity is an important part of measuring social relationship. It indicates the degree of coincidence of one person's social circle with his friends. In this problem, we define the neighbors of one node as its followers and influencers. So we have $N_u = \mathcal{A}_u^I \cup \mathcal{A}_u^F$. To measure the coincidence degree of two sets, we define IoU (Intersection over Union) as follows.

$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Finally, we define the social affinity as follows.

$$\text{Affinity}_u = \frac{1}{|N_u|} \sum_{v \in N_u} IoU(N_u, N_v) \tag{3}$$

We defined three important node features above. Next, on this basis, we will define static influence, direct influence and music influence in turn.

- **Static Influence**. According to the above analysis, we get influence strength, influence width and social affinity are all positively realated with static influence. In addition, we need to consider the importance of artist within genre, because the number of artists in different genres is various. Artist in larger genre tends to have more followers. We need to identify the important revolutionaries in minority genres. R_u denotes the rank of number of followers within genre. So the static influence is given by

$$\alpha_u = \frac{\text{Strength}_u \cdot \text{Width}_u \cdot \text{Affinity}_u}{\sqrt{R_u}}. \tag{4}$$

- **Direct Influence**. Now we consider the direct influence of influencers on followers. If a follower is influenced by more influencers, the influence of a single influencer is more likely to decrease. At the same time, direct influence should include time attenuation effect, which is because the trend of music changes over time. So the direct influence is given by

$$\beta_{v,u} = \frac{\alpha_u}{\log(|\mathcal{A}_u^I| + 1)(\delta_{u,v} + 1)}. \tag{5}$$

Here we define

$$\delta_{u,v} = \begin{cases} t_u - t_v & , \quad t_u \geq t_v \\ 0 & , \quad otherwise \end{cases}$$

- **Music Influence**. When measuring music influence, we need to consider not only the direct influence between influencer and follower, but also the mode of cascade propagation. In this way, we define the music influence between influencer and follower as

$$\Gamma_{v,u} = \beta_{v,u} + \sum_{w \in \mathcal{A}_v^F} \mu \cdot \beta_{w,u}. \tag{6}$$

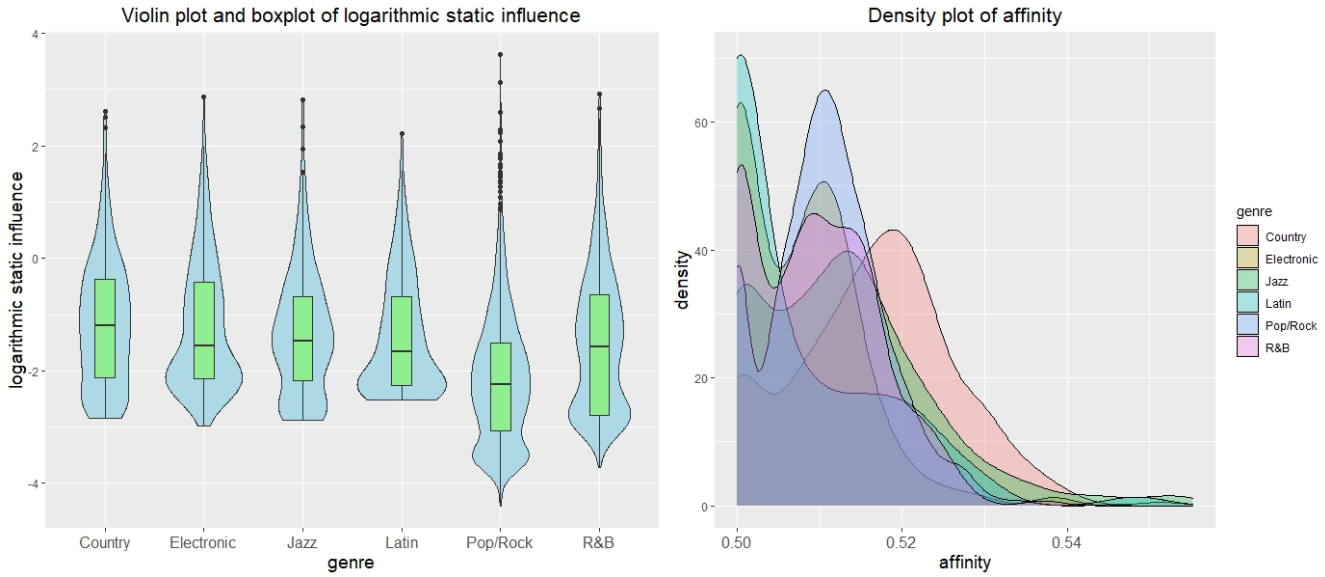
Here μ is the discount factor that needs to be defined in advance and we choose $\mu = 0.1$ as default. We can also define the total music influence of influencer set on follower as

$$\Gamma_{S,u} = \sum_{v \in S} \Gamma_{v,u}. \tag{7}$$

Now we have defined music influence measure. We take it as edge weight in network G. According to the network construction method in this subsection 2.3, we have completed the directed network of musical influence.

Before formally visualizing the subnetwork, we may as well make exploratory data analysis on the node features defined and calculated above. In Figure 2, we give the violin plot of logarithmic static influence and the density plot of affinity. In order to ensure the block size, we only select the 6 genres with the largest number. As can be seen from the left figure, static influence of *Pop/Rock* is widely distributed; *Country* and *R&B* tend to have a higher static influence on the whole. From the right figure, we can find that *Country* has a higher density in the high value area of affinity than other genres apparently; *Latin* and *Jazz* tend to be distributed in low value areas. There are many similar analysis, we will not repeat them. These visualizations can bring us intuitive understanding of node features.

Figure 2: Exploratory visual analysis of network node features.



3.2 Subnetwork Analysis

In this subsection we will give the visualization and description of directed influencer network. Because there are too many nodes and edges in the whole network, so we will create a subnetwork for visualization and analysis. We only select the top 100 static influence artists to join the sub graph nodes as the subnetwork nodes and the influence-follow action among them as edges. For one artist is not connected to others, we finally build a directed influencer network with 99 nodes and 377 edges.

We first show the top 10 static influence artists below so as to get an intuitive feeling. We list their information and node features in the Table 2.

Next, we use *Gephi*, which is a complex network analysis platform, to visualize the directed influencer subnetwork. The results are shown in Figure 3 below. We use Fruchterman Reingold algorithm proposed in this work [6] to layout the network.

In this network, node represents artist, directed edge represents influence-follow action, the thickness (or weight) of edge indicates music influence of influencer on follower and the size of node indicates the total music influence of artist in the network. Important statistical descriptions of the network are as follows: the average degree is 3.808, the average weighted degree is 11.836, the diameter of network is 8, the average path length is 2.721, the density of network is 0.039 and the average clustering coefficient is 0.156.

We also do community detection in the network and the modularity algorithm comes from this work [2]. The results show that the number of communities is possibly 5 and the modularity of network is 0.454.

Table 2: Top 10 static influence artists.

Name	Decade	Genre	Followers	Strength	Width	Affinity	Static Influence
The Beatles	1960	Pop/Rock	615	164.9	0.4454	0.5048	37.07
Bob Dylan	1960	Pop/Rock	389	131.1	0.4803	0.5068	22.57
Marvin Gaye	1950	R&B	169	80.31	0.4518	0.5086	18.45
Kraftwerk	1970	Electronic	108	67.25	0.5060	0.5140	17.49
Miles Davis	1940	Jazz	160	68.79	0.4746	0.5072	16.56
Muddy Waters	1940	Blues	113	57.81	0.4816	0.5137	14.30
James Brown	1950	R&B	154	76.67	0.5168	0.5093	14.27
Hank Williams	1930	Country	184	67.97	0.3930	0.5121	13.68
The Rolling Stones	1960	Pop/Rock	319	118.7	0.3855	0.5077	13.42
Billie Holiday	1930	Vocal	106	51.59	0.4735	0.5115	12.49

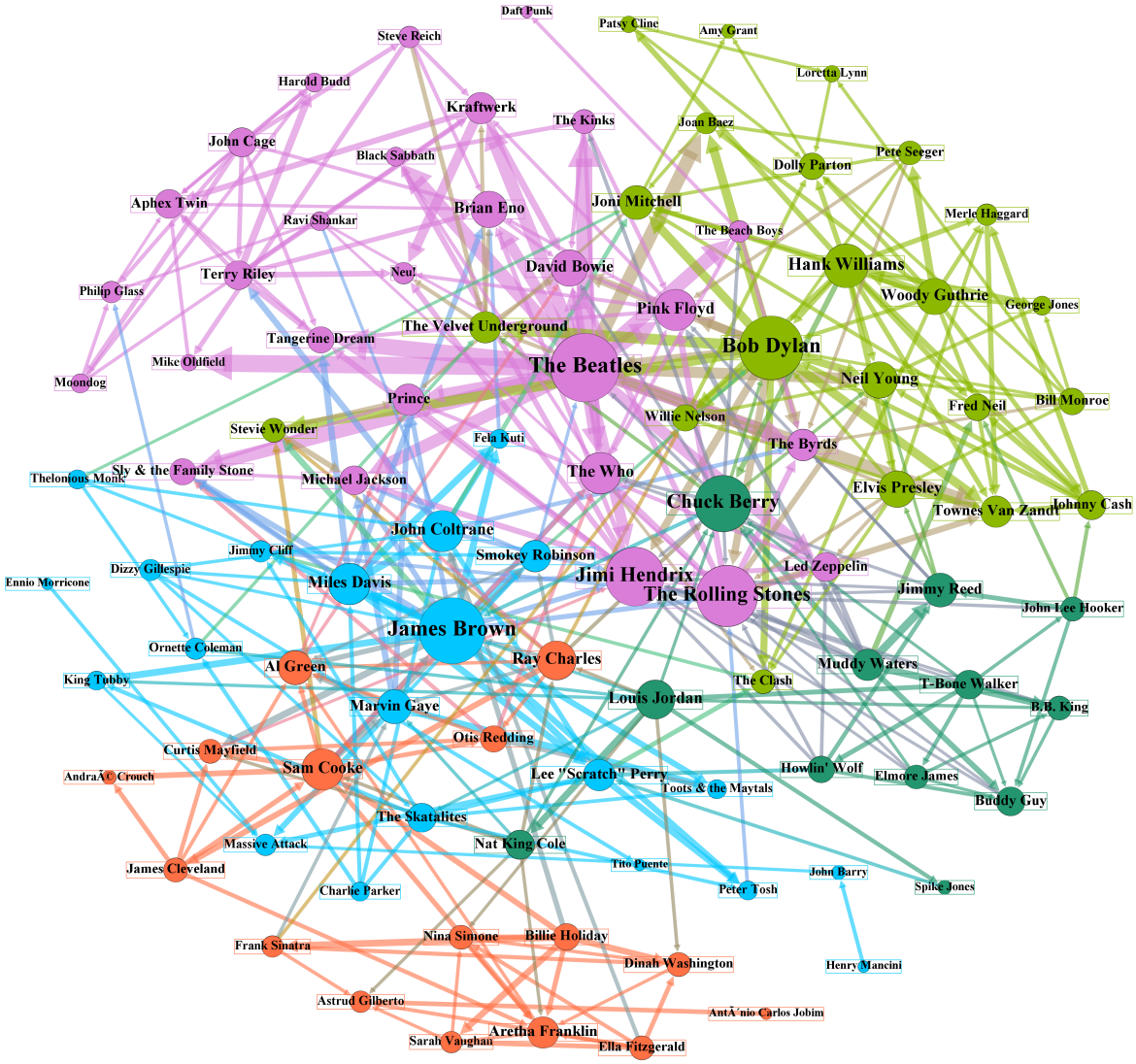
In Figure 3, the node set of each color represents one community. By observing the influential artists in each community, we analyze them as follows.

- The pink community mainly contains artists from *Pop/Rock* and *Electronic*, such as The Beatles (*Pop/Rock*), Kraftwerk (*Electronic*), The Rolling Stones (*Pop/Rock*) and Jimi Hendrix (*Pop/Rock*).
- The blue community mainly contains artists from *R&B*, *Jazz* and *Reggae*, such as Marvin Gaye (*R&B*), Miles Davis (*Jazz*), James Brown (*R&B*), Lee "Scratch" Perry (*Reggae*), John Coltrane (*Jazz*), Peter Tosh (*Reggae*) and Smokey Robinson(*R&B*).
- The light green community mainly contains artists from *Country*, *Folk* and *Pop/Rock*, such as Hank Williams (*Country*), Johnny Cash (*Country*), Merle Haggard (*Country*), Woody Guthrie (*Folk*) and Bob Dylan (*Pop/Rock*).
- The dark green community mainly contains artists from *Blues*, *Reggae* and *Pop/Rock*, such as Muddy Waters (*Blues*), Jimmy Cliff (*Reggae*), Jimmy Reed (*Blues*), T-Bone Walker (*Blues*) and Chuck Berry (*Pop/Rock*).
- The orange community mainly contains artists from *Vocal*, *Latin* and *R&B*, such as Billie Holiday (*Vocal*), Sam Cooke (*R&B*), Antonio Carlos Jobim (*Latin*), Ella Fitzgerald (*Vocal*), Ray Charles (*R&B*) and Nina Simone (*Vocal*).

We can preliminarily summarize the following inferences.

- *Pop/Rock* has three main branches and corresponding representative artists are The Beatles, Bob Dylan and Chuck Berry. The music style of the branch represented by The Beatles is related to that of *Electronic*, the branch represented by Bob Dylan is related to *Country* and *Folk* and the branch represented by Chuck Berry is related to *Blues*.
- *R&B* has two main branches. The music style of the branch represented by Marvin Gaye, James Brown and Smokey Robinson, which is the main stream, is related to that of *Jazz* and *Reggae*. Differently, the branch represented by Sam Cooke and Ray Charles is related to *Vocal* and *Latin*.

Figure 3: The visualization of directed influencer subnetwork.



3.3 Influence Maximization Algorithm

In the research of social influence analysis, how to get the seed set effectively is an important problem [8]. We not only want to know the size of influence between two nodes, but also want to get influence spread mechanism in the network. Under the assumptions in subsection 2.3, the whole network is generated by a seed set through influence spread. The artist in seed set can be regarded as revolutionaries, who played an important role in evolution and led the genre to develop in a new direction. So here we will give an algorithm based on greedy selection to obtain the seed set.

We use $\sigma(S)$ to represent the influence spread function, which equals to the total music influence given to current seed set S for influencing others in the networks, i.e. $\sum_{u\in V-S}\frac{1}{|\mathcal{A}_u^I|}\Gamma_{S,u}$. The marginal gain of node v is given by

$$\sigma(S+v)-\sigma(S)=\sum_{u\in V-S\cup\{v\}}\frac{1}{|\mathcal{A}_u^I|}\Gamma_{S\cup\{v\},u}-\sum_{u\in V-S}\frac{1}{|\mathcal{A}_u^I|}\Gamma_{S,u}=\sum_{u\in V-S\cup\{v\}}\frac{1}{|\mathcal{A}_u^I|}\Gamma_{v,u}-\frac{1}{|\mathcal{A}_v^I|}\Gamma_{S,v} \tag{8}$$

Algorithm 1 Music Revolutionary Algorithm based on Greedy Selection

Input: network $G = (V, E)$, the size of set seed k, discount factor μ
Output: seed set S

set $S = \varnothing$, $Q = \varnothing$
for each $u \in V$ **do**
    generate $\mathcal{A}_u^I, \mathcal{A}_u^F, N_u, t_u, KL_u^{\text{year}}, KL_u^{\text{genre}}, R_u$
for each $u \in V$ **do**
    compute influence strength Strength_u by Equation 1
    compute influence width Width_u by Equation 2
    compute social affinity Affinity_u by Equation 3
    compute static influence α_u by Equation 4
for each $(v, u) \in E$ **do**
    compute direct influence $\beta_{v,u}$ by Equation 5
    compute music influence $\Gamma_{v,u}$ by Equation 6
for $i = 1$ to k **do**
    for $v \in V - S$ **do**
        compute marginal gain $\sigma(S + v) - \sigma(S)$ of v by Equation 8
        push v into queue Q in decreasing order
    get queue head w=pop(Q)
    update $S = S \cup \{w\}$
return S

To obtain the seed set, we use influence maximization strategy. To deal with this problem, many efficient heuristic algorithms have been proposed [7] [10], while they always sacrifice the accuracy to guarantee the efficiency of executions. In our work, we present the Music Revolutionary Algorithm based on Greedy Selection. This algorithm refers to the Greedy Algorithm with Node Features proposed in this work [9], but we improve it. Algorithm 1 shows our full algorithm of selecting k seed nodes.

In our problem, the number of nodes and directed edges of the network is less than that of the online social network, so the running time of Algorithm 1 is not long. But when the complexity of the problem increases, such as adding more influencers and followers, we should consider using heuristic algorithm instead of greedy method. In addition, we can also use dynamic influence $\gamma_{u,S}^T$, as defined in subsection 6.2, instead of direct influence $\beta_{u,S}^T$ to describe the influence of revolutionaries more reasonably. In this way, we run the algorithm to get music revolutionaries in Section 6. Our algorithm provides a unified framework based on greedy selection. By setting different node features and influence spread mechanism, it can be applied to a variety of scenarios.

4 Music Similarity

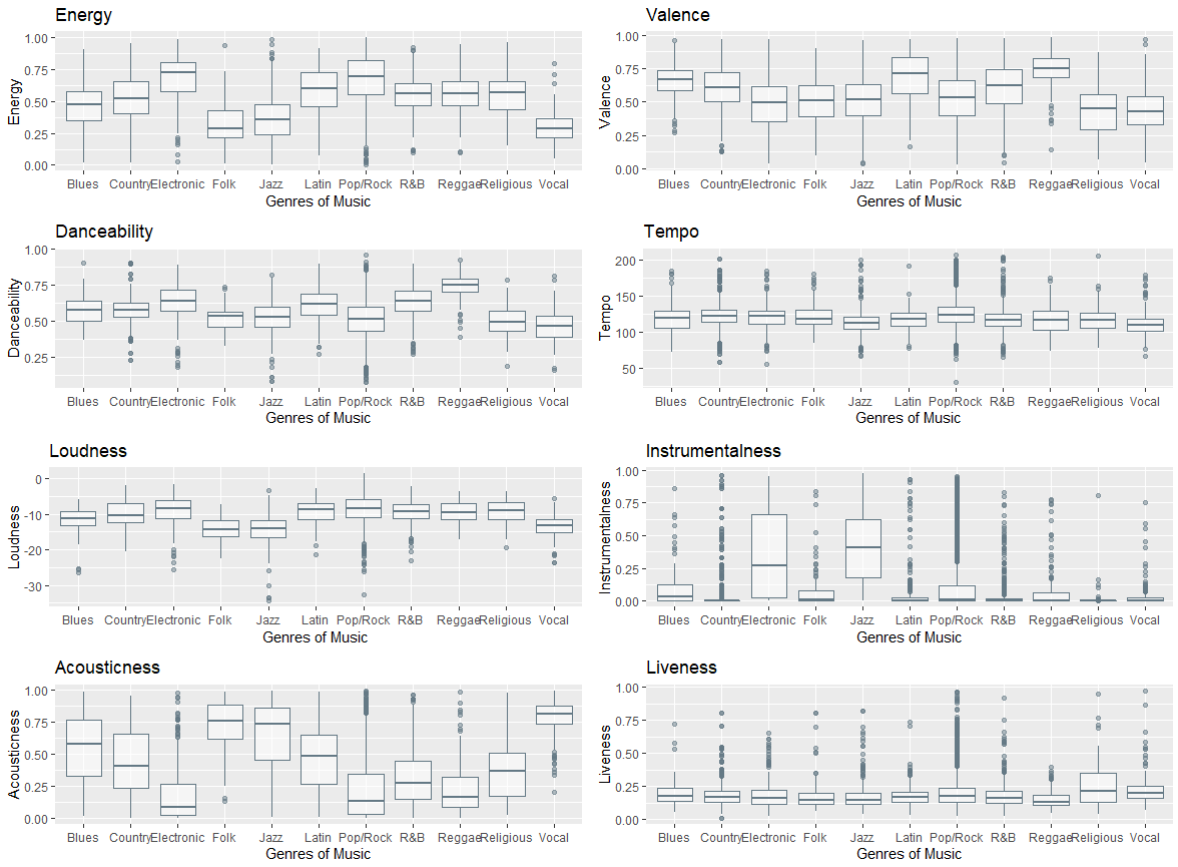
In this section, we use *data_by_artist* and *full_music_data* to develop a measure of music similarity. With this measure, a comparison between the similarity of artists within and between genres will be conducted. This helps reveal the similarities and influences between and within genres. Besides, analysis on genres' outstanding features and their change over time will be included. From the above discussions, we can discover some latent relationships between different genres. This section answers Task 2 and Task 3 in the whole problem.

4.1 Exploratory Data Analysis

Before we decide what measures to take, let's first take a quick look at our data.

In *data_by_artist*, each music has 13 characteristics. Based on the physical meanings of these variables and some common sense, we know that *duration_ms*, *popularity* and *mode* are variables that describe some unchangeable facts about music and are not suitable to distinguish a music. Besides, we can draw the boxplot of each variable for music of different genres, as Figure 4 shows. This plot can give us a more intuitive knowledge about these variables. A direct finding is that, some variables have value between 0 and 1, while some others have much larger values: the mathematical units are not uniform. This means when considering how to measure similarity between music, data need to be standardized at the beginning. We can also find that, there is a big difference between music of different genres in some of these variables, such as *energy*, *valence*, *danceability*, *instrumentalness* and *acousticness*. This reminds us that, these variables may play an important role in distinguishing different music.

Figure 4: Data description.



4.2 Measures of Music Similarity

Now we consider what measures to take to describe the similarity between music. Recall that we have 13 variables in total, but three of them are not suitable to use as measure of different music. We weed these variables out and use the remaining ten to form a feature vector for each music. From the previous subsection, we see that units of variables are not uniform. So we cannot simply use Euclidean distance, which may be confounded by variables that have large values. With our key assumption that, our data are samples from an infinite population and our sample size is large enough, we suppose that sample means and sample variances of these variables are the same as the original infinite population. Under this assumption, we can standardize our data with the following formula:

$$X^{SC} = \frac{X - \mathbf{mean}(X)}{\mathbf{std}(X)}, \tag{9}$$

where X is our data matrix, whose rows represent different songs or artists, columns mean 10 different variables. For nodes u and v, we define the Standardized Euclidean Distance between row vectors X_u^{SC} and X_v^{SC} as music distance, denoted as $d_{u,v}^{SC}$. Further, we give the direct measure of music similarity between node u and node v as

$$\rho_{v,u} = \frac{1}{1 + d_{u,v}^{SC}}, \tag{10}$$

$\rho_{u,S}$ can be similarly defined as the average value of $\rho_{u,v}$, $v \in S$. To compare music similarity, we can use both $d_{u,v}^{SC}$ or $\rho_{v,u}$ as the measure. In Section 4 and 5, we mainly use $d_{u,v}^{SC}$. In Section 6, we use $\rho_{v,u}$ for better expression. In fact, they have almost the same effect as the measure.

4.3 Similarity of Artists within and between Genres

With the above measure of similarity between music, we now conduct a comparison between artists within and between genres. Our interest is, are music of same genre closer to each other than music of different genres?

To answer this question, We only choose artists in *data_by_artist* data set whose genre is known to be our sample. To begin with, we use one most simple method to have a little try: we randomly choose two genres and use mean of the standardized Euclidean distance of every two artists in the specific genre as measure of similarity within this genre. Similarly, we use mean of the standardized Euclidean distance of every two artists between two genres as measure of similarity between the two genres. If the distance between two genres is larger than that within one genre, we come to the conclusion that music of the same genre are closer to each other than music of different genres.

This method is explicit but qualitative, we need a quantitative way to support our conclusion. Recall that in the previous key assumption, we assume that each variable follows a normal distribution. Upon this, a t-test is helpful to test our result. First of all, for every two different genres, instead of directly comparing the mean distance, we use form the computed distance of every two artists in one specific genre as one sample X, and that of artists between two genres as another sample Y. We then conduct a hypothesis test:

- $\mathbf{H}_0$: X **is not smaller than** Y**.**

If p-value is significantly small, we will have strong confidence to reject the null hypothesis and come to the conclusion that music in the same genre do have smaller distance.

Here we conducted the test on 11 genres of music. In order to have a more explicit exhibition of the result, we made a min-max normalization to our result, which restrict the distance in interval [0,1]. In this way, we can plot it with a correlation-plot, as is shown in the left graph of Figure 5. In this figure, numbers in the squares means distance between two genres, and if the test result is not significant, the corresponding square will have a cross. The result shows that our measure of similarity can pass most tests. Generally, the distance between music of the same will be smaller than music between genres.

4.4 Comparison of Genres

In the previous two subsection, we defined a measure to distinguish different genres. Under that measure, the 11 genres we selected performed well. In this subsection, we are going to dive deeper and reveal the similarity and differences between music of different genres.

First of all, we conduct a cluster analysis with the above defined distance. In every single step, we compute the distance between different groups and combine the closest two. If a group has more than one element, the distance is defined to be the average of distance to all elements in the group. Repeat the above procedure until the iteration ends. Finally, we obtain five clusters, which are shown in Figure 6. Based on the cluster and the test of similarity between genres in the previous subsection, we can get more interesting information about different genres.

We can see from the left graph of Figure 5 that, the test of *Folk* and *Vocal* music is not significant. Additionally, in our cluster result, *Folk* and *vocal* are in the same group. This result shows that *Vocal* and *Folk* music are similar

Figure 5: Similarity and influence between genres.

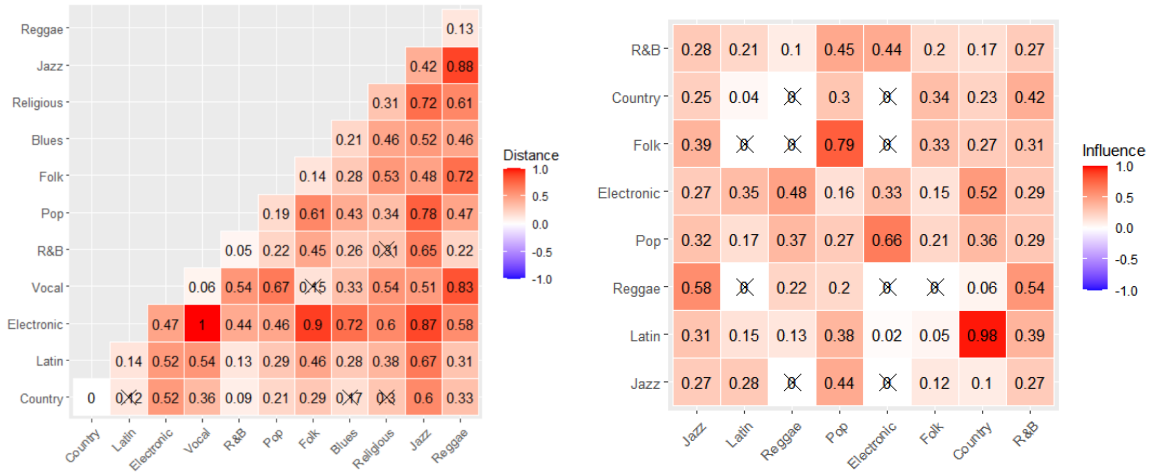
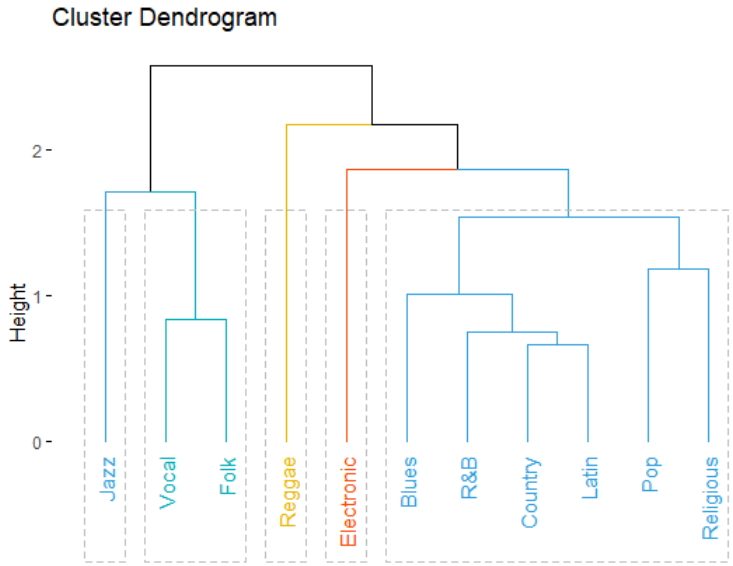


Figure 6: Hierarchical clustering results of genres.



in each feature. Similarly, *R&B* and *Religious*, *Country* and *Religious*, *Country* and *Blues*, *Country* and *Latin* also didn't pass the t-test, and all of them are in the same group, which means these genres of music are similar in our chosen 10 features.

We also noticed that, *Jazz*, *Reggae* and *Electronic* are clustered as single group. From the point of similarity test, these three genres of music passed all of their tests. Under our measure, these three genres must have some speciality over other genres. Let's first get back to our description of data in Figure 4 for an instinctive observation. It is explicit to see that, *Jazz* has the highest *instrumentalness*, which means *Jazz* music seldom using vocal. *Reggae* has the highest *danceability*, which is consistent with the fact that *Reggae* is a type of dance music. *Electronic* has the lowest *acousticness*, which is also consistent with the common sense that *Electronic* music has many electronic elements. The above analysis is qualitative and is based on our common sense.

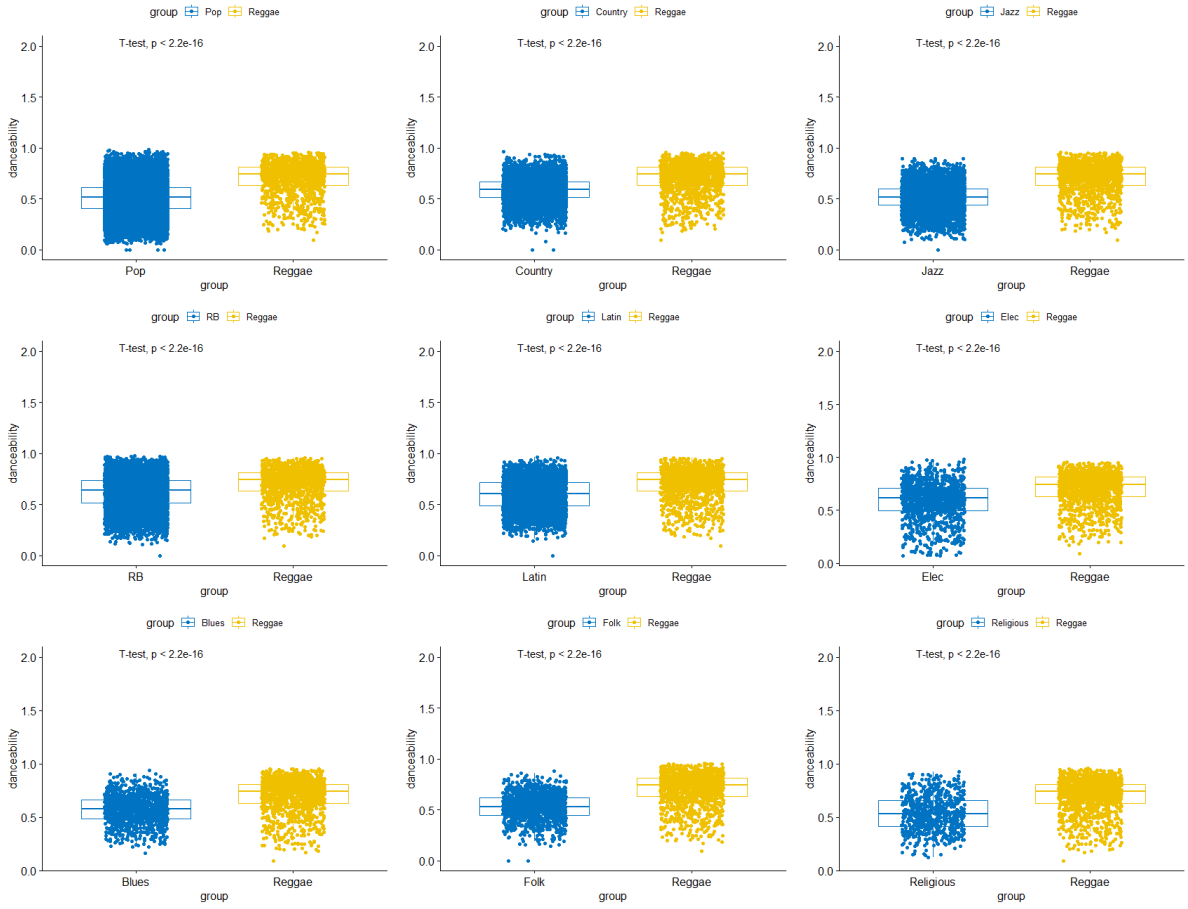
Now, we are going to test our observation. That is, we want to test quantitatively whether the three genres can be distinguished by the mentioned features. For simplicity, we take *Reggae* for example. This time, we use *full_music_data* data set for accuracy. In addition, since we have more data in *full_music_data* data set, the previous assumption of large sample normal distribution is still applicable. We then conducted t-test to test whether the

danceability of *Reggae* is significantly larger than that of other genres. Our null hypothesis is:

- H_0: **Danceability of *Reggae* is not larger than that of other genres.**

If p-value is significantly small, we can confidently reject the null hypothesis and come to the conclusion that *danceability* does distinguish *Reggae*. The result is consistent with our common sense, as is shown in Figure 7. We can repeat the procedure on the other two genres and the result will be the same.

Figure 7: Tests of large danceability of *Reggae*.



4.5 Change and Influence of Genres over Time

In this subsection, we plan to dig into change of genres over time in the perspective of similarity and influence between different genres.

First of all, we use *data_by_artist* to take a look at the change of the number of famous artists in specific music genre. This change is a reflect of rise and fall of the genre. A quick glimpse is shown in the first graph of Figure 8, which describes the change of numbers of famous artists in different genres of every ten-year period. Since our interest is in the similarity of the trend of different genres, a correlation test of every two genre's trend is conducted and the result is shown in the second graph of Figure 8. In the graph, genres that have correlation close to 1 are considered to have similar development trends, and that have correlation close to -1 are considered to have opposite trend. For example, the development of *Jazz* is similar to that of *Vocal* and *Blues*, and is opposite to that of *R&B*. This indicates that some genres of music have similar or opposite development trends, which means there may be some connection between these genres.

Additionally, we can also consider the development trend of other features. The method is exactly the same as the above. For simplicity, we only present six most important of these features. As Figure 9 shows, *Country* and

Figure 8: Change of famous artists.

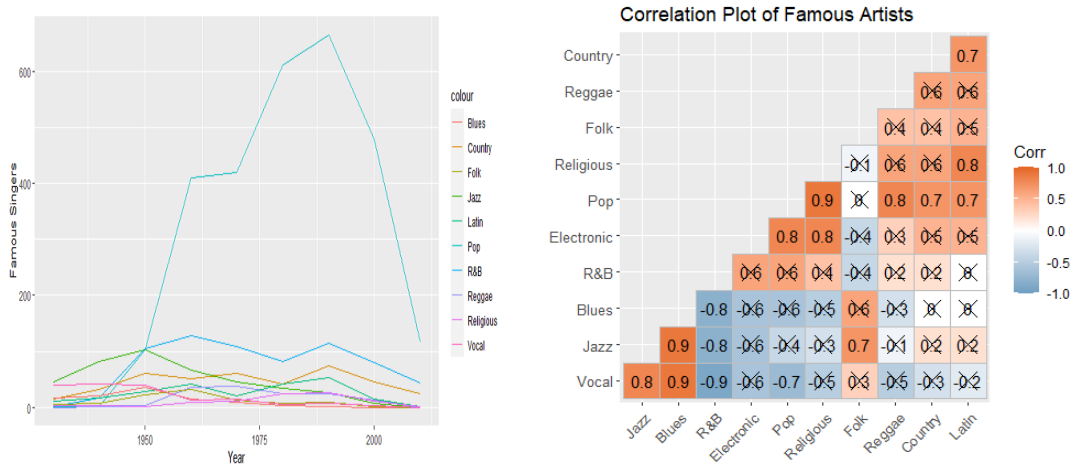
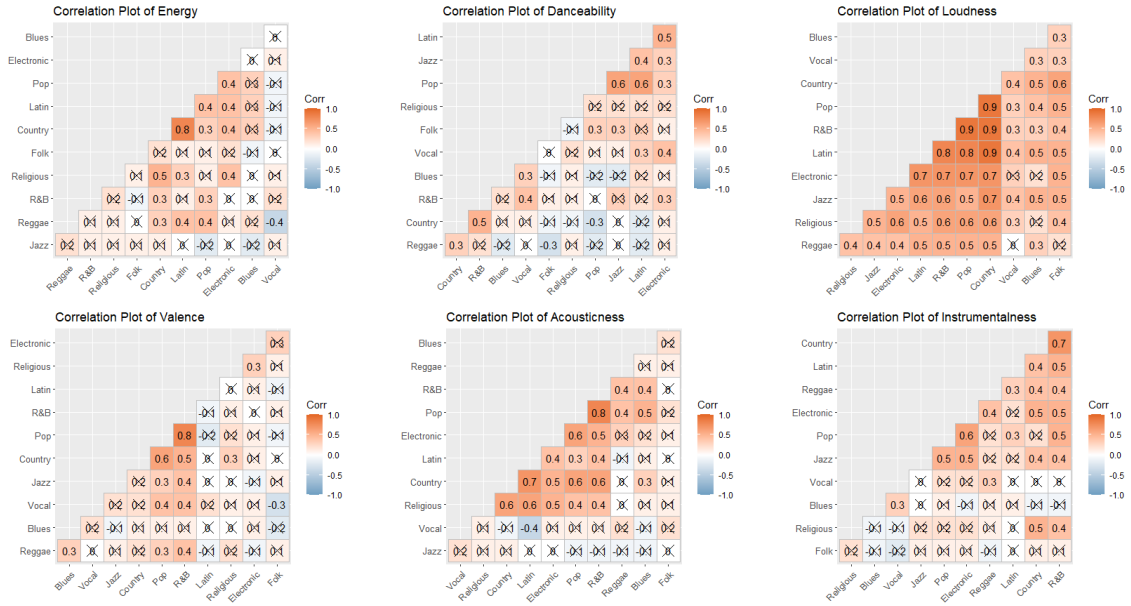


Figure 9: Correlation of changes.



Latin have similar trend of *energy* and *acousticness*, *Pop* and *R&B* have similar trend of *valence* and *acousticness*, *Country* and *R&B* have similar trend of *instrumentalness*, *Country* and *Pop/Rock* have similar trend of *loudness*, etc.

The above analysis reveals how the change of genres are connected from the perspective of similarity. We can also consider the problem from a different perspective: influence. That is, how genres influence each other? Recall that in Section 3, we defined and computed influence of different artists. We now use the previous result and compute the influence of artists in the same and different genres. Here, the influence of genre A to genre B is defined to be the average of influence of artist in genre A to artist in genre B. In order to have a more explicit display of the result, we use min-max standardization and restrict the influence to [0,1]. This can be visualized by the right graph in Figure 5. The graph represent the influence of genres on the horizontal axis to genres on the vertical axis. If there is no data to use, we mark it with 0 and put a cross on it. It is easy to see that, *Pop* has a large influence on *Folk* and *Electronic*, *R&B* has a large influence on *Reggae*, and *Country* has a large influence on *Latin*. This is consistent with the result in subsection 3.2.

5 Influence Behavior

In Section 3, we calculate static influence of the node, define measure of music influence in network and analyse the influence spread mechanism. In Section 4, we develop measure of music similarity and compare similarities and influences between and within genres. In this section, we try to explore whether influencers in fact influence the music style of their followers and whether some music characteristics are more contagious than others. This section answers Task 4 in the whole problem.

5.1 Effectiveness

To explore whether the identified influencers in fact influence the respective artists, our method is to verify whether the influence-follow action really significantly improve the music similarity between two artists, in other words, significantly reduce the distance between music characteristics. Two things need to be clear. First, when we compare whether influence-follow action can significantly reduce the distance, we can only consider the action between artists within the same genre. This is because in Section 4, we have found that there is a significant difference in music similarity between and within genres and most followers of the influencer come from the same genre. Second, the influencer has much individuality for influence-follow action. When we compare the influence of following or not on music distance, we need to choose the only one influencer.

Table 3: The coefficient estimates and corresponding p-value (in brackets) of four logistic regression. The last line of table represents the p-value obtained by one-sided t-test on the distance data. *** means $p \leq 0.001$, ** means $0.001 < p \leq 0.01$, * means $0.01 < p \leq 0.05$ and . means $0.05 < p \leq 0.1$

Name	The Beatles	Marvin Gaye	Miles Davis	Hank Williams
Genre	Pop/Rock	R&B	Jazz	Country
# Followers	553	99	83	97
# Non-Followers	2253	577	322	305
danceability	−0.328 (0.001)***	+0.162 (0.350)	−0.531 (0.036)*	−0.532 (0.027)*
energy	−0.336 (0.003)**	−0.602 (0.008)**	−0.961 (0.000)***	−0.075 (0.781)
valence	−0.125 (0.148)	+0.066 (0.754)	+0.084 (0.747)	−0.442 (0.079).
tempo	−0.126 (0.143)	−0.513 (0.023)*	−0.312 (0.173)	+0.143 (0.486)
loudness	−0.188 (0.076).	−0.111 (0.595)	−0.454 (0.101)	−1.059 (0.000)***
key	+0.071 (0.154)	+0.073 (0.515)	−0.124 (0.335)	−0.101 (0.661)
acousticness	−0.526 (0.000)***	−0.672 (0.005)**	−0.589 (0.054).	−0.002 (0.989)
instrumentalness	−0.131 (0.083).	−0.505 (0.072).	+0.026 (0.888)	−0.188 (0.266)
liveness	−0.094 (0.253)	−0.138 (0.518)	+0.044 (0.820)	+0.063 (0.733)
speechiness	−0.437 (0.000)***	−0.057 (0.724)	−0.334 (0.343)	+0.037 (0.796)
distance	<2e-16***	3.193e-05***	0.0701.	5.420e-08***

In order to make the results more convincing, we choose the artists with the largest number of followers successively from the four most popular genres. More precisely, they are The Beatles from *Pop/Rock*, Marvin Gaye from *R&B*, Miles Davis from *Jazz* and Hank Williams from *Country*. In this way, we get 4 data sets. The main variables are the absolute difference of music characteristics between artists and the selected influencer above. Before making the difference, we first standardize each characteristics within genre. In addition, we use all music

characteristics to calculate Euclidean distance between artists and influencer, similarly to Section 4. Then, we give each artist an *action* feature, When the artist is influenced by the selected influencer, *action* takes 1, otherwise takes 0. We use the above 4 data sets to run 4 logistic regression, using *action* feature as the classification dependent variable. In order to avoid multicollinearity problem, we do not add "distance" feature in logistic regression model, but perform one-sided t-test separately on it. More specifically, H_0 : $\mathbf{distance}_0 = \mathbf{distance}_1$ and H_1 : $\mathbf{distance}_0 > \mathbf{distance}_1$ are the null and alternative hypothesis. All the results are listed in Table 3.

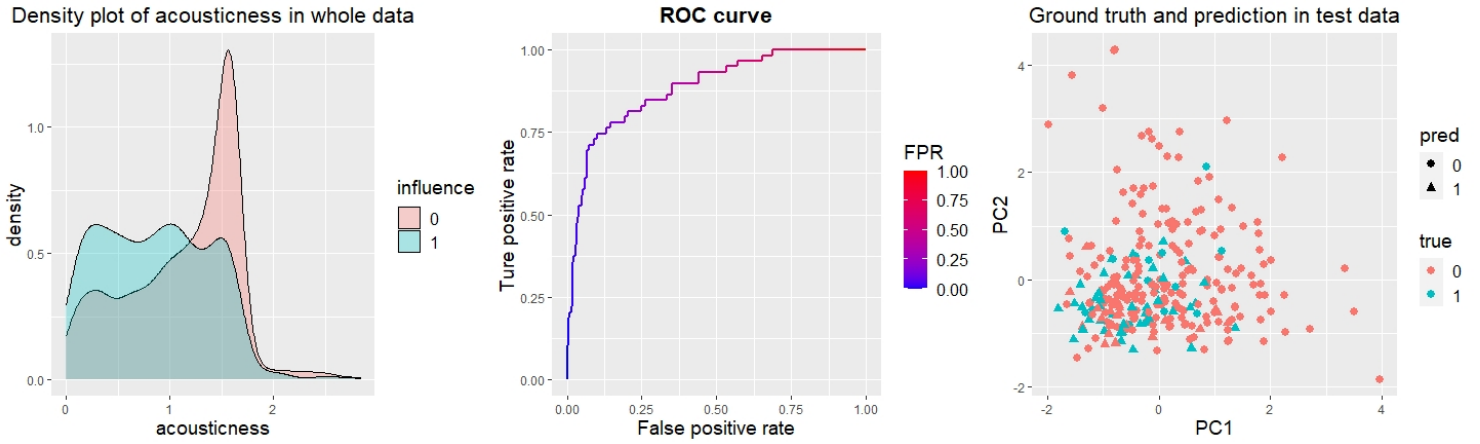
According to Table 3, most of the music characteristics are negatively correlated with *action*, which means that followers and their influencer tend to have a smaller characteristic difference, in other words, higher similarity. The p-value of some characteristics is significant and all this characteristics have negative estimates. In addition, different influencers have different influence behaviors on their followers. For Marvin Gaye, *energy*, *tempo* and *acousticness* have more important roles in influencing followers, while *loudness* and *danceability* are more contagious for Hank Williams. Both *danceability* and *energy* play an important role for three influencers. Finally, t-test is significant for The Beatles, Marvin Gaye and Hank Williams and weakly significant for Miles Davis. This shows the influencers actually affect the music created by the followers.

5.2 Characteristic Importance

The logistic regression performed in the previous subsection can preliminarily get the conclusion of feature importance. However, logistic regression is a linear model, which can't measure the effect of feature combination. Therefore, we need to use more complex nonlinear classification model to measure characteristic importance. Here we choose XGBoost model proposed in [4], which is a very popular and high performance machine learning algorithm. As machine learning algorithm often needs a large amount of data, we only show the results of The Beatles' data set. The sample size of the data set is 2806. We split 90% as training set and 10% as test set. Since the positive samples are far less than the negative samples, we enhance the positive samples. We repeat the positive samples in the training set four times to make the model learn more features of positive samples.

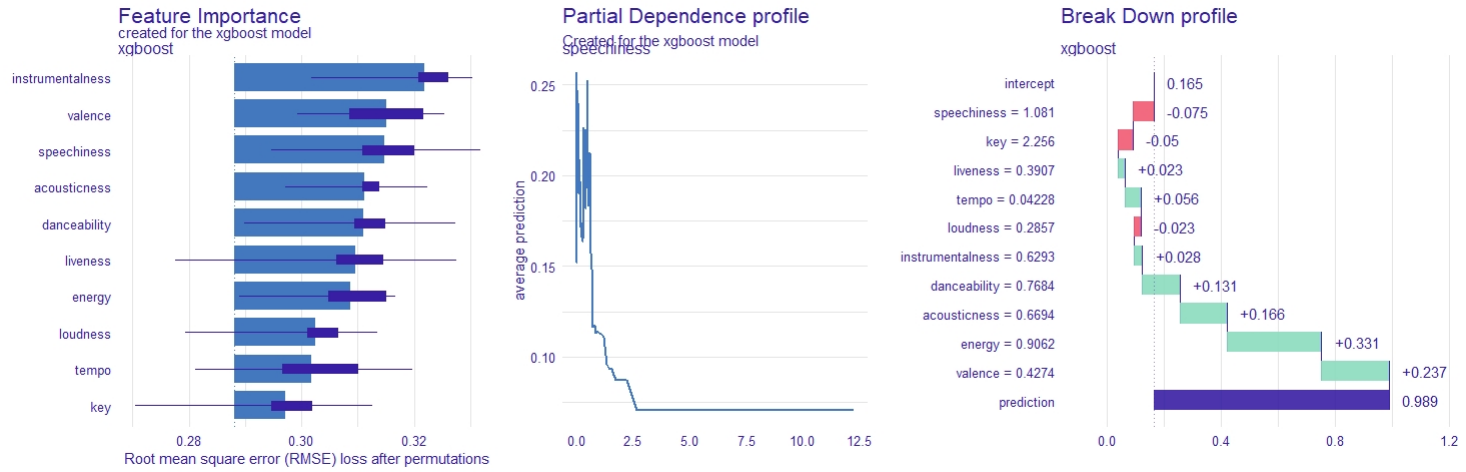
When training, we set the number of subtrees as 5000, the maximum depth of the tree as 4 and the learning rate as 0.05. Then we test the model and get that accuracy is 86.07%, true positive rate (or recall) is 74.58%, true negative rate is 89.14%, precision is 64.71%, F-score is 0.4826 and AUC is 0.8413. Overall, the classification effect on the test set is well. Some exploratory data analysis, ROC curve and comparison between ground truth and prediction are shown in Figure 10.

Figure 10: Exploratory data analysis on "asousticness", ROC curve of XGBoost model and comparison between ground truth and prediction. "FPR" means false positive rate.



Next, we give some detailed model analysis results in Figure 11. The figure on the left shows the relative importance of variables and we can find that *instrumentalness*, *valence*, *speechiness* and *acousticness* are the most important four features after considering the combination of variables, which is not exactly the same as logistic regression. The figure in the middle is partial dependence profile, which shows how does prediction value change with *speechiness* feature when other features are taken as fixed average values. The figure on the right shows how each feature works on prediction value of Jimi Hendrix, who is also a famous artist in *Pop/Rock*. There are many similar specific analyses, which we will not repeat here. In this way, we improve the interpretability of machine learning model and open the "black box" to some extent.

Figure 11: Some detailed model analysis results.



6 Musical Evolution

In this section, we use *full_music* and *data_by_year* data sets to reveal the process of musical evolution both in general and separately in one genre, which answers Task 5 and Task 6. We also build a relationship with music and the surrounding world by discussing the mutual influence effects to solve Task 7.

6.1 Identification of Revolution

Revolution in music, defined as great change or major leaps, was usually caused by emegence of a new genre or a shift in an old one. In this subsection, we try to signify revolution by analysing these data. And a revolution can reflect in two aspects: the characterists of songs, which measures songs from a professional perspective, and the popularity, which measures from social perspective.

- **Musical Characterists**. Data sets include 13 kinds of musical characteristics, of which seven focus on the background music in a song, five on the vocal performance and one shows the length of a song. We use the data in *data_by_year* to show the change of characters in general, as it's shown in Figure 12.

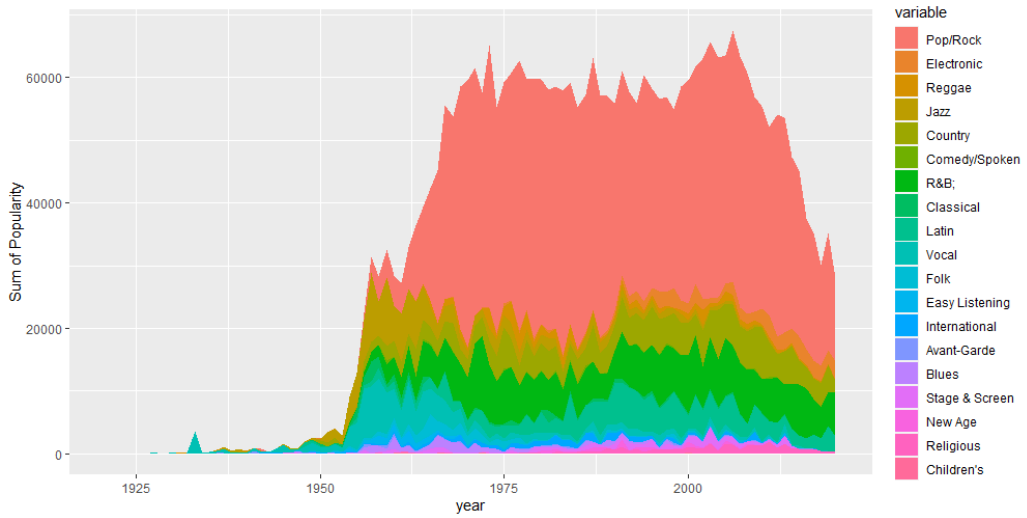
  In this figure, this data shows the average of songs'characteristics in one year. Some data has already been 0-1 scaled, which doesn't influence the trends of changing. *Mode* and *explicit* characteristics aren't inclued since they showed little change over 100 years.

  From Figure 12, we can directly see some major leaps without further statistical anlysis. For example, notice the characteristic *acousticness* in ower left subfigure, we can see a sharp drop in about 1950. Then we may guess whether there was a development in the vocal processing technology, such as technology enhancements or electrical amplification. And there will be more detailed anlysis with statistical tools later in this subsection.

Figure 12: The change of musical characteristics over years.

- **Popularity**. For this characterist, we choose the sum of popularity rather than the average since the sum reflects both the number of songs and the average popularity. And in order to show the evolution of the whole music industry as well as each genre, we plot a stowage diagram of popularity,as Figure 13 shows.

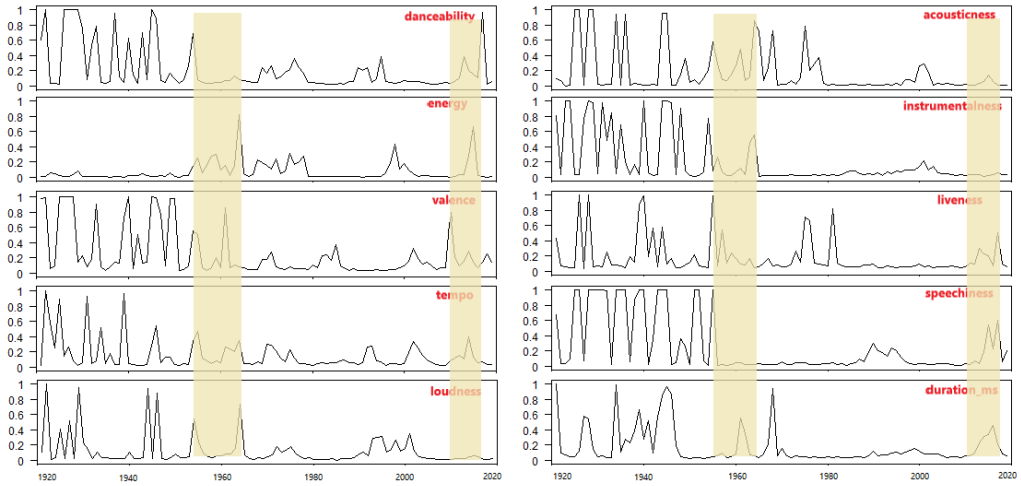
Figure 13: The change of popularity over years.



From Figure 13, we can find that the whole music industry enjoyed a rapid boom in around 1950s with the rise of *Jazz* music. And the second wave occurred during 1970s due to the upsurge of *Pop/Rock* music, which dominates the music industry for about four decades. As for each genre, it's obvious that the *Jazz* genre began to decline at about 1960. And the *Vocal* genre faded away in 1960s while the *Latin* genre sprang up around the same time.

All above analysis is just what we can directly get without rigorous calculations.To identify major leaps in a statistical way, we use the method called "Change Points Detecting". After mutiple times of trail and comparison, we use *bcp* package [3] in R to detect change points of several characteristics, the result of which is showed in Figure 14.

Figure 14: The possible change points by *bcp*.



For each characteristic, the curve shows the possibility of being change point for every year. And the accurate time for each possible change point can also be calculated through R. We won't discuss the data before 1950 because the number of songs is not big enough to provide stable data. To be precise, the curve may have a tiny shift after removing unstable data, but we just show the results of final calculation in Table 4.

Table 4: Accurate years of possible change points. x, such as 1954 for Dan., means the possibility to be change point > 0.7; while (x), such as (1954) for En., means 0.6 < posibiliy ≤ 0.7.

Char.	**years**	**Char.**	**years**
Dan.	1954,2017	Acou.	1961,1964,1965,(1968),1975
En.	(1954),(1959),1964,2015	In.	1954,(1960),(1961),1964
Val.	(1955),1961,2010,2014,(2018)	Li.	1955,(1957),1975,1976,1981,2017
Tem.	(1955),(2002),2014	Sp.	1955,2015,2017
Loud.	1954,1961,1964,(1998),(2001),(2015)	Du.	(1962),(1967),1968

According to change point detection, we can give the final conclusion:there may were two revolution over the past seven dacades, the first one during 1954 to 1964 while the second from 2010 to 2018.

- **the First Revolution**. During the first major leap, almost every characteristic we used went throught big changes. According to Figure 13, it was just the right period when *Jazz* enjoyed a short-time glory while *Pop/Rock* music began to take the dominant role. It's also the time when vocal music began to give way to *Latin* music. Together with the special features of these musical genre, it's easy to explain changes in the characteristics. For example, *danceability* and *energy* arose rapidly as a result of the raise the *Pop/Rock*, which is more vibrant and suitable to dance than *Jazz*.

- **the Second Revolution**. The second major leap focused more on musical change rather than *Vocal*. Figure 13 showed us that this period witnessed a drop in *Pop/Rock*, or even in the whole musical industry. Meanwhile,

Latin, *Electronic* and some new genre came out and became increasingly popular. With the wide use of network and lower pride to make songs, the younths take the stage to tell new generation's story, which is more nagitive and faster than before.

6.2 Revolutionary Model

In this subsection, we will give a effective method to find the revolutionaries–in other words, influencers of major change– by combinding the influence network in Section 3 and the similarity model in Section 4. We build a dynamic influence model for revolution detect in Task 5 and influencers analysis in Task 6. We define the dynamic influence an artist had on a gernre or the whole music as below:

$$\gamma_{u,S}^{T} = c \cdot \alpha_u + \sum \rho_{u_t, S_{t+1}} P_u^t \tag{11}$$

In this equation, c is a constant number to balance the static influence and the similarity. In our section, we choose $c = 1$ when T is one decade. We consider the similarity between u_t and S_{t+1} since we consider there to be a delay before the artist' music influenced the public or the musical industry.

Using dynamic influence, we run Algorithm 1 propose in subsection 3.3 to find the revolutionaries in the two major leaps. We separately calculate the dynamic influence of each artist on the whole musical genre in certain decades, such as 1950s. The results are shown in Table 5.

Table 5: Top 6 influence artists in the First Revolution.

1950s				**1960s**			
Name	**Year**	**Genre**	**D. Inf**	**Name**	**Year**	**Genre**	**D. Inf**
Miles Davis	1940	Jazz	85.09	The Beatles	1960	Pop/Rock	248.37
Elvis Presley	1950	Pop/Rock	80.66	Elvis Presley	1950	Pop/Rock	138.46
Dean Martin	1940	Vocal	64.75	Bob Dylan	1960	Pop/Rock	128.17
Johnny Cash	1950	Country	58.8	The Beach Boys	1960	Pop/Rock	120.581
Chet Baker	1950	Jazz	50.08	The Rolling Stones	1960	Pop/Rock	110.12
Miles Davis Quintet	1950	Jazz	49.61	Marvin Gaye	1950	R&B	109.39

Table 6: Top 6 influence artists in the Second Revolution.

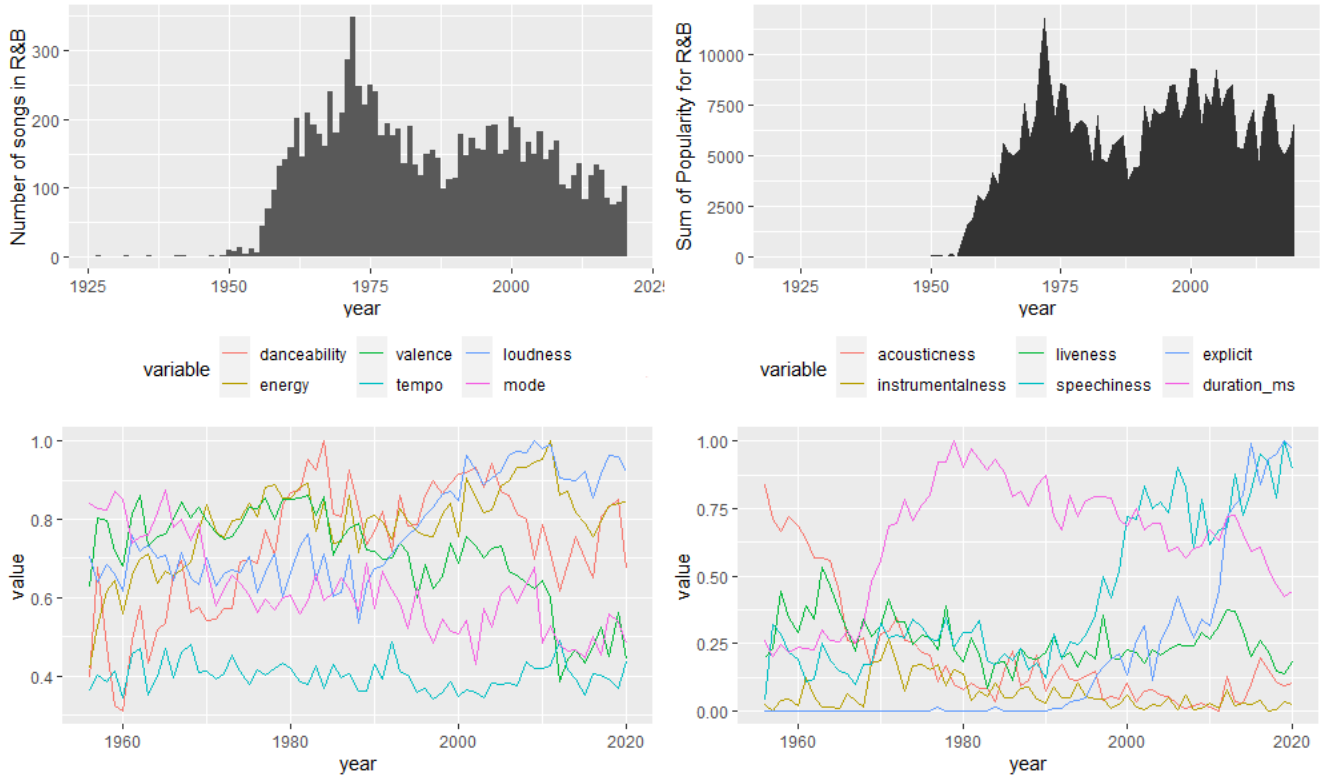
2000s				**2010s**			
Name	**Year**	**Genre**	**D. Inf**	**Name**	**Year**	**Genre**	**D. Inf**
Toby Keith	1990	Country	172.20	Calvin Harris	1990	Electronic	183.23
Brad Paisley	1990	Country	169.14	Zedd	2010	Electronic	181.02
Kenny Chesney	1990	Country	160.42	John Mayer	1990	Pop/Rock	157.99
Alicia Keys	1990	R&B	157.87	Bruno Mars	2000	Pop/Rock	156.95
Rihanna	2000	Pop/Rock	155.76	Blake Shelton	1990	Country	151.74
Rascal Flatts	1990	Country	149.92	Coldplay	1990	Pop/Rock	150.50

From these four tables, we can see not only the major revolutionaries in each major change, but also the change of the kinds of revolutionaries. For example, in 1950s, *Jazz* musicians were still the top influential group while they they has already been replaced by *Pop/Rock* in 1960s, which is consistent with Figure 13.

6.3 Evolution in One Genre: R&B

In this subsection, we choose the *R&B* genre (wrote as "R&B;" in data sets) to reveal the process of evolution over the past one hundred years. Figure 15 shows these basic characteristics of *R&B* genre.

Figure 15: Basic information about *R&B* genre.



Just as what we have done before, we use *bcp* to detect change points in this process. This time, we choose periods after 1956 in order to avoid unstable data, and the calculation results are showed in Table 7.

From the data above, we can have some direct speculation, such as: *R&B* change more on vocal aspect rath than musical, and there was a huge leaps for *R&B* at the beginning of the 1980s. Actually, as is told in musical history, a new kind of *R&B* music – Contemporary *R&B* came up at around 1980s. We still use the dynamic influence model mentioned above to reveal the dynamic influencers, the result of which is showed in Table 8.

From the table we can find an interesting phenomenon that the top influencers' list showed little repetition between decades, which implies that *R&B* genre developed at a relatively rapid speed. Also, we can realize that some names in 1980s to 2010s are the representative of the contemporary *R&B*.

Table 7: Accurate years of possible change points. x, such as 1968 for En., means the possibility to be change point > 0.7, while (x), such as (1957) for En., means 0.6 < posibiliy ≤ 0.7 and {x}, such as {1961} for Acou., means 0.5 < posibiliy ≤ 0.6

Char.	**years**	**Char.**	**years**
Dan.		Acou.	{1961}1965,1976
En.	(1957),1968,(1982),{2000},2011	In.	1968,{1972},1980
Val.	{1956},(1984),2011	Li.	1957,{1962},{1964}
Tem.		Sp.	1955,2015,2017
Loud.	(2000)	Ex.	1995,{2000},(2003),(2011),2014
Mode.	1969,(2004),2009	Du.	1968,1970,1976,1985,2016

Table 8: Top 5 influence artists on *R&B* in different decades.

Year	**Names of 5 Top Influencers**				
1950s	Ray Charles	The Platters	Jackie Wilson	Fats Domino	The Drifters
1960s	Marvin Gaye	Otis Redding	Aretha Franklin	Stevie Wonder	The Temptations
1970s	Earth, Wind & Fire	The Isley Brothers	Bill Withers	Barry White	Diana Ross
1980s	The Pointer Sisters	Luther Vandross	Tina Turner	James Brown	Rick James
1990s	Mariah Carey	Mary J. Blige	Nate Dogg	Boyz II Men	Michael Jackson
2000s	Alicia Keys	Mariah Carey	Usher	Beyoncé	Pharrell Williams
2010s	Chris Brown	The Weeknd	Beyoncé	Trey Songz	Ty Dolla Sign

6.4 Mutual Effects between Music and Surrounding Society

From the analysis above, we have already found some connections between music and other fields. For example, Figure 12 showed a sharp drop in *acousticness* in about 1950. Actually, 1950 was just the right year when Franklin S. Cooper Professor with his colleague John M. Borst finished the first Pattern Playback in the world, which laid the foundation for further track technology[1].

Alternatively, music plays an irreplaceable role in shaping the culture and society around us. Lets still take the *R&B* genre as example. *R&B*, as a typical Black music, wasnt given this name until 1947. Previously this kind of music was called as "Race Music", which was thought as insulting after the Civil War. Seen as a fruit of Black culture, *R&B* music in turn played an significant role in promoting Black culture. Many blacks lived on playing *R&B* in bars or markets, a few of which even become popular through this way, from [11].

Except *R&B* genre, other musical kinds also participate in shaping contemporary Black culture. As an article discussed, rap music, a very new kind music, constitutes a resistive occupation, employed by marginalized Black American youth to communicate thoughts and concerns that are often discounted by the dominant culture, and in doing so makes a significant contribution to Black American identities and culture, from [1].

[1]Refer to https://www2.ling.su.se/staff/hartmut/kemplne.htm

7 Conclusions

In this work, we build a series of models based on social network analysis and statistics to measure the influence-follow action between artists, genre music similarity and music evolution process.

First, we build a music influence network with node features. We define static influence of each artist and the music influence between influencer and follower to construct the directed network. We also propose a unified scheme based on greedy algorithm to find the seed set in influence network. Second, we put forward the measures of music similarity and using t-test and hierarchical clustering to compare similarities and influences between and within genres. Third, we explore the relationship between music similarity and influence behavior in the same genre by logistic regression and XGBoost model. We verify that influencers actually affect the music created by the followers significantly and some characteristics are more "contagious" than others. Last, we use the change points detection to signify revolutions. We also combine influence network together with similarity measure to build a dynamic influence model, which helps to identify revolutionaries in each revolution. In addition, by analyzing musical characteristics, we connect the music industry with the surrounding society.

For a more detailed summary, see the Summary Sheet on the front page.

References

[1] Elizabeth Pyatak Otr/Lsupa/Supsup*/Sup &Amp and Linda Muccitelli Maotr/Lsupa/Sup. Rap music as resistive occupation: Constructions of black american identity and culture for performers and their audiences. *Journal of Occupational Science*, 18(1):48–61, 2011.

[2] Vincent D. Blondel, Jean Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics Theory and Experiment*, 2008.

[3] Erdman Chandra and John W. Emerson. bcp: A package for performing a bayesian analysis of change point problems. *Journal of Statal Software*, 23(3), 2007.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *the 22nd ACM SIGKDD International Conference*, 2016.

[5] Xiaoheng Deng, Yan Pan, Hailan Shen, Jingsong Gui, Zheng Xiao, and Kenli Li. Credit distribution for influence maximization in online social networks with node features1. *Journal of Intelligent and Fuzzy Systems*, 31(2):979–990, 2016.

[6] Thomas M. J. Fruchterman and M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21, 1991.

[7] D. Kempe. Maximizing the spread of influence through a social network. *Proc.of Acm Sigkdd Intl Conf.on Knowledge Discovery & Data Mining*, 2003.

[8] Kan Li, Lin Zhang, and Heyan Huang. Social influence analysis: Models, methods, and evaluation. *Engineering*, pages 40–46, 2018.

[9] Guanfeng Liu, Feng Zhu, Kai Zheng, An Liu, Zhixu Li, Lei Zhao, and Xiaofang Zhou. Tosi: A trust-oriented social influence evaluation method in contextual social networks. *Neurocomputing*, 210(OCT.19):130–140, 2016.

[10] Youze Tang, Xiaokui Xiao, and Yanchen Shi. Influence maximization: Near-optimal time complexity meets practical efficiency. 2014.

[11] Thomas and Swiss. Black noise: Rap music and black culture in contemporary america. by tricia rose. hanover & london. wesleyan university press, 1994. xvi + 241 pp. *Popular Music*, 1995.

One-Page Document for the ICM Society

Value of Our Approach

Hello, respected ICM Society! It's my great honor to introduce our work to you. There are four main parts in our work. They are interrelated and progressive, forming a complete system. We are confident that our work can make contributions to the related research. Next, I will list some highlights and advantages of our approach.

- First, we propose a unified scheme based on greedy algorithm to find the seed set in influence network, which can be used to find revolutionaries together with dynamic influence indicator. The advantage of our influence network is its scalability. Within this framework, you can add more complex features of node/edge and mechanisms for influence spread, which can be applied to a variety of scenarios.

- Second, our work put forward a proper measure of similarity to reveal the similarity and difference of music within or between genres. This measure helps us take a close look at the relationship between music. It shows that, as time goes by, the development of music can be affected by music of other genres. Some genres may have similar development mode, while some may be opposite. This is quite exciting because it indicates the interaction of different genres. In addition, We explore the use of XGBoost to calculate feature importance, which achieves very good performance. In this way, we can analyze how a famous influencer affects his followers. This will also help us to infer how the evolution of genres took place.

- Third, the change points detecting technology provides a more vigorous method to analyze the major shift in characteristics, with which we identify two major revolutions through history. Further, it works quite well when it comes to predict the emergence of a new kind R&B in 1980s. Meanwhile, the dynamic influence model, aimed to measuring artist's influence during a certain period, will show the list of the most influential artists in different periods. It's good to discover potential talent musician by using this easily-calculate model.

Application

For other new datasets, it is very easy to apply our approach. First, according to a series of formulas in subsection 3.1, features of each node can be calculated, such as static influence. Second, you should choose a measure of music similarity, like euclidean distance or cosine distance. Third, you could use change points detecting technology like *bcp* to analyze the major shift in characteristics. Fourth, you can calculate the important indicator - dynamic influence by Equation 11. Last, according to the network of influencers and followers, run Algorithm 1, and you can get the list of the most influential people in specific periods. The rest of the analysis methods, such as t-test, clustering, classification model, can be carried out according to the way in our work. Not only that, you can also reset the node features and influence spread. Our method has a strong scalability, please rest assured to use.

Further Study

For further study, we can build up a larger influence network: social influence network, where some social and cultural indicators, such as stability of politics, social happiness, level of economic development, education level and so on will be added in. The indicators we choose should be able to reflect the levels of some aspects of our society or culture. In this way, we can study the relationship and interaction between music characteristics and social indicators. For example, we can study whether the trend of *energy* in music is connected with the economy. This relationship and interaction is worth taking note of, as it reveals the influence of rise and fall in music on the society quantitatively.

We can also pay more attention to the effect of music on social thinking. Some literature show that, the rise of some philosophical schools gave birth to some genres of music. We can quantitatively measure their correlation by putting forward similar influence networks. Moreover, we may test the hypothesis that music can also give rise to philosophy. These questions are mostly studied qualitatively before, what about building a statistical model? This will give us more insight into music and its relationship with the society and culture.