Problem Chosen	2021 MCM/ICM Summary Sheet	Control Number
D		2100112

Music Influence and Evolution Analysis Based on Networks

Summary

"Music is like a psychiatrist. It will answer you with things people can't tell you." said the Beatles, one of the most famous Pop/Rock artist. In this paper, we deeply explore the magic of music - its influence network, evolution path, and effects on the culture.

For Task 1, we build a directed network between influencers and followers. Based on network theory, we proposed three different metrics - Degree Centrality, Weighted Degree Centrality and Eigen Centrality. Then we develop a combination of these metrics as a comprehensive measure of music influence. After that, we create a subnetwork to illustrate our influence measure.

For Task 2, we first preprocess the data using Principal Component Analysis to reduce dimension and collinearity. Then we define and calculate the distance between different tracks to obtain the similarity between artists. By calculating the average similarity, we apply Mann-Whitney test. Results show that with the probability of 62.8% and p-value smaller than 0.001, artists within a genre are more similar than those between genres.

For Task 3, according to proposed music similarity measure, we find that similarities and influences within and between genres differ greatly when we analyse different genres. To distinguish one genre from others, we build a Genre Classification Tree Model. By analysing the number of influencers in different period, we explore the evolution path of genres. Based on the directed network on genre scale, we find Pop/Rock has a strong relationship with R&B, Blues and Folk.

For Task 4, we build a Similarity Bayesian Network to identify the real followers based on the music characterstic similarity in tracks. Then we apply a multivariate two-sample mean test and find there is no strong evidence (p-value > 0.1) that any music characteristic are more "contagious" than others.

For Task 5, we first analyze the rise and decline of genres and find the music revolution in 1950s. We proposed a Dynamic Programming Algorithm to detect change points in music characteristics which are consistent with the revolution. Our results show that acousticness, energy, danceability and loudness might signify revolutions. Based on the Bayesian Network, Elvis Presley and Cliff Richard represent revolutionaires.

For Task 6, to get a further insight into the evolution in Pop/Rock, we propose an Dynamic Influencer Indicator based on the lagging trend in music characteristics of the whole genre. From 1960s to 2010s, there are 10 dynamic influencers, each has his/her unique influence on the genre. Moreover, we explain the evolution of Pop/Rock.

For Task 7, three important periods are detected based on time series analysis, which show the culture-influence of music. Based on the model established, we identify social changes like countercultural movement and technological changes such as the proliferation of the Internet.

At last, we conduct sensitivity analysis, which shows the robustness of our model. We also summerize the stengths and weaknesses and provide insights to ICM society about the evolution and cultural influence of music.

Keywords: Directed Network; Bayesian Network; Change Points Analysis

Contents

1 Introduction

Nowadays, various kinds of music has increasingly become an indispensable part of human life. Thousands of music artists influence each other and form a complex music influence network. While some genres show great similarity to each other, other genres are quite different in music characteristics. Some artists are passionate revolutionaries, who lead to an emergeence of a new genre or reinvention of an existing genre. Despite culture influence, the change of music characteristics of artists also indicate external events like the proliferation of the Internet. In order to further explore the music influence network and the role music has played on the society, it is necessary to quantify music evolution.

2 Problem Restatement and Analysis

- Task 1 requires us to build a complex network between influencers and followers based on dataset "influnce_data.csv" and develop metrics to capture the music influence in the network. The key to this problem is to define directed influencer network and propose metrics that comprehensively measure the music influence of each influencers in the network.

- Task 2 requires us to use various musical characteristics in dataset "full_music_data.csv" to measure music similarity and judge whether musicians in the same genre are more similar than those in different genres. It is of great essence to define the distance between music works of artists, from which we can obtain the similarities between artists.

- Task 3 requires us to compare similarities and influences between and within genres. It is necessary to define the distance between genres and measure similarities between them. We plan to build a classication tree to distinguish one genre from others. Besides, we can show how genres changing over time and the relationship between genres using visualization.

- Task 4 requires us to get a further insight into the similarity of music characters between influencers and followers. We plan to build a Bayesian Network to find the real influencer and use hypothesis test to evaluate whether some music characteristics are more 'contagious' than others.

- Task 5 requires us to identify the characteristics and major artists that signify music evolution. To handle this problem, we propose a Change Point Detection model based on DP algorithm and match the change of music characters with the musical revolution. Then find revolutionaries via Bayesian Network.

- Task 6 requires us to develop indicators to analyse dynamic influencers and corresponding influence processes of musical evolution in one genre. In this problem, we will focus on Pop/Rock and then discuss the major contributors to the Pop/Rock evolution through decades.

- Task 7 requires us to identify the effects of social, political, or technological changes within the network and find the cultural influence of music. Based on time series analysis, we intend to find the connection between the change in music characteristics and the external events. We also detect several culture-influence of music in different time or circumstances.

The workflow of this paper is shown in Figure 1.

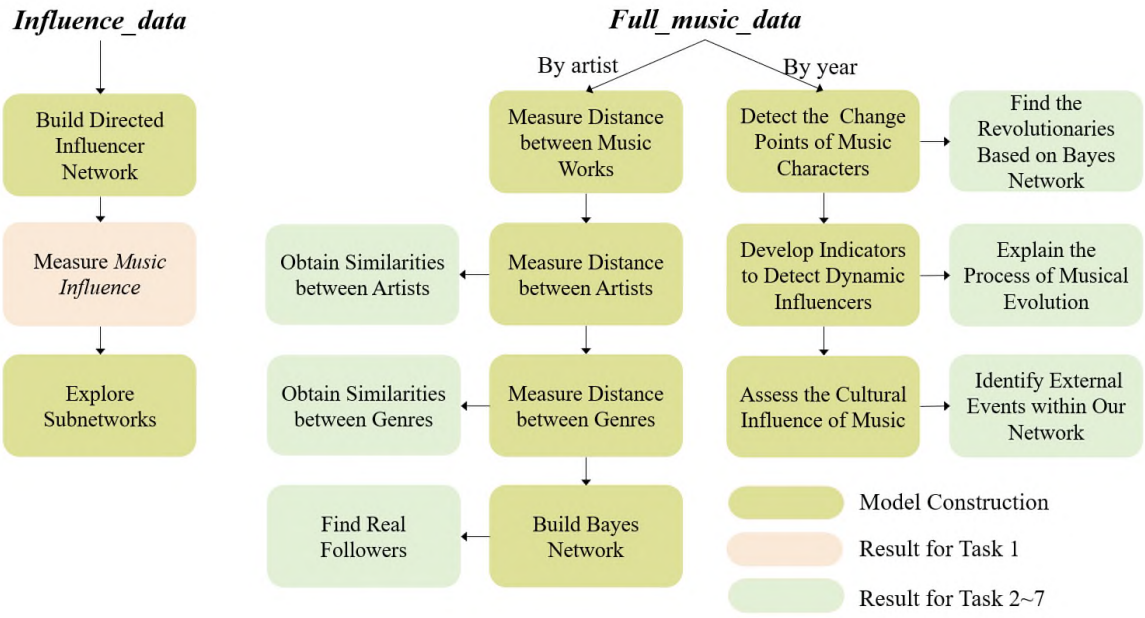


Figure 1: Workflow of the paper

3 Assumptions and Notations

3.1 Assumptions and Justifications

- **The similarity of artists is represented by the similarity of their musical characteristics.** In all available datasets, the most reliable souce of information about an artist's characteristics is tracks he/she released. Therefore, it is reasonable to use musical characteristics to represent the characteristics of artists.
- **The reinvention and revolution of exisiting genre can be can be signified by the sharp change in music characteristics.** Revolution shifts in a genre will denitely change the music characteristics, thus we can capture revolutions based on these changes.
- **In different stages of genre development, music characters changed with a linear trend over time.** This assumption is a reasonable simplification to make breakpoints identification possible.

3.2 Notations

The primary notations are shown in Table 1.

Table 1: Notations

Symbol	Definition
DC_i	the local influence Degree Centrality of i th influencer
WDC_i	weighted degree centrality for i th influencer
EC_i	eigen centrality for i th influencer
$F-Score_i$	the comprehensive score of i th influencer
$Sim_{i,j}$	music similarity between track i and track j
$Acv_{i,j}$	absolute coefficient of variation of music character j in genre i
ρ_{AB}	similarity score between artist A and artist B
μ	length of lagging year

4 Models and Solutions

4.1 Task 1

4.1.1 Directed Influencer Network

We consider influencers and followers as nodes and collect all musicians together to make up the set $V = \{v_i\}_{i=1}^{n}$. If artist i has an influence on artist j, then an edge from node i to node j will be generated. All the edges make up the set $E = \{e_j\}_{j=1}^{m}$, and the edges from node i make up set $N(i)$. Based on the dataset "influnce_data.csv", we build a complex network involving n=5603 nodes (artists) and m=42770 edges (influence) in total, as shown in Figure 2.

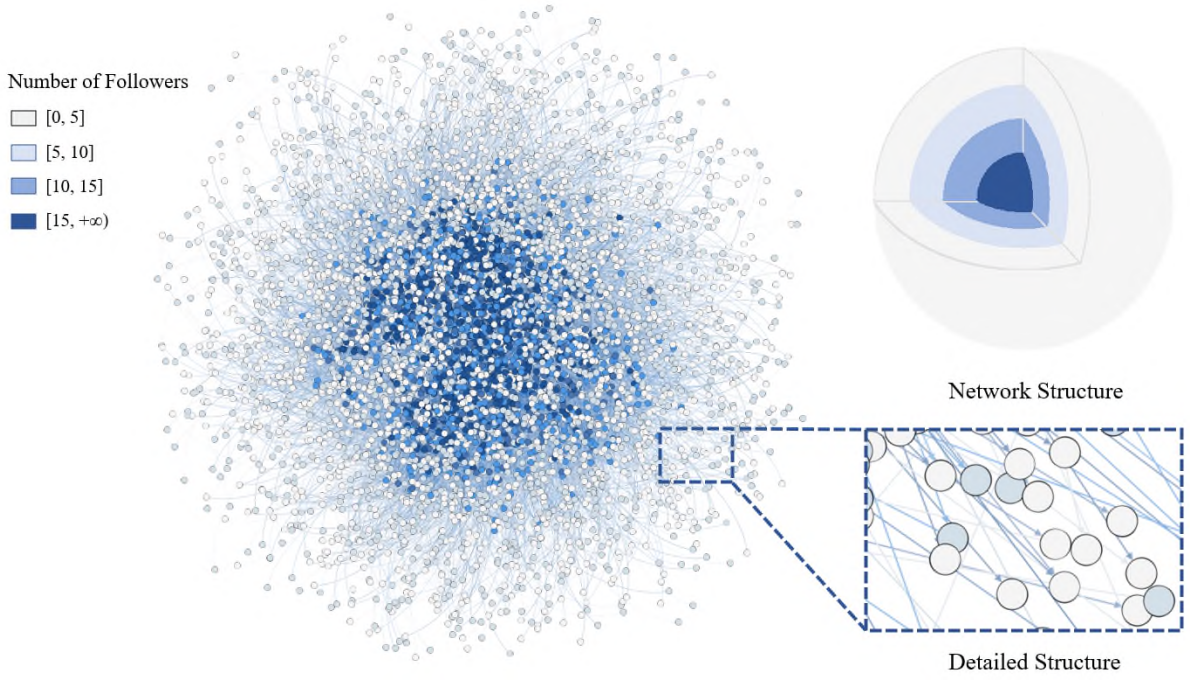


Figure 2: 3D Directed Influencer Network

4.1.2 Music Influence Measure

We now define some essential metrics to measure the music influence in the network.

Degree Centrality: Degree is an important concept in network theory. In directed graph, out degree of node v represents the number of edges from node v[1].

$$outdegree_v = \#N(v) \tag{1}$$

A naive idea to measure the local importance of the node is the out degree of the node. In other words, we use the number of followers to measure the local influence of a musician. We call the local influence Degree Centrality and define the Degree Centrality of influencer i as follows:

$$DC_i = outdegree_i/n \tag{2}$$

Where n is the number of nodes in the network.

Weighted Degree Centrality: Some genres interact with each other closely, while others seem to have little connection. As a result, we can't treat all the influences equally, and the edges of the network should share different weights. If a musician has an influence on artists of other genres, it means that the musician has a wide range of influence. Similarly, if a musician has an influence on future musical generations decades later, it indicates that the influence lasts for a long time. In the cases above, we align greater weight.

We define the genre of musician i as $G(i)$. The *yeargap* between followers and influencers is defined as the time difference between the start of their careers. When the *yeargap* exceeds the *threshold*, we call it long-time influence, otherwise short-time influence. Here we set the $threshold = 20$.

The weight matrix $W = (W_{ij})$ is defined by integrating influence range and influence duration as follows:

$$W_{ij} = \begin{cases} \frac{1}{3} * (1 + I_{\{S(i) \neq S(j)\}} + I_{\{yeargap > threshold\}}), & j \in N(i) \\ 0, & else \end{cases} \tag{3}$$

Then we propose the Weighted Degree Centrality (WDC) to modify the Degree Centrality above:

$$WDC_i = \frac{1}{n} * < W_i, 1_n > \tag{4}$$

Where W_i represent the ith row of matrix W, 1_n is a column vectors with all entries 1. DC and WDC both measure the local influence of an influencer.

Eigen Centrality: The basic idea of Eigen Centrality is to regard the influence of a node as a function of local influence of its adjacent node. In other words, the higher influence the artist's followers have on others, the greater the Eigen Centrality of the artist himself/herself is.

Eigen Centrality is defined as follows:

$$EC_i = \frac{1}{n} * < W_i, OD > \tag{5}$$

Where $OD = (outdegree_1, outdegree_2, \ldots, outdegree_n)^T$ and $W_i = (W_{i1}, W_{i2}, \ldots, W_{in})^T$ both denote a column vector. Intuitively, Eigen Centrality proportionally allocate the Degree Centrality from adjacent nodes to all nodes, which seems to "spread out" the Degree Centrality.

Comprehensive F-Score: The three different degrees above measure the music influence of artists in the network in different aspects. In order to get the comprehensive measure of each

influencer, we use weighted sum of the three degrees. In order to measure relative values, each degree is divided by the corresponding maximum value before the weighted sum:

$$F-Score_i = w_1 * \frac{DC_i}{\max\limits_k(DC_k)} + w_2 * \frac{WDC_i}{\max\limits_k(WDC_k)} + w_3 * \frac{EC_i}{\max\limits_k(EC_k)} \tag{6}$$

F-score, as a combinaiton of all three metrics, comprihensively measures the music influence of each influencer in the complex network.

4.1.3 Solutions

Based on the music influence measurement F-score (here we set $w_1 = w_2 = w_3 = \frac{1}{3}$), we extracted 10 directed influencer subnetworks from the original network, as shown in Figure 3. All metrics of these 10 influencers is shown in Table 2. The size of the core in each small subnetwork indicates the music influence of the top artist. It is clear that the subnetwork shows a radial-structure, with a great artist as the center, connecting to his followers.

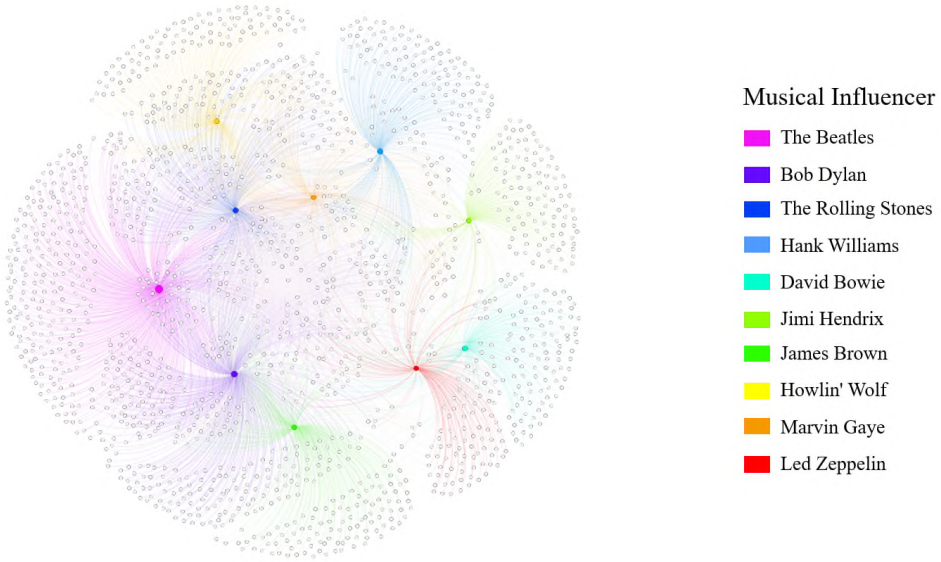


Figure 3: Directed Influencer Subnetwork

Generally, artist with higher music influence has more followers. However, Howlin' Wolf has less followers than Marvin Gaye, whereas has greater music influence. This is because followers of Howlin' Wolf are more famous and influential.

4.2 Task 2

4.2.1 Data Preprocessing

First, we preprocess the dataset "full_music_data.csv". There are 14 characteristics related to music included in dataset, including danceability, energy, loudness, etc. Through data visualization, we find that the Boolean variable "Explicit" is invalid: Out of 98430 tracks, only 3647 tracks are marked as 1, which means less than 4% tracks have explicit lyrics. As it conflicts with common sense, we believe most lyrics in tracks remained undetected. Thus, we remove "Explicit" data. After that, we standardize all continuous variables.

Table 2: Metrics of top 10 artists

Id	Name	Genre	DC	WDC	EC	F-Score
754032	The Beatles	Pop/Rock	0.11	0.14	2.3	1.00
66915	Bob Dylan	Pop/Rock	0.07	0.1	1.65	0.68
894465	The Rolling Stones	Pop/Rock	0.06	0.07	1.24	0.52
549797	Hank Williams	Country	0.03	0.07	1.57	0.49
531986	David Bowie	Pop/Rock	0.04	0.06	0.71	0.37
354105	Jimi Hendrix	Pop/Rock	0.04	0.05	0.94	0.37
128099	James Brown	R&B	0.03	0.05	1.11	0.36
276085	Howlin' Wolf	Blues	0.02	0.04	1.3	0.35
316834	Marvin Gaye	R&B	0.03	0.06	0.72	0.34
139026	Led Zeppelin	Pop/Rock	0.04	0.05	0.65	0.34

Some of the music characteristics have similar meanings, for example, "energy" and "loudness" both reflect the intensity and activity of tracks. To reduce the influence of collinearity when calculating similarity, we use PCA (Principal Component Analysis) to reduce the dimension of the data while preserving as much of the data's variation as possible. After calculation, the cumulative variance contribution rate is shown in Table 3.

Table 3: Cumulative Variance Contribution Rate

Number of Principal Component	1	2	3	4	5	6	7	8
Cumulative Variance Contribution Rate	0.23	0.37	0.49	0.59	0.67	0.74	0.81	0.85

Based on the result of PCA, we choose first seven principal component and ignore the rest. The seven new variables maintain more than 80% information of the original data.

4.2.2 Similarity Measure and Test

Inspired by the Euclidean distance, we define music similarity between track i and track j as follows:

$$Sim_{ij} = \frac{1}{\sqrt{\sum_{t=1}^{m} \left(x_{it} - x_{jt}\right)^2}} \qquad i, j = 1, 2, \ldots, m \tag{7}$$

Comparing to other measures like Mahalanobis distance, this simple measure doesn't need any assumption on data and is easy to calculate.

To explore whether artists in the same genre are more similar than those in different genres, we construct Mann-Whitney hypothesis test. Comparing to traditional tests such as two-sample t-test, Mann-Whitney statistic does not require distributional assumptions for its validity. In this problem, the distributions of music similarity within and between genres are quite far away from the normality. Thus, this statistic will bring about better results.

We first calculate the average music similarity within genre and between genres, which can

be expressed as

$$Sim_{in,i} = \frac{1}{n}\sum_{j=1}^{n} Sim_{ij} \tag{8}$$

$$Sim_{bet,i} = \frac{1}{m}\sum_{k=1}^{m} Sim_{ik} \tag{9}$$

where j, k denote individual in the same genre and different genres with artist i, and n, m denote number of above individuals, respectively.

Then we calculate Mann-Whitney statistic as:

$$U = \frac{1}{(n+m)^2}\sum_{i=1}^{n+m}\sum_{j=1}^{n+m} I\left(Sim_{in,i} < Sim_{bet,j}\right) \tag{10}$$

Based on Central Limit Theorem, the asymptotic distribution of $Z = \frac{U-\frac{1}{2}nm}{\frac{1}{12}nm(n+m+1)}$ is normal distribution[4]. Based on this, we can construct Mann-Whitney Test to find out whether artists within a genre are more similar than those between genres.

4.2.3 Solutions

Similarity between all artists are shown in Figure 4. Mann-Whitney statistic shows that with a probability of 62.8% and p-value smaller than 0.001, artists within genre are more similar than those between genres.

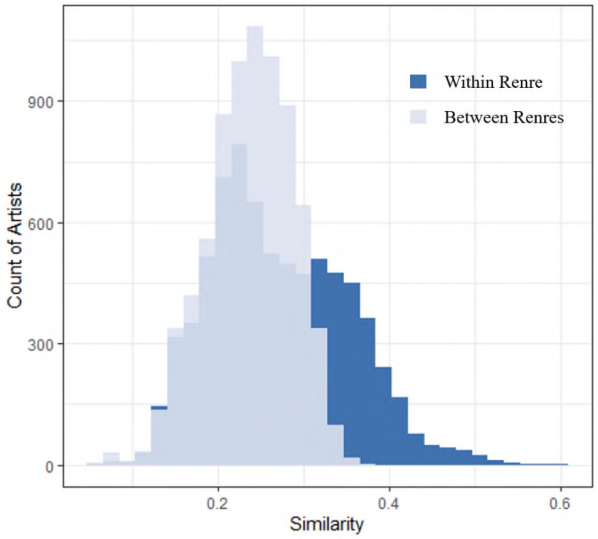


Figure 4: Similarity between artists within a genre and between genres

4.3 Task 3

4.3.1 Genre Similarity and Influence

To compare music similarity and influence on the genre level, we first merge genre in dataset "influnce_data.csv" to dataset "data_by_artist.csv", then we group the data by different genres.

We define the similarity between genres as

$$Sim_{ij} = \frac{1}{\sqrt{\sum_{k=1}^{n_g} \left(c_{ik} - c_{jk}\right)^2}} \qquad i, j = 1, 2, \ldots, n_g \tag{11}$$

where n_g is the number of genres, c_{ik} is the kth average music characteristic value in genre i.

To evaluate the similarity within genres, we calculate the Absolute Coefficient of Variation of each music character in different genres, which quantifies the variation of some characteristics within genres. Let Acv_{ij} denote the Absolute Coefficient of Variation of music character j in genre i, then:

$$Acv_{ij} = \frac{\sqrt{\frac{1}{\#G(i)} \sum_{k \in G(i)} \left(x_{jk} - \frac{1}{\#G(i)} \sum_{k \in G(i)} x_{jk}\right)^2}}{\left|\frac{1}{\#G(i)} \sum_{k \in G(i)} x_{jk}\right|} \tag{12}$$

where x_{jk} denotes the music character j of the artist k. By averaging the Acv of all continues music characters, we obtain the Average Absolute Coefficient of Variation for each genre.

4.3.2 Genre Classification Tree

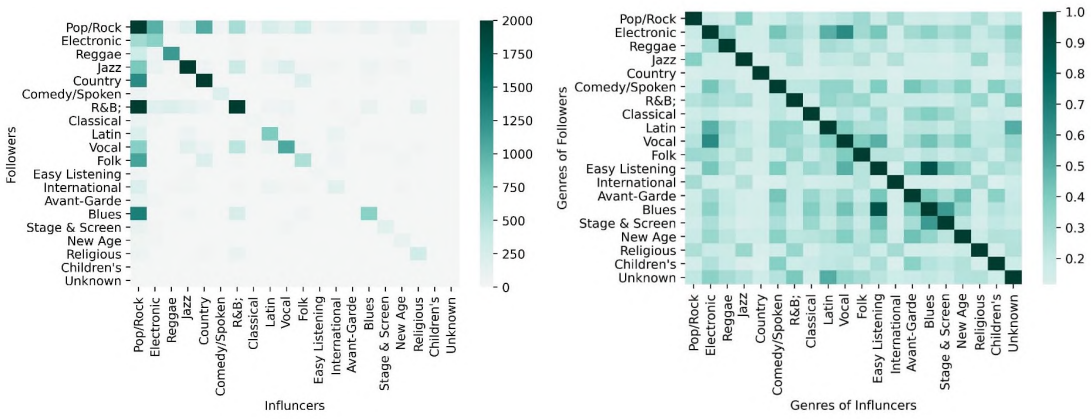
According to our assumption, when we distinguish one genre from another, the main difference lies in the creation styles - the various features of the music created by the artists. Considering each genre has its unique style, we separately analyze the distinguishing features of each genre.

Classification tree can distinguish the relative importance of features: A feature closer to the root node is more important than those closer to the leave node. The specific steps to construct a Genre Classfication Tree are listed as follows:

1. Identify the genre to be analyzed. Mark the artists in the genre as positive and the remaining artists as negative.
2. Construct the classification tree and choose the appropriate tree size to keep the simplicity of the classification standard.
3. Sort the importance of features according to their relative position in the tree, visualize the tree and offer distinguishing strategy to identify the genre from others.

4.3.3 Solutions

Figure 12 shows the Similarity Matrix and Influence Matrix between Genres. Blues and Easy Listening share some similar characteristics in musical features, whereas Electronic and Vocal are more similar. As for influence, Pop/Rock significantly influences other genres. Interestingly, Pop/Rock has a strong influence on R&B, but they do not share significant similarity in musical characteristics, which we will discuss the reason later.



(a) Influence between and within genres (b) Similarity between genres

Figure 5: Influence and Similarity between and within genres

Figure 6 and Figure 12a show the similarity and influence within a genre. Some genres like R&B and Country have more variation within the genre than others, which indicates artist's desire for true freedom of expression.At the same time, Pop/Rock and Jazz have a greater influence on themselves than others, which means they tend to maintain a consistent style.

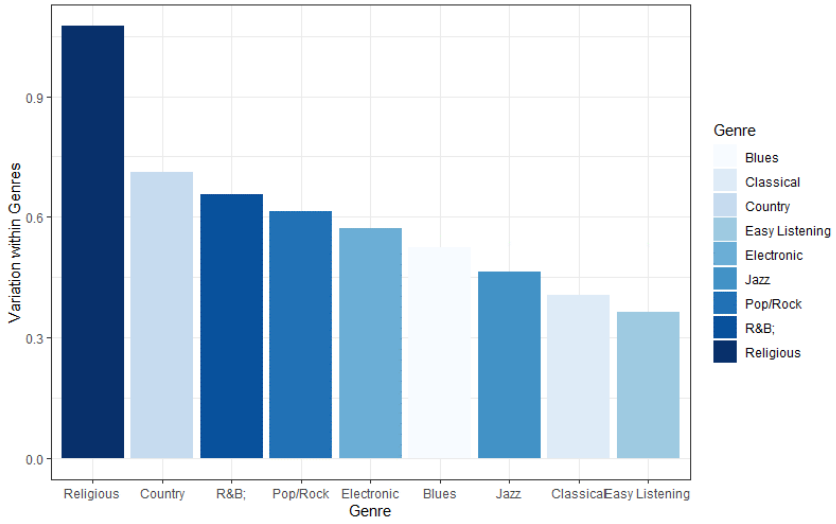


Figure 6: Similarity within genres

Set R&B as an example, we construct classfication tree as shown in Figure 7. **danceability**, **duration** and **instrumentalness** are vital features that distinguish R&B from other genres. Based on this approach, we can find the vital features of every genre.

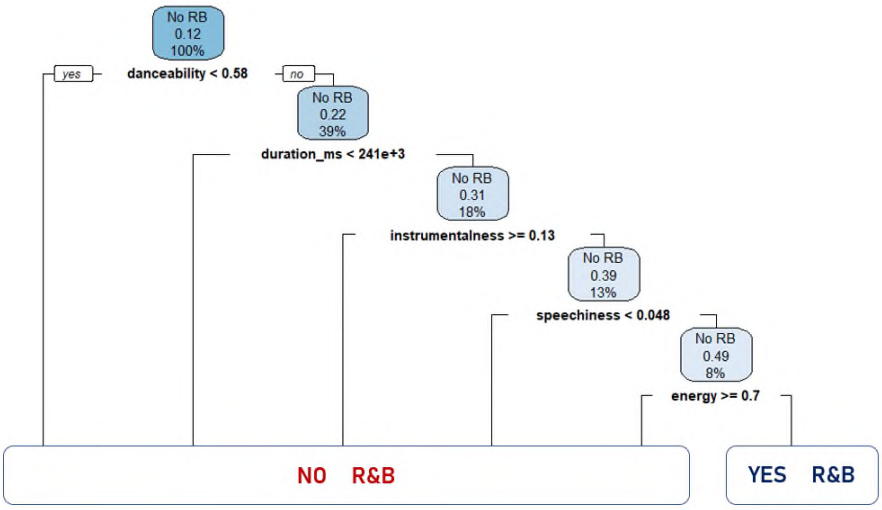


Figure 7: R&B Classification tree

We use dataset "influence_data.csv" to depict how genres change over time. Based on the number of influencers in different genres through different time period, we draw Figure 8. As shown in the figure. **Blues** and **Jazz** flourished from 1930s to 1950s, and then met a smooth decline. On the other hand, **Country** kept popular from 1930s to 1990s. **Pop/Rock**, similar to **R&B**, thrived between 1960s and 1990s.

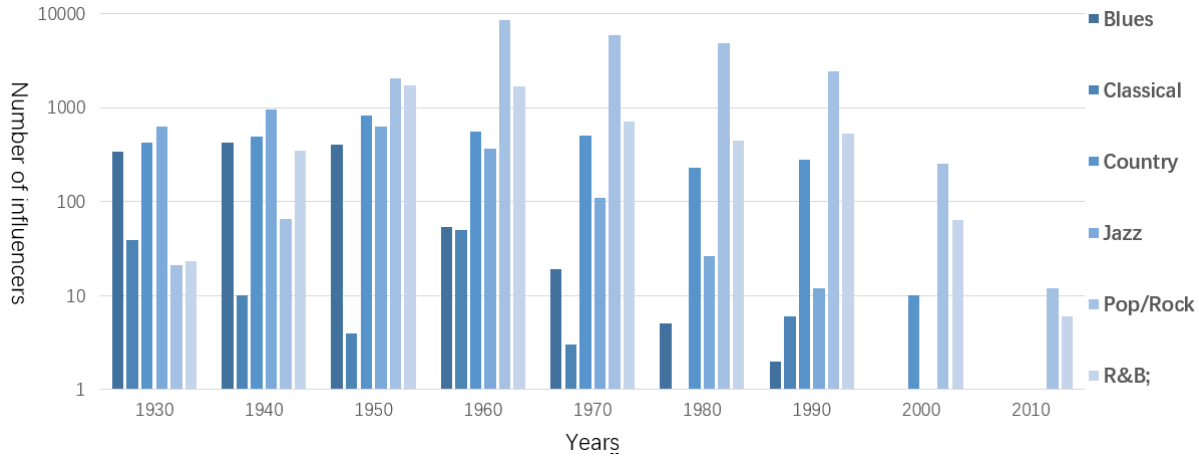


Figure 8: Genres changing over time (Part)

To further explore the relationship between different genres, we construct a directed genre network based on the Influence matrix, as shown in Figure 9. Intuitively, if a genre is closely related to another, artists tend to have a great influence on each other's tracks. Therefore, we consider "music influence" as the major measure of relationship.

The Genre Network illustrates that Pop/Rock artists have a great influence on many other genres, especially R&B, Blues and Folk etc., relating these genres together. In addition, Pop/Rock artists are deeply influenced by Electronic. Learning from Pop/Rock Musicians, artists in R&B have a number of interactions with artists in Avant-Grade.

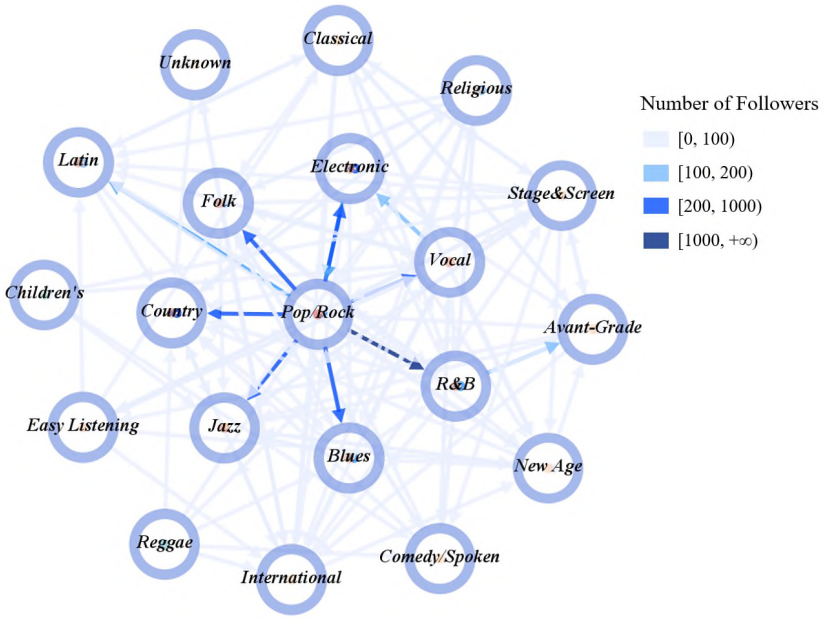


Figure 9: Genre Network

4.4 Task 4

4.4.1 Similarity Bayesian Network

This subsection aims to explore whether so-called 'influencers' truly have an influence on music created by 'followers', in other words, whether there is a significant similarity between influencers' music and pieces created by their followers.

To tackle this problem, we construct a Bayesian network based on similarity data. Here we define the similarity score between artist A and artist B as

$$\rho_{AB} = \frac{\sum_i^{n_c} \left(Cha_{Ai} - \bar{Cha}_A\right)\left(Cha_{Bi} - \bar{Cha}_B\right)}{\sqrt{\sum_i^{n_c} \left(Cha_{Ai} - \bar{Cha}_A\right)^2 \sum_i^{n_c} \left(Cha_{Bi} - \bar{Cha}_B\right)^2}} \tag{13}$$

which has similar form to Pearson correlation, n_c denotes the number of music characteristics, Cha_{Ai} denotes ith standardized music characteristic score of artist A, and $\bar{Cha}_{Ai}$ denotes the mean.

In our Bayesian Network, nodes denote artists, edges denote similarity scores. According to Fisher Z-Test[2], if A and B doesn't have any similarity given C, then

$$\frac{\sqrt{n - |Cha_C| - 3}}{2} log \frac{1 + \rho_{AB|C}}{1 - \rho_{AB|C}} \to N(0,1) \tag{14}$$

Based on the asymptotic distribution, we can calculate the similarity score between any two artists, then construct a Similarity Bayesian network. However, this traversal algorithm has NP problem. Therefore, we used Hill-Climbing algorithm with Random Starts instead. Define the Bayesian information as

$$BIC = \sum_{i=1}^{n_c} log f_{A_i}(A_i | \Pi_{A_i}) - \frac{d}{2} log(n) \tag{15}$$

Algorithm 1 Similarity Bayesian Network

```
while i ≤ Max_RandomNumber do
  Randomly generate a Similarity Bayesian network
  Calculate BIC_init
  BIC_best ← BIC_init
  while j ≤ Max_iterations do
    Randomly choose add or delete an edge
    Calculate BIC_new
    if BIC_new ≤ BIC_best then
      Net_best ← CurrentBayesianNetwork
      BIC_best ← BIC_new
    end if
  end while
end while
Rank the BIC_best of each i and find the best Similarity Bayesian network
Postprocess results and visualization
```

Where $log f_{A_i}(A_i|\Pi_{A_i})$ is the conditional density function, the pseudocode of the algorithm is shown as follows:

To improve the stability of the solution, we use bootstrap method to calculate 500 different solutions and take an average.

4.4.2 Contagious Characteristic Test

In this subsection we analyse if some music characteristics are more'contagious' than others, or they all have similar roles in influencing a particular artist's music. Based on the results in the last subsection, we tackle this task by analyzing The Beatles and their major followers: The Beach Boys, Led Zeppelin, Gram Parsons and David Bowie.

To find out if there exists characteristics significantly more contagious than others, we apply multivariate two-sample test[3]. The null hypothesis H_0 is

$$H_0 : \mu_1 = \mu_2 \tag{16}$$

Where μ_1 denotes the average music characteristic vector of all tracks released by The Beatles. μ_2 denote the average music characteristic vecter of one of their followers.

According to asymptotic statistic theory, when p is fixed and n tends to infinity, by the Central Limit Theorem and Slutsky Theorem:

$$(\bar{X}-\bar{Y})^T \left(\frac{S_1}{n_1}+\frac{S_2}{n_2}\right)^{-1} (\bar{X}-\bar{Y}) \to \chi^2(p) \tag{17}$$

where $\bar{X}$ and $\bar{Y}$ denote the sample mean respectively, and S_1 and S_2 denote the sample covariance matrix respectively.

Since the number of music characteristic (p = 9) is much smaller than the sample size, we do not need to make any distribution assumption for the data. The Chi-squared statistic follows a Chi-squared distribution with 9 degrees of freedom. By calculating the statistic, we can obtain the p-value according to the asymptotic distribution.

If the null hypothesis is accepted, we conclude that all music characteristics of followers are equally similar to their influencer, indicating no characteristics are significantly 'contagious'.

On the other hand, if the null hypothesis is rejected, we can further apply BH multiple hypothesis testing to further find out which characteristic is more 'contagious'.

4.4.3 Solutions

Without the loss of generality, we put our focus on the most popular band – The Beatles and their followers. Setting similarity score threshold as 0.21, the Similarity Bayesian Network of the Beatles and their 29 most popular followers is shown in Figure 10.

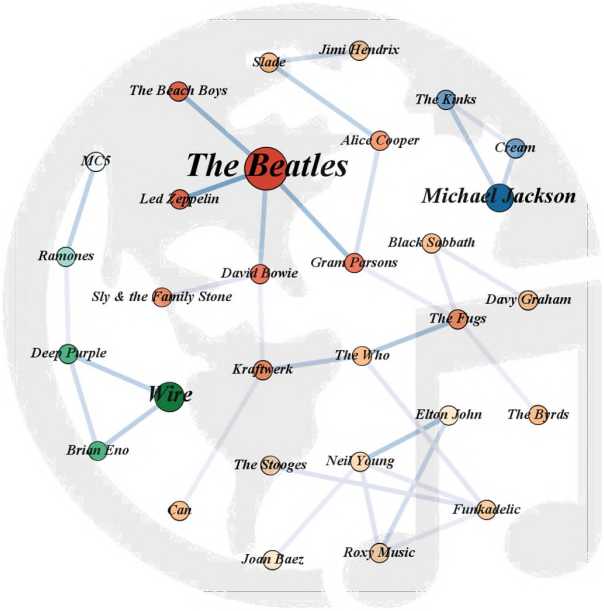


Figure 10: Similarity Bayesian Network of the Beatles and their followers

As shown in figure 10, directly or indirectly, 72% alleged 'followers' have similar composing style to the Beatles, their 'influencer'. However, 28 % artists who claim they are influenced have insignificant (lower than the threshold) similarity to their 'influencer'. In other words, based on similarity data, nearly three out of ten artists aren't affected by their so-called 'influencers'.

Table 4: Contagious Characteristic Test

Follower Name	chi-square value	p-value
Led Zeppelin	3.183	0.957
Gram Parsons	5.971	0.743
David Bowie	7.232	0.613
The Beach Boys	14.284	0.113

The results of Contagious Characteristic test is shown in Table 4. Under the significant level of 5%, all four null hypotheses are accepted, which indicates that all musical characteristics of The Beatles equally influence its followers, in other words, no characteristics are more contagious and prominent than others.

we should emphasize that in the above process, Type I error may accumulate due to multiple hypothesis tests. However, in this problem, all hypotheses are accepted. Thus, Type I errors can still be controlled within the significance level.

4.5 Task 5

4.5.1 Definition of Revolution

To tackle task 5, we first define "music revolution". The number of influencers in major genres is shown in figure 11. As time went by, some genre flourished and others went out of favor.

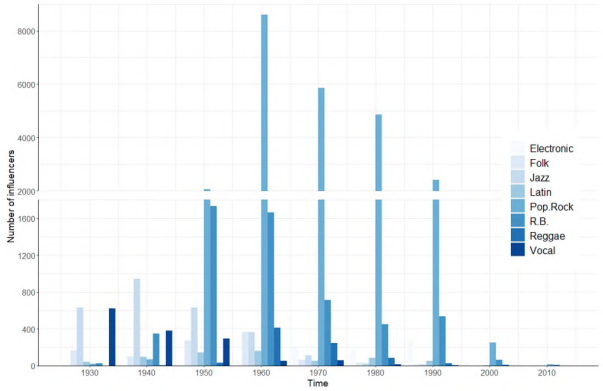


Figure 11: The number of influencers in major genres

We define the revolution time for genres as the time when there are some genres declined and others springing up or dominating. From Figure 11, we find in the 1950s, Electronic flourished gradually, while Vocal all went from boom to bust. The popularity of Pop/Rock and R&B also met a sharp increase. These phenomena indicate that 1950s is one of the revolution periods.

After defining revolution periods, we focus on finding a significant change point of music characteristics in the 1950s to explain what we discussed above. The key to this problem is turned to identify the change points in time series of the music features. Based on the change points we find, we then can identify artists that represent the revolutionaries.

4.5.2 Change Points Detection (DP Algorithm)

We define the set of the change points as $\Theta = \{\theta_i\}_{i=1}^{m}$. Let the time series of some music character $\{Y_t\}_{t=1}^{n}$ admit:

$$Y_t = a_i + b_i * \left(\frac{t}{n}\right) + \varepsilon_t \quad \theta_{i-1} + 1 \leq t \leq \theta_i \tag{18}$$

Where a_i is the intercept term, b_i measures the speed of linear trend, and the error term ε_t is a white noise process.

In order to ensure that the stability of parameters and change points, we also need to limit

$$\theta_i - \theta_{i-1} > \tau \tag{19}$$

and we suggest $\tau = [0.1n]$, which will be disucssed in detail in sensitivity analysis.

Since y_t changes with a linear trend in different stages, when ε_i follows independent and identical distribution, it is apparent to estimate parameters by minimizing the sum of squared residuals.

$$\{\theta_i\}_{i=1}^{m} = \underset{\theta \in \Theta}{argmin} \sum_{t=1}^{n} (y_t - \widehat{y_t})^2 \tag{20}$$

When the number of change points is given, the key of using Dynamic Programming algorithm in this problem is to establish the recursive relationship of the sum of squared residuals before and after a new change point is added.

Let $\delta(m,n)$ denote the sum of squared residuals associated with the optimal partition containing m breaks using first n observations, $RSS(i,j)$ denote the sum of squared residuals obtained by applying least-squares to a segment start from i to j.

Then, we can achieve the global minimization of overall sum of squared residuals by solving recursive problem as follows:

$$\delta(m,n) = \min_{m\tau \le j \le n-\tau} \delta(m-1,j) + RSS(j+1,n) \tag{21}$$

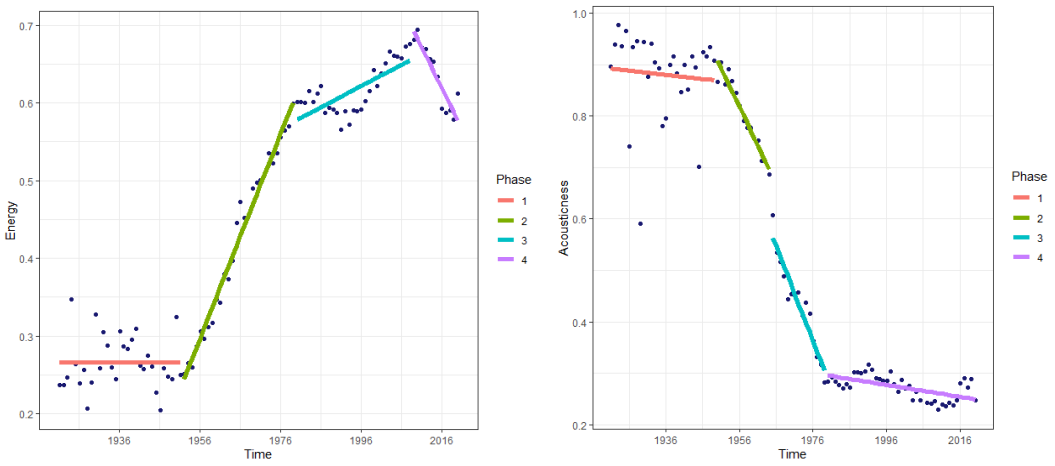
where τ is a trimming parameter that constrains the distance between two change points not to be too close.

It is important to note that the DP algorithm is $O(T^2)$ and does not depend on the number of change points. Therefore, we can quickly estimate all the change points which are consistent for real ones[5].

Obviously, we also need to penalize the number of change points. Considering the number of change points as a tuning parameter, We recommend to use BIC criterion to determine the optimal number of change points m^*.

$$m^* = \underset{m}{argmin} RSS_{overall} + \log(n) * \hat{\sigma}^2 * m \tag{22}$$

4.5.3 Solutions



(a) Energy (b) Acousticness

Figure 12: Change points detection of acousticness and energy

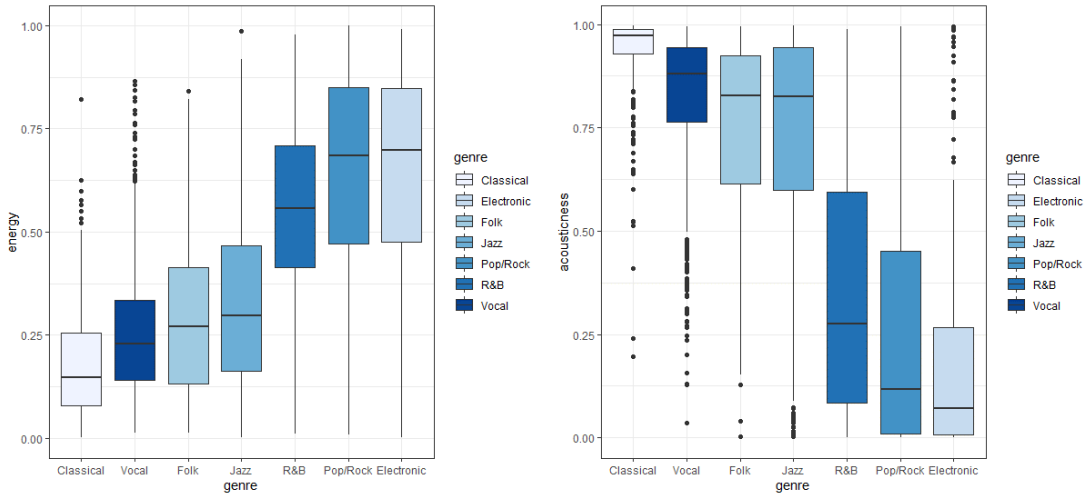
By using the DP algorithm above, we detected the change points of 11 continuous music characteristics in the dataset "data_by_year.csv", as shown in table 5.

It can be seen in table 5 that four music characters (acousticness, energy, danceability and loudness) show significant changes around 1950.

Table 5: Revolution date of music characteristics

	Revolution date							
Characteristic	1930s	1940s	1950s	1960s	1970s	1980s	1990s	2000s
Instrumentalness	1933	1946	-	1964	-	-	-	-
Duration_ms	-	1946	-	1966	-	-	-	2007
Acousticness	-	-	**1950**	1964	1979	-	-	-
Tempo	-	1947	-	-	1979	-	1996	2008
Danceability	-	-	**1950**	-	-	-	1997	2008
Valence	-	1947	-	1966	-	-	-	2005
Energy	-	-	**1951**	-	1979	-	-	2008
Liveness	-	-	1956	-	1976	-	-	2008
Speechness	-	-	1956	-	-	-	-	2006
Popularity	-	-	1953	-	1970	-	-	2006
Loudness	1936	-	**1950**	-	-	-	-	2008

To get a further insight into the revolution related to music characteristics in different genres, we draw boxplots of energy and acousticness respectively in figure 13. Obviously, the new genres (Pop/Rock, R&B and Electronic) have significant higher energy, yet their acousticness is much lower than the old ones. This is consistent with the time series analysis above. Thus, we convinced that the four music characters: **acousticness**, **energy**, **danceability** and **loudness** might signify revolutions in musical evolution.



(a) Energy (b) Acousticness

Figure 13: Energy and Acousticness in different genres

We define two types of revolutionaries: artists who make a major change to existing genres and artists who create a new genre. For the first type, we focus on Pop/Rock, which shows a dramatic change in 1950s. For the second type, we put our emphasis on Electronic, which first emerges in early 1950s.

To find out revolutionaries in our network, we first filter Pop/Rock and Electronic tracks released in 1950s, then build a Similarity Bayesian Network based on the artists of these tracks. Intuitively, artists at the core of our network are revolutionaries, as they have a major influence on artists during this period.

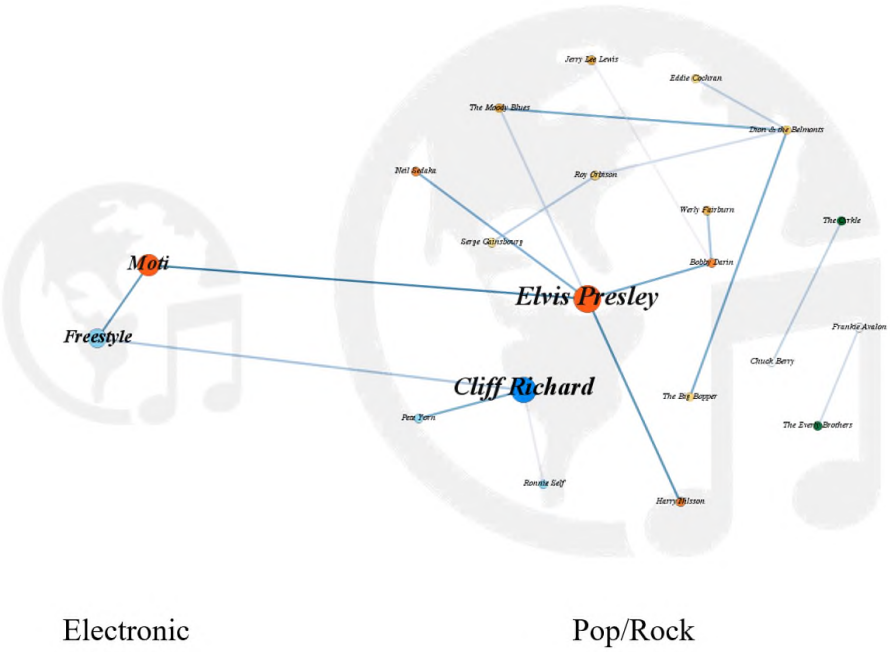


Figure 14: Similarity Bayesian Network of Pop/Rock and Electronic in 1950s

Our Similarity Bayesian Network is shown in figure 14. Elvis Presley, an influencer who has the most followers in 1950s in Pop/Rock, leads a countercultural movement in the United States. At the same time, Cliff Richard, an English artist, makes a dramatic change on the evolution of the genre. These two revolutionaries not only reinvent the existing genre, but also inspire artists like Moti and Freestyle to create a new genre – Electronic. Therefore, **Elvis Presley** and **Cliff Richard** represent revolutionaries in 1950s music revolution.

4.6 Task 6

4.6.1 Dynamic Influencer Indicator

In this task, we choose Pop Rock as the genre for research since it exists for a long time and there are several reinventions.

Intuitively, dynamic influencers dramatically change the trend of the genre in a specific period of time. Thus a dynamic influencer of a period should satisfy

1. He/She released more than 10 tracks in this period.
2. The music characteristics of these tracks are consistent with the lagging trend of Pop/Rock.

Based on the definition above, we construct an indicator of dynamic influencer as

$$d_i = \frac{1}{|A_i|} \sum_{t \in A_i} \left(\sum_{j \in F} \left(x_{i,j,t} - \bar{x}_{j,t-\upsilon} \right)^2 \right)^{1/2} \tag{23}$$

where $x_{i,j,t}$ is the music characteristics j of tracks released by the ith artist in the year t, $\bar{x}_{j,t}$ is the average value of artists, F is the set of music characteristics, υ is the lag year. Here we set $\upsilon = 1$, which means the dynamic influencer will lead the music trend a year after tracks are released.

4.6.2 Solutions

Changes of music characteristics in Pop/Rock are shown in figure 15 Preliminary data analysis shows that most Pop/Rock songs are released after 1956 (the dotted line). Before 1956, the fluctuation of musical characteristics is caused by small sample size. Therefore, we only analyze changes after 1956. We set $F = \{acousticness, danceability, energy, loudness, valence\}$, as only these characteristics show significant fluctuations after 1956.

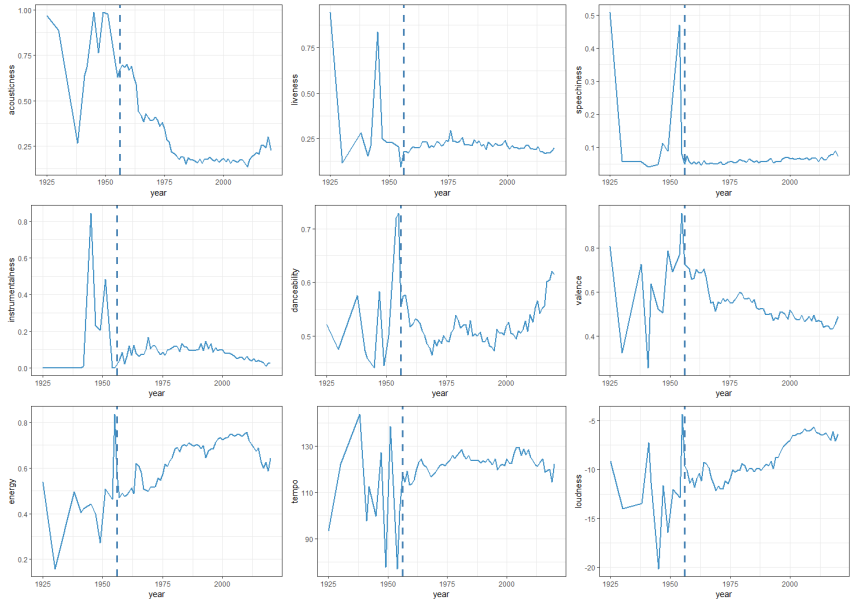


Figure 15: Changes of music characteristics in Pop/Rock

By calculating the value of dynamic influencer indicator of 99 most famous artists, we find out top 10 dynamic Influencers that have giant impact on Pop/Rock. The influencer and his/her active period is shown in figure 16.

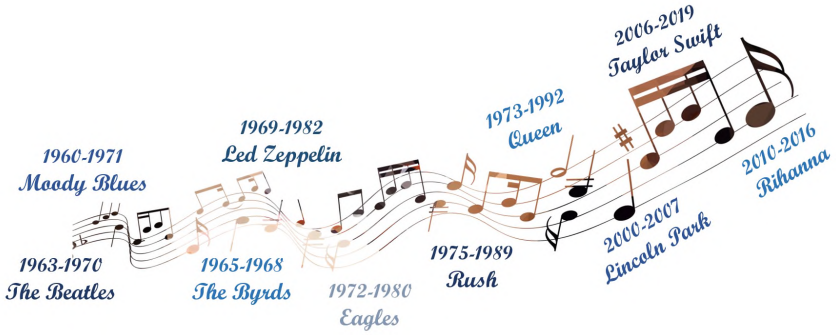


Figure 16: Changes of music characteristics in Pop/Rock

According to the figure 15, Some music characteristics of Pop/Rock have changed a lot over time. Acousticness and valence both experience a significant decline after 1956, while energy and loudness show a consistent increase.

The change above is mainly caused by the change of dynamic influencers. With the technology enhancements and electrical amplification, the Beatles, Led Zeppelin and Eagles poured more intense emotion and passion into Pop Rock music in the last century. These artists tend to express sadness or depressed emotions. However, in recent 20 years, some great musicians, like Lincoln Park and Taylor Swift, switched the tone of Pop Rock music from sadness to happiness. They make pop music smooth, relaxing and much more suitable for dancing.

4.7 Task 7

4.7.1 Cultural Influence of music

Based on the models in the previous section, we detect three important periods in musical evolution, as shown in figure 17. We discuss the culture-influence processes as follows:

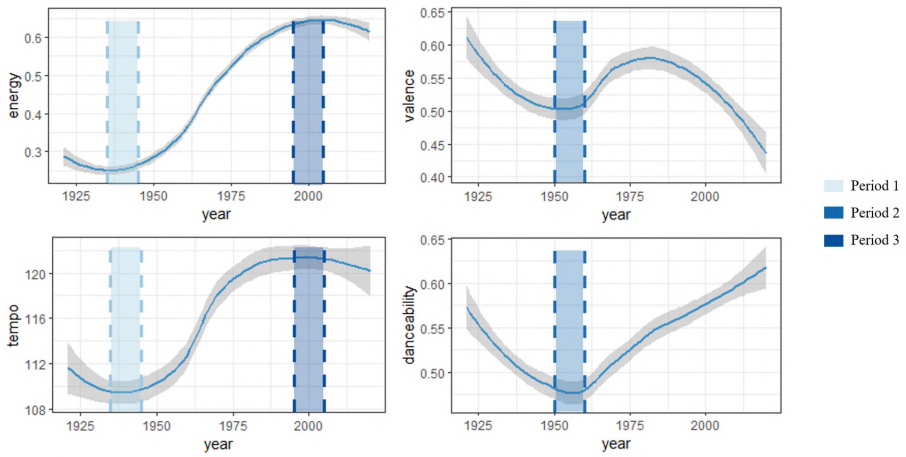


Figure 17: Culture-influence processes

Period I (1935-1945): This Period was shaped by the moods of the Great Depression. In these period, lots of artists begin to create music with high energy and fast tempo, as shown in Figure 17. Reinvention and fast growth of Pop/Rock and the glamorous beginning of Reggae form an upward look of culture.

Period II (1950-1960): Following the detrimental effects of World War II, people was about to embark on a musical journey and substantially change the music style, valence and danceability saw a sharp increase in music released in this period – the golden age of Pop/Rock. The whole culture becomes hopeful and energetic.

Period III (1995-2005): By late 1995, many young people were getting tired of the Pop/rock were inundating the airwaves with, the energy and tempo of music reached the peak and began to decline. A mature music culture is gradually formed.

4.7.2 Changes identified within the network

Social and Political Changes: According to Similarity Bayesian Network of Pop/Rock artists in 1950s. As shown in figure 18, 1950s' Pop/Rock is diversified in several subgenres, which are rather different from each other (as there are no edges between artists in different subgenres). This feature of network identifies the countercultural movement in 1950s – people have increasing desire for true freedom of expression and the diversity of people become more and more prominent.

Technological Changes: According to the Wikipedia, we know that the emergence of the Internet mainly leads to the rise of the Electronic, as the Internet made it easier to spread the software for producing Electronic music. Such trends can be identified through change of numbers of influencers in Electronic over time, as shown in figure 19. Influence reaches its peak due to the proliferation of the Internet.

Analysis above shows that based the models established, we can find the cultural influence of music in time and circumstances, and identify the effects of social, political or technological changes.

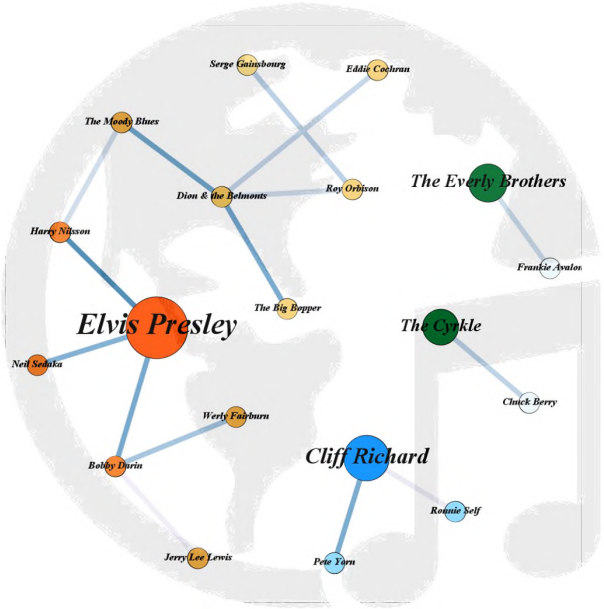


Figure 18: Familiarity Beyesian Network of Pop/Rock in 1950s

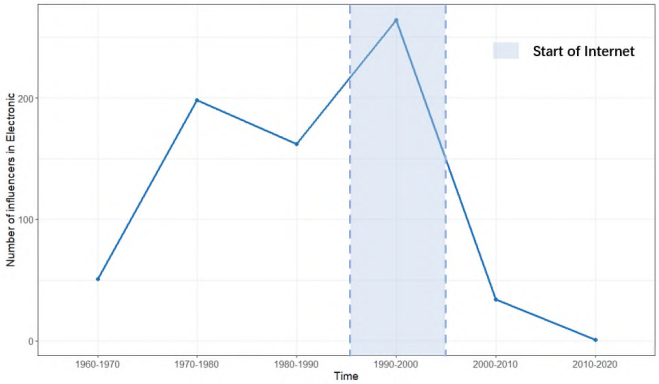


Figure 19: Number of influencers in Electronic

5 Sensitivity Analysis

In Task 5, our Change Point Detection model involves a trimming parameter τ, which might have an effect on optimal change points detection. To test the robustness of the model, we apply a change to the trimming parameter and check whether the revolutionary in 1950s can be well detected.

Table 6: Influence of τ on change point detection

	τ			
Characteristic	[0.05n]	[0.10n]	[0.15n]	[0.20n]
Acousticness	1950 1964 1979	1950 1964 1979	1950 1964 1979	1950 1964 1979
Danceability	1928 1951 1997 2008	1950 1997 2008	1950 1997 2008	1950 2008
Energy	1951 1983 1994 2008	1951 1979 2008	1951 1979 2008	1951 1979 2008
Loudness	1936 1951 1980 2008	1936 1950 2008	1935 1950 2008	1950 2008

As shown in table 6, the proposed model is very robust to change points when τ is greater than or equal to [0.1n], which implies the value we selected is reasonable.

6 Strengths and Weaknesses

6.1 Strengths

- **Effective models**: For different problems, we build several models including Bayesian network, Change Point Detection model, which makes the analysis of each problem detailed and convincing.
- **Statistical tests**: Instead of simply comparing the value, we apply Mann-Whitney Test and Multivariate Mean Test to make our conclusions are reliable.
- **Vivid visualizations**: We use many figures to show our results, make it easier to capture the key information.

6.2 Weaknesses

- Our analysis does not cover all genres in some models. Given space limitations, part of our analysis mainly focuses on two representive genres: Pop Rock and R&B.
- We make a strong assumption on time series, which doen't always hold in reality. We assume the error term follows independent normal distribution, however, it often follows a weakly dependent stationary process in reality.

References

[1] Chen, Guanrong. Introduction to complex networks : Models, structures and dynamics [M]. Higher Education Press, 2012.

[2] Lawley, D. N. "A generalization of Fisher's z test." Biometrika 30.1/2 (1938): 180-187.

[3] Rencher A C , Christensen W F . Methods of Multivariate Analysis[J]. Technometrics, 2002, 38(443):76-77.

[4] McKnight, Patrick E., and Julius Najab. "Mann-Whitney U Test." The Corsini encyclopedia of psychology (2010): 1-1.

[5] Bai J , Perron P . Computation and analysis of multiple structural change models[J]. Journal of Applied Econometrics, 2003.