Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
D		2107714

A Network Model for Music Influence, Similarity and Evolution

Music is developed under the mutual influence of artists from various genres. In the last ninety years, lots of great music has been created. We develop a network model based on some advanced algorithm to excavate the **significant evolutionary and revolutionary trends** of artists and genres. This paper is categorized into the main three-part, focusing on influence, similarity and evolution, for music genres and artists respectively.

Firstly, we process the data provided by the ICM Society. We sort out the number of followers and influencers with respect to artists and the music genres' basic information, which contains 14 main features. Then, we build a network model and analyze the modularity, density and average degree, which could represent the influence level of certain artists. For the network model part, we get the rank of most influential artists from 1930 to 2010. The top 6 artists are: ***The Beatles, Bob Dylan, The Rolling Stones, David Bowie, Led Zeppelin and Jimi Hendrix***. All of them are Pop/Rock artists in 1960. Apart from that, we also get the top 1 most influential artist for each year and each genre separately. We find that the artists in the Pop/Rock genre are most influential and in the genre table, we can also get that the artists in 1960 are most influential, which is consistent with results get from the network diagram. The value of modularity greater than 0.44 indicates a certain degree of modularization had reached by the network diagram. We get 0.7030, 0.6950, 0.6180, 0.5260, 0.6060, 0.6570, 0.7180, 0.8990 and 0.9000 for each network diagram (in order of years from 1930 to 2010), which shows our network diagram is significant.

Next, we expand our model and establish a classification system by using a hierarchical clustering model to analyze the similarities and differences among those music genres. The highest correlation coefficient is 0.8473 for Pop/Rock and Religious. Moreover, the characteristic which could distinguish a genre is **speechiness**, since it has a low correlation with others characteristics. For a certain genre, the attraction must be focused on the **danceability, energy, valence, tempo and loudness**. Furthermore, combining with the network model, we consider the time and analyze the **evolution** processes of musical evolution and **revolution** that occurred over time in one genre.

Then, we could combine our quantitative analysis with the qualitative analysis by studying the literature background such as social, political, cultural and technological factors along with the music genres evolution. For instance, **regional immigration, Bop school, hippie youth movement, people's thoughts on war, economy**, the technological breakthrough of **music synthesizer** promote the development of Country, Jazz, R&B and Pop/Rock, respectively. The key decade is the 1960s when *The Beatles* dominated the music community.

In the end, we write a report to ICM society about how we excavate the important value by our network model to help them understand our recommended further study of music and its effect on culture.

Key Words: Network model; Music Influence; Similarity; PCA; Hierarchical Clustering; Evolution

Contents

1 Introduction

1.1 Background

Nowadays, music is a common thing around us, it can bring different emotions to people. Music is both the voice and memory of individuals and societies, providing a soundscape through which to listen and understand people, groups and their histories.

It is well established that there are many influences to artists when they composing music, including their natural creativity, current social or political events, the usage of a new musical instrument or tool, or other personal experiences. But how to evaluate the different influences of these things? It is necessary to consider the different results of these influences and develop mathematical models and methods to make a quantitative analysis.

1.2 Restatement of the problem

In order to evaluate the music influence among artists and genres, as well as the music revolution. We are required to answer the following questions.

1. Create a (multiple) directed network(s) of musical influence to describe the musical influence and analyze what the musical influence the network(s) reveals.

2. Develop a measure to explore music similarity and find out whether artists within the genre are more similar than artists between genres.

3. Compare similarities and influences between different or the same genres.

4. Create a model to find out that whether the similarity data indicate the identified influencers in fact influence the respective artists.

5. Find out whether there are characteristics in the network that can mark the revolution of music evolution and which artists represent the revolutionaries of the network created.

6. Create a model to analyze the influence process of a genre's musical revolution over time. Indicate what reveal dynamic influencing factors is and explain how genres or artists have changed over time.

7. Show that how does our work expresses the cultural influence of music in time or environment or Explain how to identify the impact of social, political or technological change within a network.

1.3 Overview of model

Firstly, based on the data from the files given by ICM Society, we clean, select and standardize the data. More precisely, we sort out the number of followers and influencers with respect to artists and the music genres' basic information, which contains 14 main features. Secondly, we build a network model and analyze the **modularity, density and average degree**, which could represent the influence level of certain artists. Then by using the **principal component analysis skill**, we shift our focus to some main factors and build a similarity model based on cosine distance. Next, we expand our model and

establish a classification system by using a **hierarchical clustering model** to analyze the similarities and differences among those music genres. Furthermore, combining with the network model, we consider the time and analyze the influence processes of musical evolution that occurred over time in one genre. Finally, after a lot of quantitative analysis, we introduce qualitative analysis techniques to conclude and summarize the model, factors like cultural, social, political or technological changes are taken into consideration. The main framework of our model is showed as Figure 1.

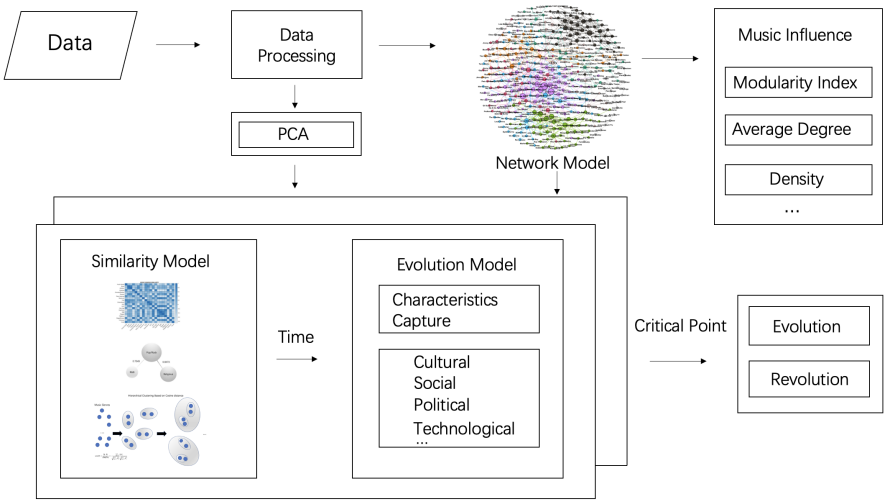


Figure 1: Main Framework of the Model

2 Assumptions and Notations

2.1 Assumptions

In our model, we make the following assumptions.

- The music genre of artists would not change suddenly.
- The main factor with respect to the change of music genre for artists would be the influence of other artists.
- There is no overlap of the genre for a certain song, that is, the classification and features of songs are clear.
- The basic time unit in our model is considered as one year.
- The average level of songs from a certain genre could represent the whole features of the genre.

2.2 Notations

Notations	Interpretation
x_1	Danceability
x_2	Energy
x_3	Valence
x_4	Tempo
x_5	loudness
x_6	Mode
x_7	key
x_8	Acousticness
x_9	Instrumentalness

Table 1: Notations list 1

Notations	Interpretation
x_{10}	Liveness
x_{11}	Speechiness
x_{12}	Explicit
x_{13}	Durations(ms)
x_{14}	Popularity
e_{ii}	Ratio of edges
a_i^2	Ratio of edges
$cut(C_i,\ \bar{C}_i)$	Edges
$vol(C_i)$	Edges

Table 2: Notations list 2

3 Data Processing

The data sets given by ICM provide us lots of information for analysis. The attached file *"influence_data"* contains influencers and followers for 5,854 artists in the last 90 years. The attached file *"full_music_data"* provides 16 variable entries, including musical features such as danceability, tempo, loudness, and key, along with other related information for each of 98,340 songs. However, This is a huge amount of data with lots of redundant and useless data, which could be the main obstacle for later analysis. Namely, data processing by cleaning, selecting and standardizing is needed.

3.1 Data Cleaning

In order to improve the quality of data, we should detecting and removing errors and inconsistencies from data especially when multiple data sources need to be integrated [1]. Hence, the following approaches are addressed to solve the problems.

- Perhaps due to some statistical negligence, places with missing data appear from time to time. Therefore, for variables with a large amount of data missing, we just delete it, which could be explained by the fact that small data cannot provide enough and valuable information for our modeling and analysis. For variables with a small amount of data missing, we use the interpolation method to fill the missing values. **For instance, some data of the genre of music are missed in the file, which causes the problem that the classification of certain songs fails**. Since this data type is not value but text, the small samples of songs are deleted, which is also consistent with our assumptions. This method is widely used for the file *"full_music_data"*.

- Abnormal values mean that if a value in a set of data is more than twice the standard deviation of the average, we call it the abnormal value. Statistically, we can use a box-plot to identify the abnormal values. For the abnormal value, we fix it with the average value of its two adjacent observations. This method is used for summarising the information of the artist, influencer and follower, respectively.

3.2 Data standardization

This step aims to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis. In other words, to avoid the situation that those variables with larger ranges will dominate over those with small ranges, we need to standardize the data. Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.

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

For instance, the feature "duration" in the *"full_music_data"* will dominate the conclusion since it has huge magnitude compared with other features. Therefore, it is really necessary to make sure that the data varies in the same scale by standardizing the data.

3.3 Data Synthesis and Classification

Firstly, applying statistical skills, we find that there are 20 genres in total, which is showed below Table 3 (alphabetical order).

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

Table 3: Music Genres

Secondly, group the songs from the same genre, and calculate the average value, we could get a matrix containing the basic information of the genres and corresponding features as showed in Figure 2.

characteristics / genre1	danceability	energy	valence	tempo	loudness	mode	key	acousticness	instrumentalne	liveness	speechiness	explicit	duration_ms	popularity
Avant-Garde	0.5292	0.2606	0.4611	112.3491	-15.6447	0.6132	5.4774	0.7590	0.1956	0.1676	0.0578	0.0000	212581.2963	23.9794
Blues	0.5692	0.4461	0.6527	117.7901	-11.5145	0.7731	5.0358	0.6187	0.0835	0.2078	0.0661	0.0016	216722.1807	23.2962
Children's	0.6912	0.4466	0.7132	120.3264	-9.5493	0.8393	4.5893	0.6576	0.0092	0.2772	0.0893	0.0000	129730.8036	23.5714
Classical	0.3370	0.1811	0.3265	106.3310	-19.6245	0.7594	4.8848	0.9384	0.5107	0.1946	0.0639	0.0000	288746.0991	12.5774
Comedy/Spoken	0.5569	0.5586	0.4739	105.3215	-14.0186	0.7812	5.1003	0.7811	0.0034	0.6022	0.6483	0.2219	218990.0851	22.6140
Country	0.5861	0.4968	0.6027	119.8932	-10.5729	0.9331	5.3583	0.4663	0.0270	0.1875	0.0462	0.0039	194508.3968	35.8670
Easy Listening	0.4422	0.3452	0.4157	111.5115	-14.4738	0.6990	4.8629	0.7046	0.5347	0.1716	0.0462	0.0000	179973.9078	18.4369
Electronic	0.5939	0.6467	0.4319	119.4566	-9.4331	0.5458	5.5248	0.2247	0.3535	0.1950	0.0772	0.0796	280763.7703	50.8671
Folk	0.5293	0.2815	0.5222	115.7696	-15.0260	0.7967	5.1346	0.7501	0.0719	0.1839	0.0640	0.0007	195383.4037	24.2054
International	0.5628	0.4295	0.6416	116.0415	-11.7279	0.6774	5.1014	0.7013	0.1613	0.2101	0.0936	0.0233	261695.2411	20.7055
Jazz	0.5179	0.3325	0.4926	112.9748	-14.7081	0.5814	4.8801	0.7302	0.3867	0.1883	0.0567	0.0004	306111.8941	21.0431
Latin	0.5985	0.5546	0.6654	119.1537	-9.0300	0.6806	5.1938	0.4798	0.0440	0.1940	0.0579	0.0120	225522.8847	42.5272
New Age	0.3750	0.2073	0.2174	111.1397	-18.7148	0.6863	4.4760	0.8528	0.7106	0.1446	0.0448	0.0000	290956.9594	37.4834
Pop/Rock	0.5097	0.6481	0.5328	123.3224	-9.1830	0.7404	5.1842	0.2534	0.0914	0.2175	0.0593	0.0456	239647.6143	40.9154
R&B	0.6209	0.5474	0.6094	116.5362	-9.6816	0.6240	5.3247	0.3552	0.0341	0.1933	0.0764	0.1001	247040.9086	39.8331
Reggae	0.7068	0.5203	0.7164	116.3342	-10.3048	0.6411	5.6502	0.2855	0.1018	0.1948	0.1352	0.0246	241471.2674	36.7361
Religious	0.5387	0.6005	0.4632	120.5381	-8.0322	0.8113	5.2000	0.3436	0.0128	0.2477	0.0635	0.0000	264423.0127	41.7620
Stage & Screen	0.3095	0.2550	0.2262	104.0281	-16.8539	0.7051	4.8327	0.7609	0.4917	0.1834	0.0657	0.0000	210351.0558	29.8223
Unknown	0.6085	0.6722	0.7308	120.7925	-8.5981	1.0000	6.0769	0.1665	0.0066	0.1524	0.0327	0.0000	207280.0000	41.6154
Vocal	0.4563	0.2575	0.4267	110.0456	-13.6578	0.7144	5.2360	0.8182	0.0225	0.2126	0.0593	0.0000	191296.3266	20.8524

Figure 2: Data classification of the genres and features (Average Level)

Thirdly, standardizing the data by using the formula in 3.2, we could get the below Figure 3.

characteristics / genre	danceability	energy	valence	tempo	loudness	mode	key	acousticness	instrumentalne	liveness	speechiness	explicit	duration_ms	popularity
Avant-Garde	-0.0261	-1.0931	-0.3613	-0.4732	-0.8961	-1.0400	0.8756	0.7427	0.0132	-0.5096	-0.2832	-0.4741	-0.4023	-0.6124
Blues	0.3538	0.0735	0.8968	0.5044	0.2874	0.3819	-0.3282	0.1527	-0.4944	-0.0893	-0.2207	-0.4447	-0.3075	-0.6772
Children's	1.5125	0.0764	1.2936	0.9602	0.8505	0.9705	-1.5456	0.3162	-0.8311	0.6377	-0.0444	-0.4741	-2.2984	-0.6511
Classical	-1.8526	-1.5933	-1.2449	-1.5546	-2.0365	0.2598	-0.7400	1.4974	1.4413	-0.2269	-0.2371	-0.4741	1.3408	-1.6939
Comedy/Spoken	0.2367	0.7813	-0.2770	-1.7360	-0.4301	0.4537	-0.1524	0.8359	-0.8575	4.0382	4.1904	3.6210	-0.2556	-0.7419
Country	0.5145	0.3927	0.5684	0.8823	0.5572	1.8046	0.5510	-0.4885	-0.7508	-0.3013	-0.3716	-0.4013	-0.8159	0.5152
Easy Listening	-0.8531	-0.5608	-0.6593	-0.6237	-0.5606	-0.2766	-0.7998	0.5141	1.5499	-0.4682	-0.3711	-0.4741	-1.1485	-1.1381
Electronic	0.5878	1.3353	-0.5532	0.8039	0.8838	-1.6390	1.0048	-1.5044	0.7287	-0.2230	-0.1366	0.9946	1.1581	1.9380
Folk	-0.0257	-0.9617	0.0396	0.1414	-0.7188	0.5919	-0.0590	0.7055	-0.5471	-0.3390	-0.2364	-0.4609	-0.7959	-0.5910
International	0.2932	-0.0310	0.8240	0.1902	0.2262	-0.4689	-0.1495	0.5003	-0.1419	-0.0646	-0.0121	-0.0443	0.7217	-0.9229
Jazz	-0.1335	-0.6409	-0.1546	-0.3608	-0.6277	-1.3225	-0.7526	0.6217	0.8793	-0.2934	-0.2918	-0.4661	1.7382	-0.8909
Latin	0.6315	0.7562	0.9799	0.7495	0.9993	-0.4407	0.1026	-0.4316	-0.6738	-0.2336	-0.2829	-0.2530	-0.1061	1.1470
New Age	-1.4912	-1.4286	-1.9609	-0.6905	-1.7758	-0.3893	-1.8544	1.1373	2.3467	-0.7506	-0.3818	-0.4741	1.3914	0.6685
Pop/Rock	-0.2120	1.3442	0.1096	1.4985	0.9555	0.0910	0.0763	-1.3838	-0.4586	0.0120	-0.2722	0.3677	0.2171	0.9941
R&B	0.8451	0.7106	0.6123	0.2791	0.8126	-0.9438	0.4593	-0.9557	-0.7185	-0.2404	-0.1425	1.3733	0.3863	0.8914
Reggae	1.6606	0.5405	1.3150	0.2428	0.6340	-0.7916	1.3469	-1.2488	-0.4118	-0.2248	0.3030	-0.0195	0.2589	0.5976
Religious	0.0641	1.0446	-0.3473	0.9982	1.2852	0.7214	0.1194	-1.0044	-0.8150	0.3287	-0.2403	-0.4741	0.7841	1.0744
Stage & Screen	-2.1135	-1.1285	-1.9034	-1.9684	-1.2426	-0.2225	-0.8820	0.7509	1.3551	-0.3439	-0.2234	-0.4741	-0.4533	-0.0582
Unknown	0.7273	1.4955	1.4097	1.0439	1.1231	2.3995	2.5102	-1.7492	-0.8431	-0.6693	-0.4735	-0.4741	-0.5236	1.0605
Vocal	-0.7192	-1.1131	-0.5872	-0.8872	-0.3268	-0.1395	0.2176	0.9918	-0.7708	-0.0387	-0.2718	-0.4741	-0.8894	-0.9090

Figure 3: Standardized Data

The above steps are significant for later analysis, the process of model development could be divided into three categories, which is Network Construction, the Similarity Model of music and Evolution.

4 Network Construction

4.1 Evaluation Index

We use Modularity[2] and Conductance[3] unsupervised evaluation indicators to evaluate the merits and demerit of communities in network.

- Modularity

  Modularity is used to evaluate the ratio between communities' edges to network's total edges. Defined as below:

$$Q = \sum_{i=1}^{k} \left(e_{ii} - a_i^2 \right) \tag{2}$$

  e_{ii} refers to the ratio of i^{th} community's inner edges to network's total edges and a_i^2 refers to the ratio of the number of linked edges between community i and other communities to the network's total edges. According to (2), $Q = [-0.5, 1]$. The larger the Q, the more obvious communities can be found in the network.

- Conductance

  Conductance is used to evaluate the ratio between number of linked edges between communities to community's inner edges. Defined as below:

$$Conductance(C_i) = \frac{cut(C_i, \bar{C}_i)}{minvol(C_i), vol(\bar{C}_i)} \tag{3}$$

  $cut(C_i, \bar{C}_i)$ refers to the edges between communities and $vol(C_i)$ refers to the edges in a community. According to (3), $Conductance(C_i)$ = (0,1]. The smaller the $Conductance(C_i)$, the more connected in the community, the less connected between communities, the better the result of community division.

4.2 Diagram and network model

By applying diagram and network model, we get the number of influencers and followers of all artists as well as the genre of that artist. The top 6 most influential artists are shown in Table 4: (e.g. *The Beatles* has 615 followers and influenced by other 31 artists, is the most influential artist).

Artist Name	Follower	Influencer	Genre	Year
The Beatles	615	31	Pop/Rock	1960
Bob Dylan	389	29	Pop/Rock	1960
The Rolling Stones	319	39	Pop/Rock	1960
David Bowie	238	25	Pop/Rock	1960
Led Zeppelin	221	24	Pop/Rock	1960
Jimi Hendrix	201	32	Pop/Rock	1960
...	...	...	...	...

Table 4: Top 6 Most Influential Artists

Table 4 shows that the top 6 most influential artists are All of them are Pop/Rock artists in 1960.

Apart from that, we can also get the top 1 most influential artist for each year and each genre separately:

Year	Artist Name	Follower	Genre
1930	Hank Williams	184	Country
1940	Miles Davis	160	Jazz
1950	Marvin Gaye	169	R&B
1960	The Beatles	615	Pop/Rock
1970	Sex Pistols	153	Pop/Rock
1980	Metallica	112	Pop/Rock
1990	Radiohead	55	Pop/Rock
2000	Avril Lavigne	14	Pop/Rock
2010	Frank Ocean	3	R&B

Table 5: Top 1 Most Influential Artist Each Year

Analysis from Table 5, the artists in the Pop/Rock genre are consistently most influential from 1960 to 2000. Consistent with the trend shown in Table 4.

Genre	Artist Name	Follower	Year
Comedy/Spoken	Spike Jones	20	1930
Country	Hank Williams	184	1930
Folk	Pete Seeger	47	1930
Vocal	Billie Holiday	106	1930
Blues	Muddy Waters	113	1940
Easy Listening	Henry Mancini	19	1940
Jazz	Miles Davis	160	1940
Stage & Screen	Ennio Morricone	30	1950
Religious	James Cleveland	17	1950
R&B	Marvin Gaye	169	1950
Latin	AntÃ´nio Carlos Jobim	34	1950
International	JoÃ£o Gilberto	27	1950
Children's	Alvin & the Chipmunks	3	1950
New Age	Mike Oldfield	16	1960
Pop/Rock	The Beatles	615	1960
Reggae	Lee "Scratch" Perry	48	1960
Avant-Garde	Terry Riley	34	1960
Classical	Steve Reich	24	1960
Electronic	Kraftwerk	108	1970
Unknown	The Wonder Stuff	1	1980

Table 6: Top 1 Most Influential Artist Each Genre

Analysis from Table 6, the artists with most influence focus on 1930 to 1960. In 1950, there are most kinds of influential genres and in 1960, the number of followers is the largest.

4.3 Network Diagram

- Data Used **Edge file:**

  In the edge file, there are two main data set: source and target. In each edge file, we use all influencers in a certain year as the source and all followers of the source's artists as the target.

  Node file:

  In the node file, there is only one used data set: id. In each node file, we use non_redundant influencers in the corresponding edge file as id.

- Result

  Average Degree: Represent the average of the connecting edges of each node.

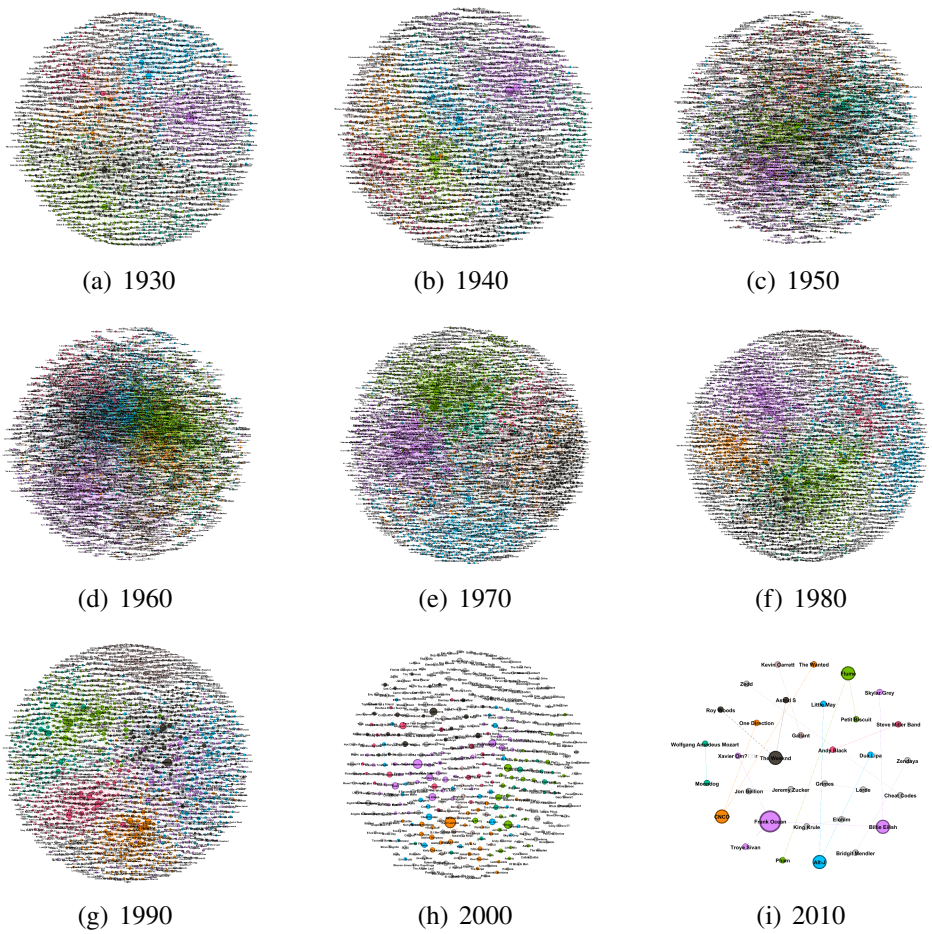
  Density: Represent the density of the diagram.

  Modularity: Measure the degree of modularization of the network diagram. A value greater than 0.44 indicates a certain degree of modularization had reached by the network diagram.

  By using the data mentioned before, we can get the following measured parameters and network images:

Year	Average Degree	Density	Modularity
1930	1.8470	0.0010	0.7030
1940	2.0410	0.0010	0.6950
1950	2.6350	0.0010	0.6180
1960	3.811	0.0010	0.5260
1970	2.9880	0.0010	0.6060
1980	2.7810	0.0010	0.6570
1990	2.3060	0.0010	0.7180
2000	0.9240	0.0020	0.8990
2010	0.6060	0.0190	0.9000

Table 7: Directed Networks By Year



(a) 1930 (b) 1940 (c) 1950

(d) 1960 (e) 1970 (f) 1980

(g) 1990 (h) 2000 (i) 2010

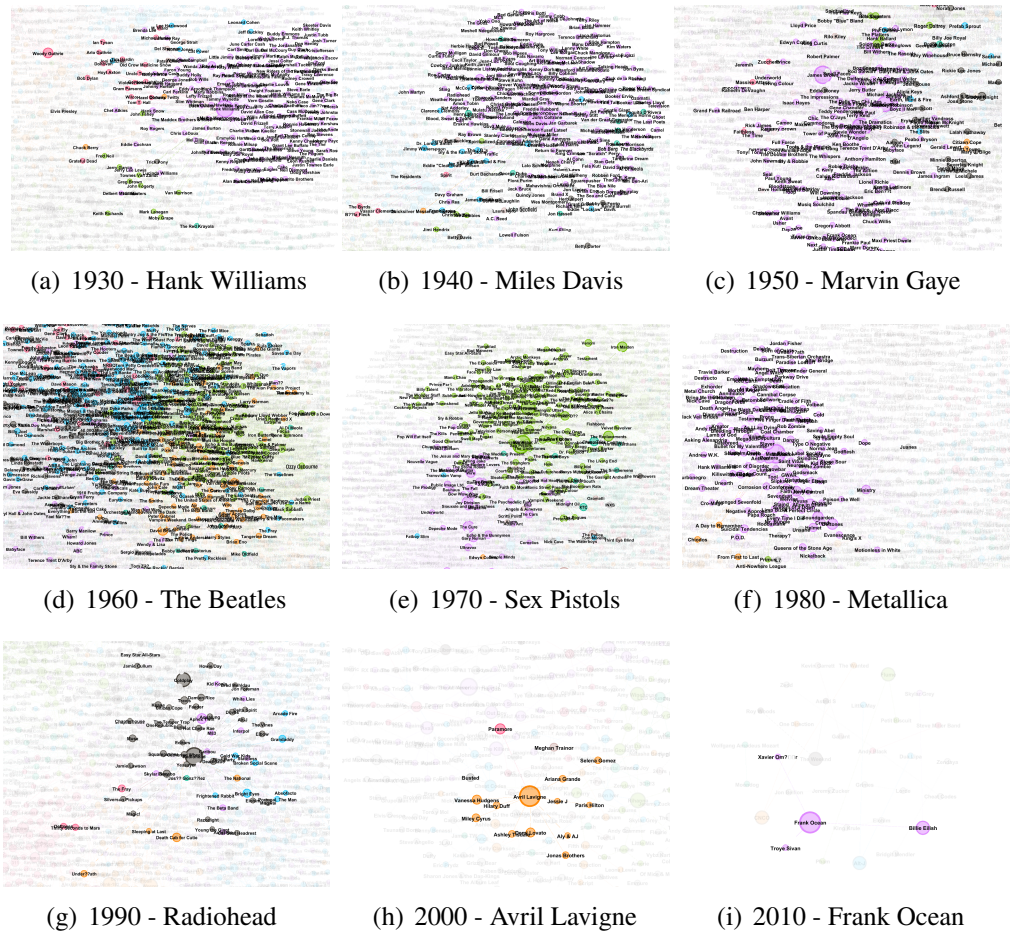
Figure 4: Network Images By Year

number of edges: Represent "music influence".

nodes: Represent artists.

Figure 4 shows "music influence" among all artists in each 10 year from 1930 to 2010, nodes are artists and edges between nods are “music influence”. We could find the artist has most complicated edges (music influence) for other years, all detailed images are shown in Figure 5.

Taking year 1960 as an example, *"The Beatles"* is the most influential artist. As shown in Figure 5 (d), the biggest blue circle represents *“The Beatles”*. It has the most complicated edges (music influence) and influences a large number of other nodes (artists), which means *"The Beatles"* has the most mutual influencers (influencers + followers).



(a) 1930 - Hank Williams (b) 1940 - Miles Davis (c) 1950 - Marvin Gaye

(d) 1960 - The Beatles (e) 1970 - Sex Pistols (f) 1980 - Metallica

(g) 1990 - Radiohead (h) 2000 - Avril Lavigne (i) 2010 - Frank Ocean

Figure 5: Detail Network Images

The layout of all images in Figure 5 accord with the result shown in Table 5. For instance, in 1960s, *"The Beatles"* dominates the music community and it is easy to remind us of the background of that period. Based on this idea, we really need to continue our study on the characteristics of each music genre so that the conclusion will be more precise. Hence, the next section mainly focuses on the characteristics of genre based on the similarity model.

5 The Similarity Model of Music

5.1 PCA analysis

Since the data has so many dimensions, we introduce the Principal Components Analysis (PCA) to reduce the dimensions.

- Standardize the data
  We have 14 feature, $x_1x_2...x_{14}$, representing the danceability, energy...popularity, respectively, and 20 music genres in total. Once the standardization is done, all the variables will be transformed to the same scale. Let a_{ij} be the entry of the data matrix. Standardize the data from a_{ij} to $\hat{a}_{ij}$

$$\hat{a}_{ij} = \frac{a_{ij} - \mu}{s_j} \tag{4}$$

where $i = 1, 2, ...20$; $j = 1, 2, ..., 14$

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} a_{ij}; s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_{ij} - \mu_j)^2} \tag{5}$$

- Calculate the correlation coefficient matrix $R = (r_{ij})_{14\times14}$, $r_{ij} = \frac{\sum_{k=1}^{n} a_{ki}a_{kj}}{n-1}, i, j = 1, 2, ...14, r_{ii} = 1$ which is visualized as Figure 5.1.

- Calculate the eigenvalues λ and eigenvector u.

- Choose the principal components
  By calculating the contribution of each component, we could figure out the contribution rate and cumulative contribution rate.

$$b_j = \frac{\lambda_j}{\sum_{k=1}^{14}\lambda_k}, j = 1, 2, ..., 14 \tag{6}$$

$$\alpha_p = \frac{\sum_{k=1}^{p}\lambda_k}{\sum_{k=1}^{14}\lambda_k} \tag{7}$$

where p is the number of components we choose, which is shown in Figure 7.

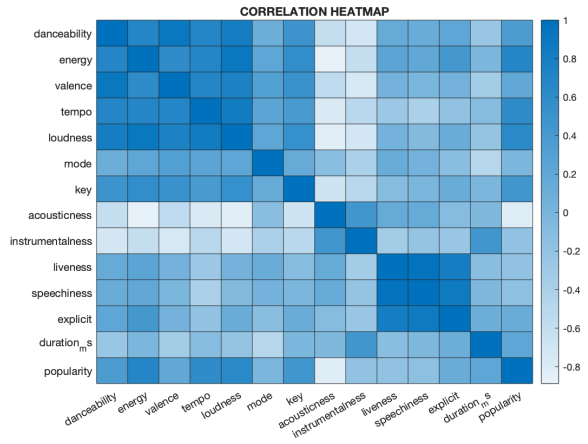


Figure 6: correlation matrix

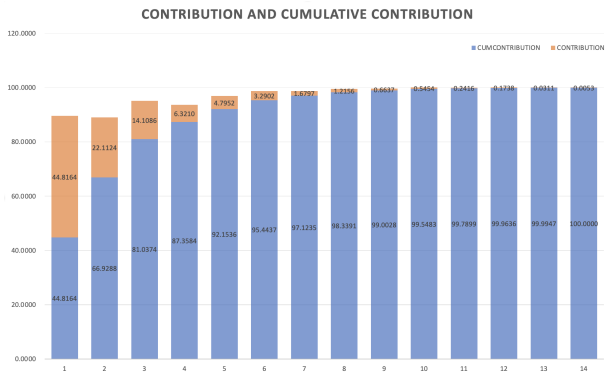


Figure 7: Contribution rate and Cumulative contribution rate

- Calculate the scores

$$Z = \sum_{j=1}^{p} b_j y_j \tag{8}$$

Result of PCA: As the correlation map shows, there exists a very closed relationship between some factors. If we use these factors directly, it will cause the overlap of information, which could also affect the final result. By PCA, we could reduce the dimension of the data and analyze the data more precisely. As the cumulative contribution rate shows, we choose the first 4 components, that is the $\alpha_p = 0.85$.

Eigenvalues	Contribution Rate	Cumulative Contribution Rate
6.2743	44.8164	44.8164
3.0957	22.1124	66.9288
1.9752	14.1086	81.0374
0.8849	6.3210	87.3584

Table 8: The First 4 Components

x_1	x_2	x_3	x_4	...	x_{14}
0.3423	0.0738	-0.0815	-0.4312	..	0.2453
0.3626	0.0910	0.1955	0.1568	...	0.6484
0.3277	0.0099	-0.2514	-0.3675	...	-0.2808
0.3316	-0.2278	-0.0318	-0.0623	...	-0.1457

Table 9: The Corresponding Eigenvectors (Excerpt)

By calculating the scores, we could get the score which is showed in table 10.

$$Z = 44.8164y_1 + 22.1124y_2 + 14.1086y_3 + 6.3210y_4 \tag{9}$$

Unknown	0.9470
Comedy/Spoken	0.8314
Electronic	0.8229
Pop/Rock	0.7999
R&B	0.7927
Reggae	0.7881
Religious	0.7600

Table 10: The scores

With the top scores of PCA, the corresponding genres are unknown, Comedy/Spoken, electronic, Pop/Rock, R&B, Reggae and Religious, next, we will mainly explore these genres. Once the PCA is done, the complexity of the data has been reduced and we could consider the features which share closed relations so that the similarity model could be more precise. PCA could increase the efficiency of the data given the processes taking place in a smaller dimension [4]. Then, the next section is about how we use the cosine distance as a measurement for the similarity among music genres after PCA.

5.2 Measurement of Similarity

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space, which could be explained by the following formula:

$$\text{Similarity} \ = \cos(\theta) = \frac{\mathbf{X} \cdot \mathbf{Y}}{||\mathbf{X}||||\mathbf{Y}||} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2} \sqrt{\sum_{i=1}^{n} Y_i^2}} \tag{10}$$

where X and Y are two non-zero vectors. Graphically, the cosine similarity measures the direction of two angles in the inner space, which is a judgment of orientation and not magnitude. We introduce this measurement to our model to explore the similarity among music genres. The relative data is Table 3, which is the standardized data. After the calculation, we could visualize the similarity by using the heat map, which is the Figure 5.2.

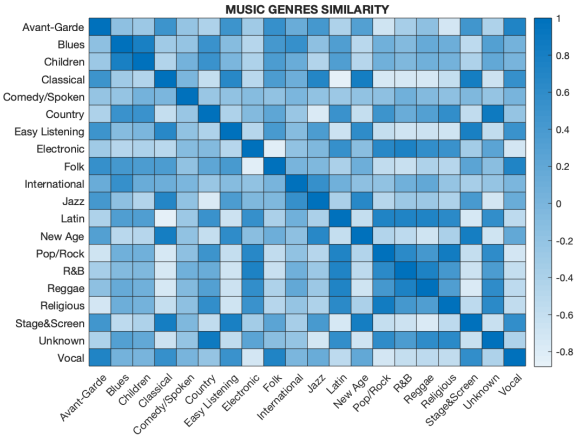


Figure 8: Music Genres Similarity

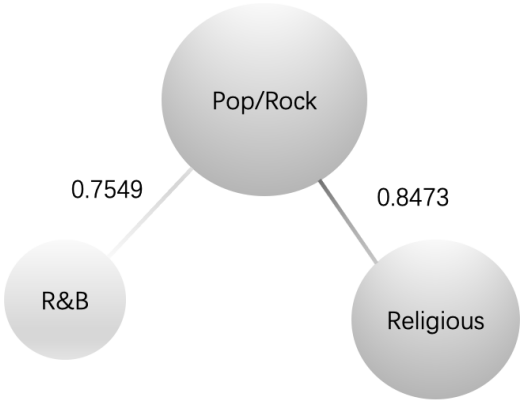


Figure 9: Highest Coefficient

It is obvious that Pop/Rock, R&B, Reggae and Religious are concentrated, which means that those music genres share a very close similarity. Numerically, the similarity between the R&B and Religious is 0.7549 and the similarity between the Pop/Rock and religious is 0.8473. However, the highest similarity lies between the country and unknown up to 0.8753. Since we could not figure out what genres the unknown music would be, it might be inconclusive for us until we consider the timeline for further study.

The cosine distance gives us a very good measurement for clustering. The next section will present how we use the cosine similarity to clustering the music genres.

5.3 Hierarchical Clustering

The hierarchical clustering calculates the similarity between two types of data points, combines the two most similar data points among all data points, and iterates the process repeatedly. To put it simply, hierarchical clustering is to determine the similarity between the data points of each category by calculating the distance between them and all the data points. The smaller the distance, the higher the similarity. The two nearest data points or categories are combined to generate a cluster tree.

The below graph Figure 5.3 shows the basic rationale of hierarchical clustering based on cosine similarity.

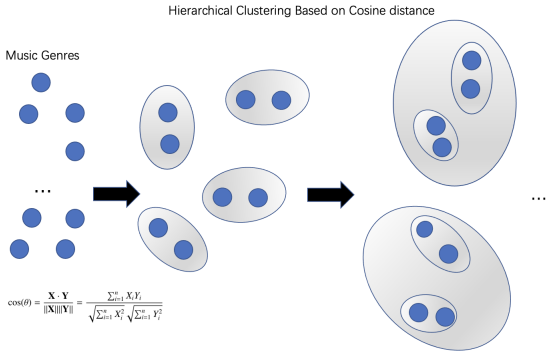


Figure 10: Rationale of Hierarchical Clustering

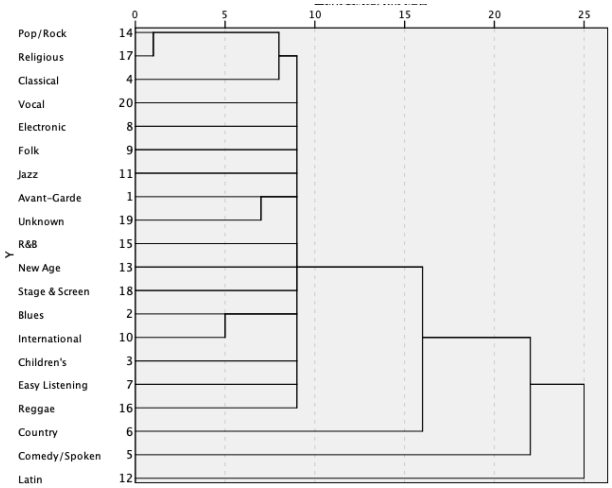


Figure 11: Hierarchical Clustering Result

The result from Figure 11 shows that Pop/Rock is has a closed relationship with the Religious, with distance ranging from 0 to 1, while the Latin music lies far away from other music genres, with relative distance up to 25, which is consistent with our common sense. However, in order to explore the music quantitatively without being affected by the subjective sense, we need to take the time into consideration. And the next section will explore that how artists or their corresponding genres change over time.

6 Evolution of the Music Genres

6.1 Changes in the influencers and followers

To examine the evolutionary and revolutionary trends of artists and genres, we do visualizations using line charts.

Through the processing of the data in the *influence_data* data set, by calculating the number of influencers and followers in each genre and sorting by time, the music influence changing with time can be observed and determined the peak years in different fields based on the greatest number of new talents.

	Influencer																		
year	Avant-Garde	Blues	Children's	Classical	Comedy/Spoken	Country	Easy Listening	Electronic	Folk	International	Jazz	Latin	New Age	Pop/Rock	R&B;	Reggae	Religious	Stage & Screen	Unknown	Vocal
1930	0	17	0	3	2	13	3	0	6	2	43	11	0	1	1	0	3	8	0	37
1940	0	19	0	4	1	31	6	0	7	8	75	17	0	2	16	1	3	1	0	39
1950	1	30	1	2	5	47	6	0	21	8	82	29	0	99	99	3	2	10	0	37
1960	3	13	0	5	12	43	1	4	30	18	59	37	3	372	116	36	7	3	0	9
1970	2	4	0	1	3	52	0	9	9	11	23	17	10	364	82	36	6	6	1	10
1980	0	1	0	1	6	31	0	31	6	5	10	23	7	483	54	19	22	1	1	4
1990	1	1	0	1	3	49	1	61	3	3	3	27	1	370	68	12	8	0	0	1
2000	1	0	0	0	0	8	0	22	0	1	0	3	0	110	33	4	3	0	0	0
2010	0	0	0	1	0	0	0	1	0	0	0	0	0	10	3	0	0	0	0	0

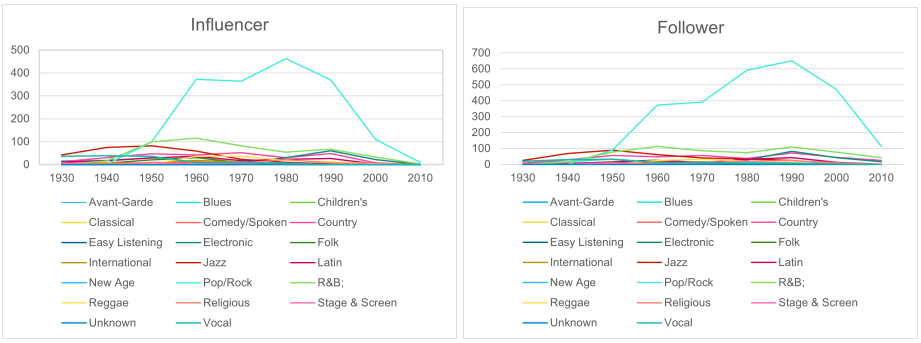
Figure 12: Number of influencers in different genres each year

Figure 12 marks the greatest number of influencers. Each genre has one or two peak years. From the data in the above figure, we get the line chart in the left side below through analysis.

Follower																				
year	Avant-Garde	Blues	Children's	Classical	Comedy/Spoken	Country	Easy Listening	Electronic	Folk	International	Jazz	Latin	New Age	Pop/Rock	R&B;	Reggae	Religious	Stage & Screen	Unknown	Vocal
1930	0	10	0	0	1	8	0	0	1	1	25	1	0	1	1	0	1	0	0	21
1940	1	17	0	2	1	29	5	0	4	2	68	6	0	2	15	0	0	1	0	31
1950	1	34	2	2	1	56	5	0	17	5	89	15	0	90	77	2	2	11	0	32
1960	3	13	0	6	10	48	3	3	28	17	62	26	2	373	113	26	7	2	0	11
1970	2	8	1	1	3	57	1	9	12	16	41	10	11	392	86	32	11	10	1	14
1980	0	4	0	1	8	38	0	29	7	11	34	32	11	592	74	21	21	4	2	6
1990	1	2	1	4	8	72	1	82	8	9	26	43	6	650	110	23	24	5	0	7
2000	1	0	0	0	5	45	0	43	2	3	7	13	6	471	78	10	10	5	0	3
2010	0	0	0	0	0	25	1	16	0	0	0	1	0	113	43	0	1	0	0	0

Figure 13: Number of follower in different genres each year

Figure 13 marks the greatest number of followers which indicates the new generation and flourish of each genre. From the data in the above figure, we get the line chart below in the right side.



(a) Line chart for influencers (b) Line chart for followers

Figure 14: Line chart for influencers and followers in different genres each year

From Figure 14, it can be seen that each field's influencer and follower have a rise and fall time from 1930 to 2010. The change in the number of influencers will influence the variation of the follower. The influencer and follower almost change synchronously after comparing two pictures in Figure 14.

The slopes in Figure 14 indicate the developing speed of each field with time. When one field becomes hot, there are always some other fields fading. Most of the fields only spring up one or two decades. This may because many artists change their genres frequently with the stream. For those genres that have similarities or connections like Pop and R&B, they could have similar change trends. In the figures above, they are kept in a relatively high number from 1960 to 1990.

6.2 Changes in the full songs

Combine with the *full_music_data* data set, dividing different types of songs in their period and counting the total number, Figure 15 gives the total amount of songs and Figure 16 shows the trend of each genre, which almost meets the changing trend above.

Full songs																				
year	Avant-Garde	Blues	Children's	Classical	Comedy/Spoken	Country	Easy Listening	Electronic	Folk	International	Jazz	Latin	New Age	Pop/Rock	R&B;	Reggae	Religious	Stage & Screen	Unknown	Vocal
1920	1	6	0	8	0	0	0	0	0	0	9	0	0	1	2	0	0	1	0	9
1930	3	65	0	144	0	51	0	0	0	17	138	1	0	9	2	42	0	9	0	669
1940	41	150	0	461	3	64	22	12	18	204	216	123	16	91	7	69	1	157	0	546
1950	75	185	0	1072	62	663	390	10	273	460	3136	307	0	425	393	24	41	121	0	2486
1960	58	461	34	315	104	1242	353	5	644	183	2080	530	1	6017	1807	183	16	154	0	1700
1970	6	177	7	81	32	1224	22	86	290	159	561	503	12	10897	2347	422	15	68	1	257
1980	30	94	0	136	67	1100	14	62	49	233	404	832	97	10508	1421	260	83	138	0	228
1990	15	82	3	84	28	1217	6	347	70	124	259	1191	90	7336	1608	169	148	159	12	215
2000	10	34	10	40	25	1176	5	324	45	72	119	718	50	7582	1594	205	228	203	0	125
2010	4	2	2	11	8	871	12	486	8	8	20	496	5	4647	1159	47	178	48	0	45

Figure 15: Full songs in different genres each year

Figure 16 also shows that when one field becomes popular, other fields will have fewer songs, which has similarities with the decreasing number of artists. This finally leads to the decline of one genre.

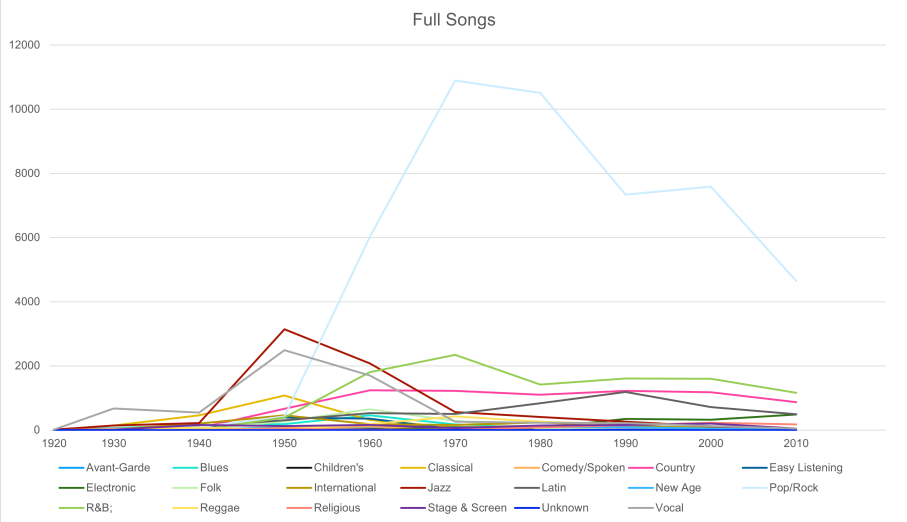


Figure 16: Number of songs in different genres each year

6.3 Changes in Pop/Rock Over time

Looking at the certain genre, we chose the distinct one, which is Pop/Rock. The Pop/Rock music has a sharp rising time of artists around 1950 to 1990 with a peak in 1990 and falls down after that. From Table 4, we know that *The Beatles* has the most followers in Pop/Rock and start activation in 1960. From the data set, they released 120 songs in 1967, which is the year with the most releases and their most popular song is released in 1969. According to Table 5, between 1960 to 2000, the most influential artists belong to Pop/Rock. This may lead to the rising of Pop/Rock during that time.

Using the *full_music_data* data set, after standardizing all data into 0 to 1 scale, we plot the change of all characteristics with time in Figure 17.

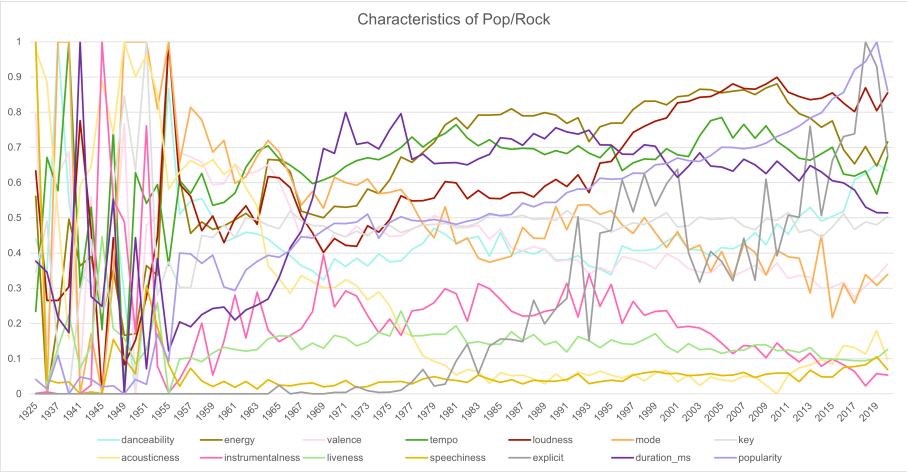
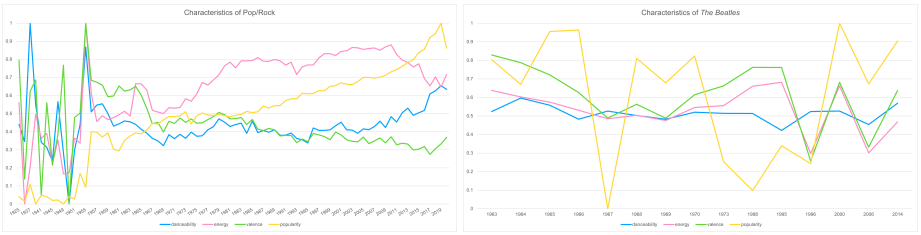


Figure 17: Characteristics of Pop/Rock

It can be seen that the general tendency of popularity is increasing. This indicates the rising position of Pop/Rock in history. Although the number of the full song has decreased, it still more than other genres, which meets the previous result.



(a) Line chart for Pop/Rock (b) Line chart for The Beatles

Figure 18: The characteristics between one genre and artist

For Pop/Rock music, we further focus on **danceability, energy, valence and popularity** in the whole field and the most influential artists with the time change. From the figure above, we find that the change of the single artist is not the same with the whole genre.

Every musician might have a trough period and popular season, while the whole area may only have a small variation due to their changes. The characteristics of *The Beatles* keep almost the same from it began. After the 1960s, Pop/Rock's energy slowly grows until 2010. Then it falls down. The danceability and valence tend to be stable during this period. However, every genre will experience the boom goes to the bust, then slowly from the bust to the second boom with evolutions over time.

7 Strengths and weaknesses

7.1 Strengths

- Our model can be widely used in engineering, management, social science and other related fields, especially interdisciplinary frontier research.
- It is an innovative use of the features of different music genres when analyzing the similarity.
- The value of modularity greater than 0.44 indicates a certain degree of modularization had reached by the network diagram. We get 0.7030, 0.6950, 0.6180, 0.5260, 0.6060, 0.6570, 0.7180, 0.8990 and 0.9000 for each network diagram (in order of years from 1930 to 2010), which shows our network diagram is significant.
- This model is especially suitable for mining a large amount of information from a complex system. Multidimensional problems would be solved by our model as well.
- Valuable information could be found by analyzing the time evolution along with the network model.

7.2 Weaknesses

- There are many ways to measure the similarity, perhaps the cosine distance is not the best choice since it only justifies the direction but not the magnitude.
- Singular values can make a big difference for hierarchical clustering and the data are likely to be clustered in chains by mistake.
- The assumption that the average level of features could represent the whole level of a music genre might fail so that our model precision could be affected.
- Perhaps the time unit (year) is too long since the technological progression would make a difference in the short term.

8 Model Extension

Our model could be applied to many areas, such as social media networks, online shopping investigation and linguistics. Let's take social media as an example. It is widely known that social media has a basic function showing the following and followers, which is almost the same as the followers and influencers in our music model. Based on this, we could build a network model to capture **how valuable information transfers among users, which could bring our attention to the scenes of major events around the world**.

9 Conclusions

- The parameter "number of edges" indicates 'music influence' in the network. Different artists have a different number of edges, the more number of edges an artist has, the more influential the artist is. The "music influence" reveals that an artist's composition is influenced by another or several artists' compositions. In another word, artists' compositions are influenced by each other. For instance, the most influential artist *"The Beatles"* from 1930 to 2010. They have the most mutual influencers (influencers + followers), so the node representing ***"The Beatles"*** has the most complicated edges.
- The **Pop/Rock, R&B, and Religious** share a very closed relationship with each other, and the coefficient between the Pop/Rock and Religious is up to 0.8473, which is a very high level.
- The characteristic which could distinguish a genre is **speechiness**. Songs that have a large value in speechiness could be more likely the genre of Comedy/Spoken. Pop/Rock always has a higher value in instrumentalness and energy. And the genre's popularity changes over time due to their characteristics' change and the influencer's and follower's variation in different fields.
- It is true that the 'influencers' actually affect the music created by the followers. For example, ***David Bowie***, who is a follower of *"The Beatles"* in the Pop/Rock region. It has similar characteristics in instrumentalnes and danceability. Moreover, it is also obvious that some music characteristics are more 'contagious' than others, such as danceability, energy, valence, tempo

and loudness. For example, Pop/Rock has a larger value in these characteristics compared with some soft music like Classical.

- The characteristics that might signify revolutions in musical evolution from these data are energy, valence and popularity. The changes in the first two may indicate the mood of the artists during that stage in one genre. For example, when in the rising time, there is always more energy in the music which can inspire the audience. While popularity reflects the real status of one genre in the whole field. A decreasing value in popularity reveals the decline of a genre. Moreover, artists like *The Beatles* could represent revolutionaries in our network model.

- We analyze the influence processes of musical evolution that occurred over time for Pop/Rock. The speechiness and instrumentalness remain steady but the mode and acousticness of Pop/Rock music change a lot in the past decades. The popularity of one genre can be one of the most direct indicators that reveal the dynamic influencers. Take an artist for example, *"The Beatles"*, who has the most follower in Pop/Rock history, decrease their energy, valence and loudness sharply in 1996. That is the period when Pop/Rock is not so popular. After that, with music diversified development, they increased these values to create more energetic music. Besides, their popularity continuously increasing with the time change. For all genres, they all have their rising and falling seasons over time with other new genre appears in the evolution.

- In the 1930s, country music enjoyed high popularity in America. Most country music tunes were simple and smooth, which was a combination of traditional American folk, Gospel, bluegrass and other forms of music and had a profound impact on the development of later music. **Regional immigration** played an important role. During that period, many Appalachians came to Atlanta to work in the factories, and with their music. It was this time that country music became mainstream music. Jazz music is a beautiful line in the 1940s, and the influence remains till now. And the greatest revolutionary of Jazz is the foundation of **Bop school**, which insisted that jazz should be improvised essentially. Based on our model, we could found that there is a major change that happens in the 1960s, it is consistent with the youth movement called **hippie** [5], which are seen as the ideal embodiment of the counterculture movement of the time. The rise of the hippie movement greatly influenced the development of music that followed. **People's thoughts on war** affected the R&B in the 1950s. Marvin Gaye's music soulful music soothed many war-torn families. In the 1970s, Britain's **economy** was in a recession because of the effects of the global economy, unemployment was high, and young people had little money and want to spare their redundant energy. Sex Pistol was first on stage as an anarchist, which promote the punk rock movement after the hippies [6]. With the breakthrough of **music synthesizer technology**, heavy metal music attracted people's attention, mainly represented by Metallica, their music has a lot of complex arrangements and a lot of gorgeous guitar solos.
  After discussing music genres in combination with data and other literature backgrounds, we can find that Pop/Rock music has occupied the music community for the longest time and has the greatest influence and the artists from this music genres are still will-known among people over the world.

10 Report to Integrative Collective Music (ICM) Society

Welcome to this music influence evaluation submit, and this page will help you to learn how to use our network model to know the value of music influence among artists and genres.

First of all, we hope you can do a discriminant analysis based on our thesis. Our evaluation is based on the network model, which is created by importing existing data sets and output different diagrams' layout to tell the value of music influence among artists and genres. Our model does not deal with musical influence outside of artists and genres. For the music influence among other areas, we will give you some advice.

Using our network model, you can see the different artists have a different number of edges, the more number of edges an artist has, the more influential the artist is. The edges here are music influential among artists. You can find the most influential artist in each year. For example, in 1960, the edges surround *"The Beatles"* is the most complicated, which means *"The Beatles"* is the most influential artist in 1960.

Using our similarity model, you can find the similarities among genres. For example, **Pop/Rock, R&B, and Religious** songs share a very closed relationship with each other, and the coefficient between the Pop/Rock and Religious is up to 0.8473, which is a very high level. Apart from that, you can also use **speechiness** of songs to distinguish a genre.

Music culture[7] can reflect the civilization and economic development of a nation. If a country's economic development is in good condition, its GNP (Gross National Product) [8] will be relatively good, and people's living standards will also be improved. At this time, people may have more surplus time for recreation besides work and daily necessities. Music culture develops gradually in people's constant entertainment. Through the continuous edification of music culture, people's quality is generally improved. Therefore, we can use our model to evaluate the music influence among economic development and music culture.

Apart from that, we also plan to use our model to research in three other areas in the future[9]:

- Geography Location:The origin geography location of music not only influence the music genre but also influence the lifespan of itself.
- Requirement:The music genres are usually formed with the social requirement. For example, music in Jazz genre fit the requirement of people who like dancing. Out of this kind of requirement, rock music and jazz music were born.
- Necessity:Sometimes, a genre is formed because society needs it. For the people who are forced to work against his will, a work song is a comforting tool. Similarly, rap music is raised in the 90s because of Tupac Shakur and Biggie Smalls against the brutality of the police and the unfair treatment of African Americans.

In the end, hope this model can give you some help in evaluating the value of musical influence.
Thanks for reading.

References

[1] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, 2000.

[2] G. M. Newman M E J, "Finding and evaluating community structure in networks," *Physical Review E*, p. 69 (2): 026113, 2004.

[3] M. J. Shi Jianbo, "Normalized cuts and image segmentation," *Departmental Papers (CIS)*, p. 107, 2000.

[4] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman, "An overview of principal component analysis," *Journal of Signal and Information Processing*, vol. 4, no. 3B, p. 173, 2013.

[5] J. R. Howard, "The flowering of the hippie movement," *The Annals of the American Academy of Political and Social Science*, vol. 382, no. 1, pp. 43–55, 1969.

[6] P. Friedlander, *Rock and roll: a social history*. Routledge, 2018.

[7] J. Toynbee, "Music, culture, and creativity," *The cultural study of music*, pp. 102–12, 2003.

[8] "Discuss the influence of music culture."

[9] "Music as influence: How has society been shaped by musical genres throughout history?."

A Appendix

The programmes we used in our model are as follows.

Input matlab source:

```
%PCA code
data=SimilaritydataS2;%
a=table2array(data);
b=normalize(a,"range");
r=corrcoef(b);%calculate the correlation coefficient matrix
[vec1,lamda,rate]=pcacov(r);%vec1 is the eigenvector
contr=cumsum(rate);
f=repmat(sign(sum(vec1)),size(vec1,1),1);
vec2=vec1.*f;
num=4;%The number of components we choose
df=b*vec2(:,1:num);
tf=df*rate(1:num)/100;
[stf,ind]=sort(tf,'descend');%sort the vector
stf=stf';
ind=ind';
xvar={"Avant-Garde","Blues","Children's","Classical","Comedy/Spoken","Count
yvar = {"danceability","energy","valence","tempo","loudness","mode","key","
h=heatmap(xvar,yvar,r);%draw the heatmap
h.title("CORRELATION HEATMAP");

%Similarity(cosine distance)
a=zscore(a);%Standardize the data
cos=[];%initialize the cos matrix
%cosine similarity
for i=1:20
    for j=1:20
        x1=a(i,:);
        x2=a(j,:);
        cos(i,j)=x1*x2'./sqrt(sum(x1.^2).*sum(x2.^2));
    end
end
x=cos;
xvar ={"Avant-Garde","Blues","Children","Classical","Comedy/Spoken","Country
h=heatmap(xvar,xvar,x);
h.title("MUSIC GENRES SIMILARITY");
```