Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
D		2104738

From Boogie to Indie: The Evolution of Modern Music

Summary

Classics never die, though the modern music industry has seen accelerating changes in styles throughout the past decades. Amidst this rapid change, past hits continue to shape new music, either from direct sampling, or through indirect influence. Sometimes, music also encapsulates public sentiment and may reflect contemporary social events. To see what influences the creation of music, in this paper we look into how artists inspire each other and investigate the major changes and general trends in music evolution through time.

We first implement **PageRank algorithm**, a centrality method commonly used to rank webpages, to determine the *music influence* of artists in a directed network of musical influences.

We then develop measures of *music similarity* by cross-referencing data obtained from Spotify's API. Our data processing comprises of **linear normalization** and the **Box Cox Transformation** to reshape the skewed distribution of characteristics. After performing a PCA check, we split each datapoint into two vectors and utilize **AHP** (Analytic Hierarchy Process) to assign weights to each characteristic/vocal.

With the cleaned data, we employ the **weighted norm** of the difference between two vectors to measure their **characteristic distance** and **vocal distance**. The total distance is a weighted sum of these two. To obtain the most informative/indicative weights, we devised and optimized an **objective function**, by minimizing distance between music of a single artist.

Built on our definition of *music similarity*, we perform various analyses on music evolution: 1) we inquire into the influences and similarities between genres both qualitatively and quantitatively; 2) we investigate whether influencers really affect their followers by comparing each artist's expected average to their followers and non-followers; 3) we model the *contagiousness* of each characteristic using conditional expectation and rank them.

To study the major leaps, we formulate quantitative definitions of musical revolutions and identify three recent revolutions: 1963-1969, 1976-1982, and 2012-2018. We establish a model, using **cosine distance,** to determine the **directional influences** of each artist during a revolution, and identify the revolutionists; we research into the **historical events** behind major musical revolutions, and **contextualize** how music "is the soundtrack of history."

To test the stability of our model, we vary the dampening factor d in our influence ranking model. As the ranking does not fluctuate with the variation of d, our model proves to be stable.

Throughout the paper, we also use multiple data visualization tools, such as **networks**, **heat map**, and **radar chart**, to better illustrate our findings. To help future researchers conveniently replicate our results, we publish our codes and essential results to a GitHub repository.

Keywords: PageRank; Centrality; Box Cox; AHP; Optimization; Graph Theory; Data Analysis

Contents

1 Introduction

1.1 Restatement of the Problem

In this problem, we are given four datasets. The first dataset describes the artists identified by other artists as sources of influence, while the second dataset provides a list of songs and their musical characteristics. The remaining two datasets group the songs by year and artist respectively. By analyzing the datasets, we aim to achieve the following objectives in this paper:

- Visualize the main features of the influencer-follower relations and develop a **music influence** metric for each artist, measured iteratively with the **PageRank algorithm** by the number of followers an artist has and how influential (as measured by *their* followers) the followers in turn are.
- Develop a metric for measuring **music similarity** between music pieces, based on the differences in their characteristics as given in *full_music_data*. We define music similarity to be the weighted root mean square of the difference between characteristics and vocals of two pieces of music.
- Determine if identified influencers actually influence their followers' music, and if some music characteristics are more 'contagious' than others. We say that a music characteristic is **contagious** if followers, compared to the other artists, bear strong similarity to their influencers in this characteristic.
- Identify **revolutions** both in individual genres and across all genres. We define a revolution to be a period when music similarity to previously created music drops significantly (and thus music created in this period is hugely different to previous work).
- Identify **revolutionaries** in revolutions. We define revolutionaries to be artists in periods of revolution who are highly influential (as measured by PageRank) and whose pre-revolutionary work bears strong similarity to the post-revolutionary music when compared to other artists.
- Explain trends in our data by establishing connections to historical socioeconomic development and major events.

1.2 Discussion of existing literatures on similarity and network

Many models have been proposed to determine the similarity of audio tracks. Prior research can be divided into three major categories: symbolic representations, acoustic properties, and subjective or 'cultural' information[5]. P. Knees et al.[4] devised a MIR model to compute the similarity of music using context-based music retrieval and indexing. R. Neapolitan et al.[6] summarized the technique of using the closeness in user activity to identify similarity in music, a powerful tool widely used in modern recommendation algorithm.

The notion of similarity is intimately connected to the concept of distance. A. Kassambara[2] summarized multiple metrics for measuring spatial distance, such as Euclidean distance and Manhattan distance, both of which constitute a special case of Minkowski distance. Pearson Correlation, Kendall Correlation, and Spearman correlation are also applied to determine the closeness of variables. These measures of distance serve as the foundation for cost functions in optimization problems.

The topic of determining the influence of musicians has been a fairly subjective field. Most journalists and bloggers analyze this question by considering the artists' style and personal background, while quantitative evidence tends to be largely overlooked. Adam Berenzweig et al.[5] pointed out that expert opinion is also widely used as a measure of influence and heritage of artists. Moreover, though famous musicians have enjoyed extensive coverage, the influence of smaller, less popular artists tend to be downplayed or ignored.

1.3 Our Work

Though existing models of evaluating musical similarity are abound, most techniques either focus on processing the raw audio file, or require user information as input. These approaches do not align well with the situation we face, where models should work with secondary data regarding the characteristics of music. Hence, we shall devise mechanisms to measure similarity that cater to our situation, based on the notion of distance and closeness.

Often, these classical definitions of distance constitute a foundation on which models that better serve a particular problem are designed and refined. Thus, out work regarding the distance function will start by adopting some existing definitions of distance, and then focus on optimizing its effectiveness.

We also explore the influence of musicians from a network theoretic point of view. Employing the combinatorial model of directed graphs, we perform centrality-related algorithms on the network of artists. This approach allows us to (i) delve into this topic more objectively and quantitatively, (ii) shed light on the less famous artists previously underrepresented in existing literature.

2 Assumptions

We make the following simplifying assumptions:

- If an artist has an influencer and a follower, then the artist's influencer also exerts a degree of influence on the artist's follower. This is generally accurate, since an artist's work will contain elements of their influencer, which will in turn influence their follower.
- An artist produces songs that are generally similar to each other. This is justified as artists usually specialize in one or few genres – which means the songs are similar – and the variability of the songs are limited by their creativity. While an artist may produce songs that are of different genres and thus are less similar, we are limited from performing further analysis on this by the fact that *full_music_data* does not specify genres.
- The data provided is an unbiased representation of the entire music industry, i.e., no genre is underrepresented / overrepresented in the data compared to the actuality.

3 Symbols

Table 1 lists the important symbols used in this paper:

Table 1: notation

Symbol	Explanation
N	The total number of artists
A_i	The ith artist
$\mathrm{PR}_n(A_i)$	The PageRank of artist A_i in the nth iteration
$\mathrm{PR}(A_i)$ or $\mathrm{PR}_\infty(A_i)$	The PageRank of artist A_i (to which $\mathrm{PR}_n(A_i)$ converges)
$I(A_i)$	The set of all influencers to artist A_i
d	The dampening factor
σ_{ij}	The *separation* of artist A_j from artist A_i
$\mathrm{dist}(p,q)$	The *distance* between two music, artists, or years, p and q
J	The cost function for the *music similarity* metric
$\mathrm{S}(p,q)$	The *music similarity* between two music, artists, or years, p and q
m_i	The ith piece of music in *full_music_data*
$M(A_i)$	The set of music produced by artist A_i
$\mathbf{c}_{A_i}$	The *characteristics vector* of artist A_i
$\mathbf{v}_{A_i}$	The *vocal vector* of artist A_i
α	The weight for $\lvert\mathbf{c}_i - \mathbf{c}_j\rvert^2/7$ in J, the *music similarity* metric
β	The weight for $\lvert\mathbf{v}_i - \mathbf{v}_j\rvert^2/5$ in the *music similarity* metric
G_i	The ith genre
$\mathcal{I}(G_i, G_j)$	The *influence* of genre G_i on genre G_j
$\mathrm{G}\ (A_i)$	The main genre associated with artist A_p (as given in *influence_data*)

4 Influencers and Followers – A Directed Network

4.1 Ranking Influence with PageRank Algorithm

4.1.1 A Simple Model

In the first part of the problem, we aim to gain insight into how artists influence each other, given data about artists and their influencers/followers. To do so, we will establish a measure of musical influence and determine the artists who have the most influence on others.

Firstly, we consider the network of artists. A directed edge is connected from a follower to their influencer. To determine which artists are the most influential, our measure will satisfy the points below:

If artist P influences artist Q, then

- The more follower P has, the higher is their music influence, as they are influencing more artists;
- The higher the music influence Q has achieved, the higher is the music influence of P as a result, since the influence of P is carried on through the wider influence of Q.

To satisfy these, we implement the **PageRank algorithm**. We define the music influence of an artist A to be their PageRank in the graph, which we denote by $\mathrm{PR}(A)$. Let N be the number of artists. Initially, each artist has PageRank $1/N$, and in each iteration the total PageRank of all artists sum to 1. In each iteration, each artist transfers their PageRank to all their influencers, spread equally among them, so that

$$\begin{cases} \mathrm{PR}_0(A_i) = \dfrac{1}{N} \\ \mathrm{PR}_n(A_i) = \displaystyle\sum_{X_i \in I(A_j)} \frac{\mathrm{PR}_{n-1}(A_j)}{|I(A_j)|} \end{cases}, i = 1, \dots, N$$

where the A_i's are the artists, $\mathrm{PR}_n(A_i)$ is the PageRank of A_i in the nth iteration and $I(A_i)$ is the set of influencers of artist A_i.

We can see that this model satisfies the two points above: an artist's PageRank increases with the number of followers and their influencers' PageRank. In each iteration, the PageRank is redistributed from followers to their influencers. After sufficiently many iterations, the PageRank of artist A_i will converge (to $\mathrm{PR}(A_i)$).

4.1.2 Dampening Factor

While this simple model ensures that net influencers in general receive a higher PageRank than net followers in the same period, earlier artists will receive a disproportionately large fraction of the total PageRank. This is because they have more generations of followers (followers of followers, etc., which all contribute to their PageRank), and they have less generations of influencers since some of them are too early to be included in the data. In the extreme case, the earliest artist in the data will have no influencer present and thus their PageRank will be non-decreasing throughout the iterations.

Therefore, although earlier artists may have more influence as their influence accumulates through generations, we need to consider the decay of their reputation as time progresses (e.g. as their music style becomes less common), which our model currently does not address, by introducing a dampening factor d $(0 < d < 1)$:

$$\mathrm{PR}_n(A_i) = \frac{1-d}{N} + d\left(\sum_{X_i \in I(A_j)} \frac{\mathrm{PR}_{n-1}(A_j)}{|I(A_j)|}\right)$$

In this improved model, in each iteration the total contribution of PageRank from an artist's followers is dampened by a factor of $0 < d < 1$, and each artist has a PageRank of at least $\frac{1-d}{N}$.

This reduces the inflation of earlier artists' PageRank due to repeated contribution from followers as discussed above. Thus, the smaller d is, the more reduced are the earlier artists' influence due to seniority.

4.2 Visualizing Artist Influence

4.2.1 Directed Network of Influence

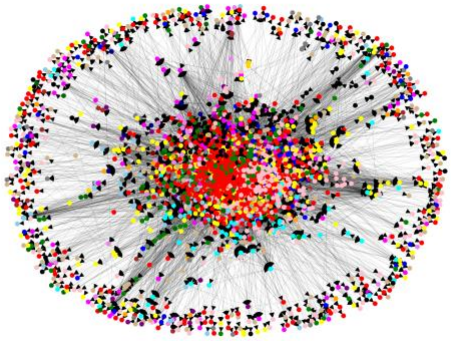

Figure 1: full network given in influence_data, each color represents a genre

Fig. 1 shows the directed network of artist influence, using the full *influence_data* data set. Each node represents an artist; a directed edge is directed from a follower to an influencer. The color of the node represents the main genre of the artist.

As can be seen, the network is extremely large and complex due to the sheer number of influencer-follower relations. To extract more meaningful information about the structure of the network and the interpretation of the influence measure (PageRank), we now look at an arbitrary subset of artists:

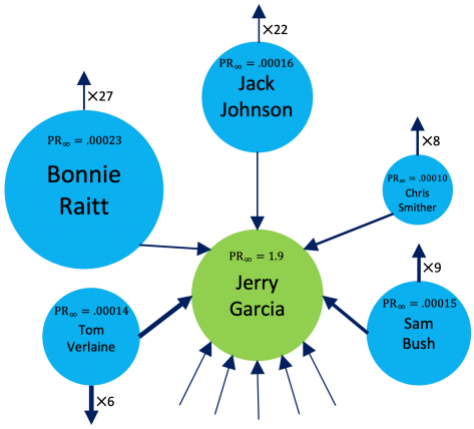

Figure 2: an arbitrary subnetwork with node size proportional to PageRank and edge width proportional to contribution

In Fig. 2, we consider the followers of an arbitrary artist, Jerry Garcia. The size of each node is proportional to the PageRank of the artist; each edge directed to it is a follower (e.g., Jerry Garcia has 10 followers, Bonnie Raitt has 28 influencers), and their thickness is proportional to the follower's contribution of PageRank to the artist ($\mathrm{PR}(A_j)/|I(A_j)|$, if the follower is A_i). From the illustration, we observe the following:

- An artist's influence depends on their followers' influence. For example, Sam Bush contributes a higher PageRank than Chris Smither, though both have similar influencers; this reflects Jerry Garcia carries more influence through Sam Bush, since Sam Bush is more influential and will more successfully spread Jerry Garcia's influence.
- An artist's influence depends on how many influencers their followers have. For example, despite having a higher PageRank, Bonnie Raitt contributes less PageRank to Jerry Garcia than Tom Verlaine; this reflects that Jerry Garcia does not influence Bonnie Raitt much, amongst Bonnie Raitt's large number of influencers.

This suggests that by using the PageRank algorithm, we have accurately captured the influence process in the music industry.

4.2.2 Revealing Top Influencers

We run 100 iterations of the PageRank Algorithm with $d = 0.5$ and sort the artists by their PageRank $\mathrm{PR}_{100}(A_i)$. The five influencers with greatest PageRank are: **The Beatles, Cab Calloway, Bob Dylan, Louis Jordan, and Billie Holiday** (each with $\mathrm{PR}_{100}(A_i) > .0038$).

We measure the validity of our results by comparing it with established rankings. The top five artists ranked by *Rolling Stone* and their PageRank ranking are shown in the table below:

Table 2: comparison of PageRank ranking and Rolling Stone ranking

Artists	The Beatles	Bob Dylan	Elvis Presley	The Rolling Stones	Chuck Berry
Rolling Stone ranking	1	2	3	4	5
$\mathrm{PR}(A_i)$ ranking	1	3	14	7	11

4.3 Shortest Path Matrix

In our following work, it will be helpful to have a notion of the degree of influence any given artist has on some other artist. It is rational to assume that if artist A is an influencer on artist B, who is an influencer on artist C (who is not a follower of artist A), then artist A will have a stronger influence on artist B, while having a diminished influence on artist C through the influence of artist B.

In the influencer-follower directed graph, a directed edge is connected from artist A_i to artist A_j iff A_i is a follower to A_j. With the previous paragraph in mind, we define the **shortest path** from artist A_i to A_j, to be the directed path (if it exists) that connects A_i to A_j passing through the least nodes (artists). The **separation** of artist A_j from artist A_i, σ_{ij}, is equal to the number of directed edges in the shortest path from artist A_i to A_j ($\sigma_{ii} = 0$). If there is no such shortest path, then $\sigma_{ij} = \infty$.

5 Data Normalization and Transformation

5.1 Normalization and Transformation Methods

5.1.1 Linear Normalization

Since some raw data (e.g., duration) is given in different units, we need to normalize all values. Using [0,1] linear transformation, we introduce the following normalization map LN:

$$\mathrm{LN}(x) = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x_{max}, x_{min} denote the maximum and minimum in the index (the column) of x, respectively.

5.1.2 Box Cox Transformation

Plotting the histogram for each index, we find that some distributions are heavily skewed.

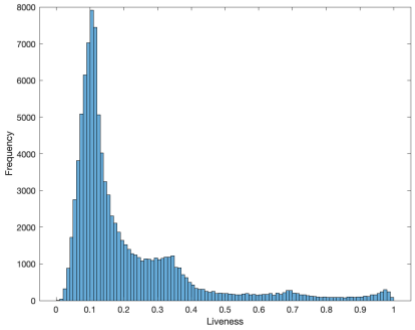

Figure 3: The Distribution of "Liveness" is skewed

For a random variable X, its **skewness** quantitatively measures how asymmetric the distribution is, given quantitatively by

$$\widetilde{\mu_3} = \mathbb{E}\left[\left(\frac{X - \mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3}$$

To estimate the population skewness, we use the sample skewness, given by:

$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{3/2}}$$

As a skewed data distribution may distort the relative influence of this index, we must reshape the data to distribute the values **more evenly**. We thus introduce the Box Cox[12] transformation $\mathrm{BC}(x)$:

$$\mathrm{BC}(x) = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log x, & \text{otherwise} \end{cases}$$

The histogram below illustrates how the Box Cox Transformation reshapes the data, spreading it out more evenly. Note that this transformation can only be applied to positive values.

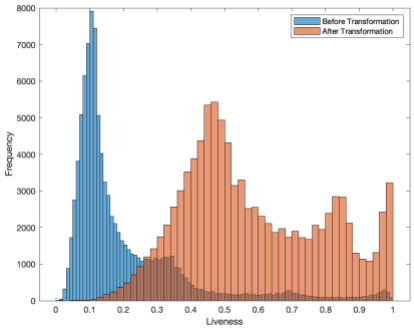

Figure 4: comparison of the distribution of "liveness" before and after transformation

5.2 Performing Data Normalization and Transformation

With these two methods described above, we devise the following procesure to process the data in *full_music_data.csv:*

For each column (except "mode" and "explicitness", which have binary 0-1 data):

1. We measure sk, the skewness of this set of data;
2. If $|sk| \leq 2$, we apply the [0,1] linear normalization to the data
   If $|sk| > 2$, we
   (i) shift the data to make every value positive (to prepare for the Box Cox); then
   (ii) apply the Box Cox Transformation; and
   (iii) apply the [0,1] linear normalization to the data.

5.3 Assigning Weights to Characteristics

To prepare for our definition of the distance function, we first determine the weights of each component in the characteristics vector and the vocals vector.

Strongly correlated variables will be overrepresented in the distance function, resulting in double-counting, bias, and higher sensitivity.[1] To prevent this phenomenon, we first perform a PCA (Principal Component Analysis) on the normalized data to investigate the correlation among variables, results shown below.

Principal Component	PC1	PC2	PC3
Variance Explained	24.3949	23.7850	14.6219

Since the first three Principal Components can only explain a small percentage of the variance, the variables are not strongly correlated.

[1] Tavory, A. (2017, November 07). In supervised learning, why is it bad to have correlated features? Retrieved February 08, 2021, from https://datascience.stackexchange.com/questions/24452/

We proceed to quantify the weights of different components. Comparing the relative significance of 12 components simultaneously may result in bias and inconsistency. Hence, we split the data into two groups (characteristics and vocals), and then employ the **AHP** (Analytic Hierarchy Process) to determine weights through comparing **pairwise** relative importance.[9] Comparison matrices are given in Table 3.

Table 3: Comparison matrices of different characteristics for AHP

	Danceability	**Energy**	**Valence**	**Tempo**	**Loudness**	**Mode**	**Key**
Danceability	1.0000	0.7407	0.3536	1.1111	1.2500	0.6897	3.3333
Energy	1.3500	1.0000	0.8696	1.1111	1.6667	1.0000	4.0000
Valence	2.8284	1.1500	1.0000	1.4286	1.6667	1.6667	10.0000
Tempo	0.9000	0.9000	0.7000	1.0000	1.4286	0.8333	2.5000
Loudness	0.8000	0.6000	0.6000	0.7000	1.0000	0.7692	1.6667
Mode	1.4500	1.0000	0.6000	1.2000	1.3000	1.0000	5.0000
Key	0.3000	0.2500	0.1000	0.4000	0.6000	0.2000	1.0000

	Acousticness	**Instrumentalness**	**Liveness**	**Speechiness**	**Explicit**
Acousticness	1.0000	0.8000	5.0000	0.6250	1.0000
Instrumentalness	1.2500	1.0000	3.3333	0.8696	1.2500
Liveness	0.2000	0.3000	1.0000	0.5000	0.6667
Speechiness	1.6000	1.1500	2.0000	1.0000	1.1111
Explicit	1.0000	0.8000	1.5000	0.9000	1.0000

Now we check for inconsistency. The $\mathrm{CR} = \mathrm{CI}/\mathrm{RI}$ values for characteristics and vocals are 0.01932 and 0.04406, respectively. Since $CR < 0.10$ for both matrices, we consider comparison process consistent[9] and assign weights accordingly, as shown in Table 4.

Table 4: weights for each characteristics

Component	Danceability	Energy	Valence	Tempo	Loudness	Mode
Weights	0.1223	0.1704	0.2602	0.1360	0.1052	0.1644

Component	Acousticnes	Instrumentalness	Liveness	Speechiness	Explicit
Weights	0.2326	0.2458	0.0877	0.2472	0.1866

For notational convenience, let $\mathrm{W}_{char}, \mathrm{W}_{voc}$ be two diagonal matrices, with the weights of characteristics and those of vocals on the diagonal. Hence, the weighting process becomes:

$$\mathbf{c} \to \mathrm{W}_{char}\,\mathbf{c} \text{ and } \mathbf{v} \to \mathrm{W}_{voc}\mathbf{v}$$

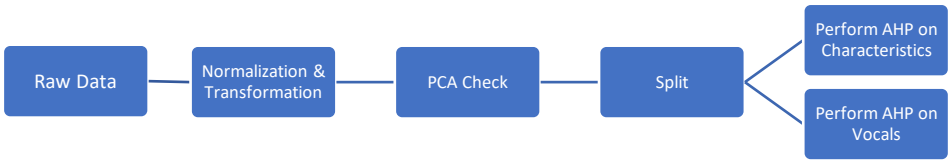

Figure 5: our data processing procedure

6 Music Similarity

6.1 Distance Function

If p is a piece of music, an artist or a year, then the **characteristics vector** $\mathbf{c}_p$ of p is a vector, where the nth entry is the normalized and transformed score (as in Section 5) of p in the nth characteristic column (as given in *full_music_data*, *data_by_artist* or *data_by_year* respectively). The **vocal vector** $\mathbf{v}_{A_i}$ of an artist A_i is defined similarly for the vocal columns.

We define the **distance function** $\text{dist}(p, q)$ between two music, artists, or years, p and q to be as follows:

$$\text{dist}(p,q) \text{ (with weights } \alpha,\beta) \coloneqq \sqrt{\alpha^2 \cdot \frac{\left|\mathrm{W}_{char}(\mathbf{c}_p - \mathbf{c}_q)\right|^2}{7} + \beta^2 \cdot \frac{\left|\mathrm{W}_{voc}(\mathbf{v}_p - \mathbf{v}_q)\right|^2}{5}}$$

where α, β are constant weights such that $\alpha, \beta \geq 0$ and $\alpha + \beta = 1$
$\mathbf{c}_p$ is the *characteristics vector* of p
$\mathbf{v}_p$ is the *vocal vector* of p

We divide the squared norm of $\mathrm{W}_{char}(\mathbf{c}_p - \mathbf{c}_q)$ and $\mathrm{W}_{voc}(\mathbf{v}_p - \mathbf{v}_q)$ by 7 and 5 respectively because the characteristics vector contains 7 features (danceability, energy, etc.) and the vocal vector contains 5 features (acousticness, instrumentalness, etc.).

Since we assume that an artist tends to produce similar music, we thus expect distance function to **minimize** the distance of music produced by the same artist. To find the optimum weights α and β, we introduce the cost function below for α and β, which we wish to optimize:

$$J(\alpha,\beta) \coloneqq \sum_{i=1}^{N} \left(\sum_{\substack{m_j, m_k \in M(A_i), \\ j<k}} \text{dist}(m_j, m_k) \text{ (with weights } \alpha,\beta) \right)$$

where J is the cost function
N is the total number of artists
m_i is the ith piece of music in *full_music_data*
A_i is the ith artist
$M(A_i)$ is the set of music produced by artist A_i

We optimize the cost function using MATLAB optimization toolbox, which yields us $\alpha^2 = 0.1265, \beta^2 = 0.4125$.

6.2 Music Similarity

We proceed to define the music similarity function $S(p, q)$ between two music, artists, or years, p and q. The larger the distance between p and q is, the less similar they are, and vice versa. Hence, $S(p, q)$ should be a strictly decreasing function of $\text{dist}(p, q)$. We further require

the value of $S(p,q)$ to be comparable with that of $\text{dist}(p,q)$. Hence, we define the **music similarity** $S(p,q)$ between two music, artists, or years, p and q to be:

$$S(p,q) = e^{-\text{dist}(p,q)}$$

The majority of the randomly chosen pairs of music (p,q) have $0.1 \leq \text{dist}(p,q) \leq 0.5$, and the distribution of $\text{dist}(p,q)$ is centered at around 0.3. Hence, a typical value of similarity will lie between $e^{-0.5} = 0.61$ and $e^{-0.1} = 0.90$.

However, we will not use this music similarity metric for the rest of this paper, but instead continue to employ the distance function $\text{dist}(p,q)$ when comparing two music, artists, or years.

7 Comparing Genres

7.1 Measuring Influences & Similarities

To investigate the influences of music between and within genres, we develop the following model. For each genre of artists, we consider the artists from all genres whom they identified to have an influence on them. If a particular genre A has a strong influence on genre B, then we expect more artists from genre B to identify artists from genre A as influencers, and vice versa.

In this definition, multiplicity matters, since if an artist from genre A has influenced many artists in genre B, then we should enumerate the presence of this influencer proportionately many times (instead of just counting the absolute number of influencers on this genre). Therefore, we will **count all (influencer, follower) ordered pairs** separately.

Moreover, we shall consider the relative influence of one genre to a given genre, compared to the combined influence that the whole music industry (i.e., all genres) on the given genre. We define the *influence* that a genre has on a genre (not necessarily different) as follows:

$$\mathcal{I}(G_i, G_j) := \frac{\left|\left\{(A_p, A_q) : \left(A_p \in I(A_q)\right) \wedge \left(\text{G}\left(A_p\right) = G_i\right) \wedge \left(\text{G}\left(A_q\right) = G_j\right)\right\}\right|}{\left|\left\{(A_p, A_q) : \left(A_p \in I(A_q)\right) \wedge \left(\text{G}\left(A_q\right) = G_j\right)\right\}\right|}$$

where G_i is the ith genre; A_p is the pth artist
$\mathcal{I}(G_i, G_j)$ is the *influence* of genre G_i on genre G_j
$I(A_p)$ is the set of influencers on artist A_p
$\text{G}\left(A_p\right)$ is main genre associated with artist A_p (as given in *influence_data*)

After calculating $\mathcal{I}(G_i, G_j)$ for all possible ordered pairs of genres (G_i, G_j), we arrive at the heatmap below:

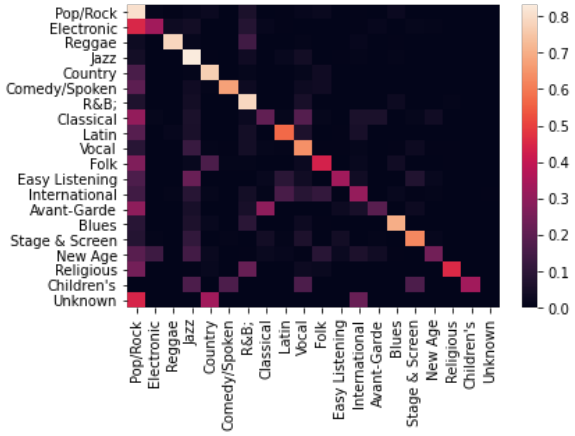

Figure 6: Influences between Genres

In Fig. 6, the numerical value of the cell in the ith row and jth column is equal to $\mathcal{I}(G_j, G_i)$ (the influence of genre j on genre i). A brighter cell represents a higher influence. We have the following observations from the heatmap:

- Artists from a given genre are most strongly influenced by other artists in the same genre. This is shown by the strong intensity of the cells in the main diagonal (where $i = j$) compared to elsewhere.
- The pop/rock genre, and to a lesser extent the jazz genre, has a relatively strong influence across all genres. This is seen from the stronger intensity of the cells in the respective columns. Interestingly, electronic artists are most strongly influenced by pop/rock artists, which may be due to the relatively young age of the electronic genre.

7.2 A Broader View of Genres

7.2.1 Representatives of Genres – The "Composite" Music Piece

In this section, we compare the characteristics of a genre to other genres and across time. Genres are typically distinguished by the characteristics in their music – we shall now look at how to identify the characteristics most common to a genre.

For each music m_i in *full_music_data*, its corresponding **music vector** $\mathbf{m}_i$ is the vector where the nth entry is the normalized and transformed score in the nth characteristic or vocal column in *full_music_data*. We define the **genre-average vector** $\mathbf{g}_i$ for a genre G_i to be as follows:

$$\mathbf{g}_i \coloneqq \text{average}\left(\{\mathbf{m}_i : \mathrm{G}(A(m_i)) = G_i\}\right)$$

where $\mathrm{G}(A_i)$ is the main genre of the artist A_i (as given in *influence_data*)
$\mathrm{A}(m_i)$ is the artist who produced music m_i

The entries in $\boldsymbol{g}_i$ thus represents the **average characteristics (and vocals)** of the songs in genre G_i.

7.2.2 Distinguishing Characteristics of Genres

We begin by presenting a radar chart overview of the characteristics/vocals across genres:

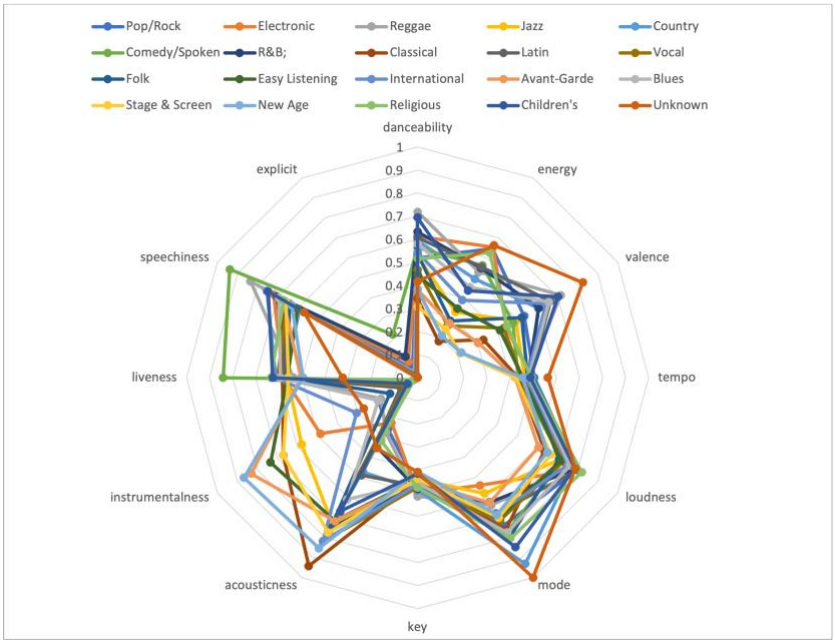

Figure 7: Characteristics/Vocals across Genres

Fig. 7 illustrates that genres differ by a large margin in some characteristics/vocals (e.g., instrumentalness), but tend to behave similarly in some other characteristics/vocals. We now inquire into this observation from a quantitative approach.

For each genre G_i, we compare $\mathbf{g}_i$ to the weighted average $\mathbf{g}_i'$ of all other genre vectors, weighted by the number of music pieces in each genre. For the nth characteristic (or vocal), we shall consider the ratio of the nth entry in $\mathbf{g}_i$ to the nth entry in $\mathbf{g}_i'$. This represents how much the nth characteristic (or vocal) of genre G_i **differs from the music in the other genres.**

Table 5 shows the fraction by which a genre's "composite" music is above or below the average of music of other genres in each characteristic: for example, reggae music is roughly 34% more danceable, while jazz music has roughly 40% less energy, compared to the average music in the other genres.

Table 5: Comparison of Selected Genres to Others on Selected Characteristics

	danceability	energy	valence
Pop/Rock	-6.34%	+51.93%	-1.15%
Electronic	+13.88%	+18.76%	-19.09%
Reggae	+34.47%	-4.79%	34.71%
Jazz	-1.97%	-40.20%	-9.20%
Country	+12.55%	-8.35%	+13.74%

Then, we determine the three columns, along which genre characteristics have the **highest variance**. These are instrumentalness, energy and explicitness; they are thus the characteristics that best discriminate genres.

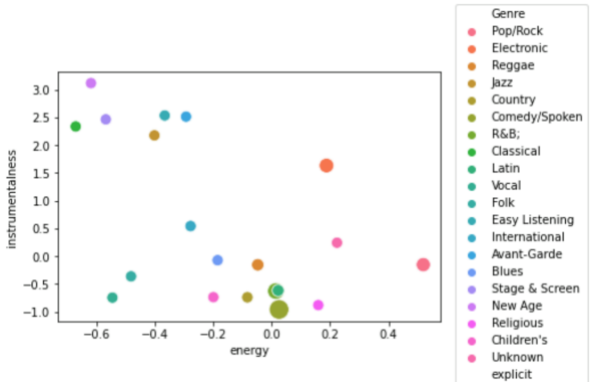

Figure 8: Genres Plotted along Characteristics of Greatest Variance

In Fig. 8, the position of the genres corresponds to their scores in instrumentalness and energy, while the size of their points is proportional to explicitness

7.2.3 Genre Evolutions through the Ages

Because the concept of "genre" serves as a **categorization** of music, each genre **encodes some characteristics** that distinguishes itself from the rest. However, a curious question to ask is: Do genres change over time? In particular, do the characteristics of a genre **stay stationary**, or do they exhibit statistically significant change?

To investigate how genres change over time, we

- organize the normalized music data by grouping all music with the same year and genre
- summarize the characteristics using **genre-average vector** $\mathbf{g}_i$

To statistically test the stationarity, we then formulate the following hypotheses:

Let $TS_{c,g}$ denote the time series for characteristic c of genre g.		
	$H_0(\text{Char } c, \text{Genre } g)$:	The time series TS is non-stationary
	H_a:	The series is stationary

We apply the **ADF (Augmented Dickey–Fuller) Test** to test our hypotheses. The results from the ADF Test demonstrate that for most characteristics of most genres, there is not enough evidence to reject the null hypothesis **that the series is a unit root process**.

Thus, we consider most series non-stationary, concluding that, statistically, **most genres do change over time.** This conclusion agrees with the visualization of the characteristics, illustrated in Fig. 9:

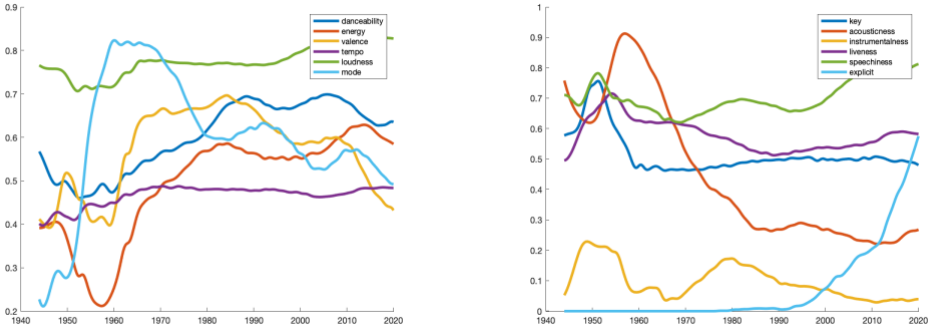

Figure 9: Changes of R&B Music over Time

In Fig. 9, the different characteristics of the "average" R&B music, i.e., the entries in its average genre vector ($\mathbf{g}_{\mathrm{R\&B}}$) in each year is plotted against time. By analyzing the trends in individual characteristics, we find that the fluctuations in the data correspond to the evolution of R&B music:

- Speechiness has been generally increasing from the 1990s; this corresponds to the integration of rap culture into R&B, as exemplified by artists such as R. Kelly and Janet Jackson who are popular during this era.
- Explicitness has increased rapidly since the 1990s and accelerated after 2010; this corroborates with the normalization of swearing following the advent of social media, as well as the popularity of streaming services, which means artists do not have to rely on radio play for chart successes and can thus not comply to the more stringent radio censorship rules[8]
- There is a steady fall in acousticness, which was most rapid from 1960 to 1990; this is explained by the incorporation of disco music into R&B music at around 1970, which popularized the use of electrical instruments such as electric piano and synthesizers.

8 How Influencers Change their Followers

8.1 Measuring Influence on Music

Based on our definition of $\mathrm{dist}(p, q)$, we investigate whether the 'influencers' actually affect the music created by the followers. If yes, then the music created by the influencers should share more similarities to their followers than other artists. Hence, we hypothesize that the average distance between each influencer A_i and his followers, $D_{follow}(A_i)$, should be less than the average distance between A_i and non-followers, $D_{nonFollow}(A_i)$. Formally, for each artist A_i,

$$D_{follow}(A_i) = \frac{1}{|I(A_i)|} \sum_{A_j \in I(A_i)} \mathrm{dist}(A_i, A_j) \leq \frac{1}{N - |I(A_i)|} \sum_{A_j \notin I(A_i)} \mathrm{dist}(A_i, A_j) = D_{nonFollow}(A_i)$$

where $\mathrm{dist}(A_i, A_j)$ denotes the distance between "average" music of A_i and A_j .

The "**artist-average**" music vector of A_i is denoted by **characteristics vectors**

$$\mathbf{c}_{avg_{A_i}} = \frac{1}{|M(A_i)|} \sum_{p \in M(A_i)} \mathbf{c}_p$$

and **vocal vectors**

$$\mathbf{v}_{avg_{A_i}} = \frac{1}{|M(A_i)|} \sum_{p \in M(A_i)} \mathbf{v}_p$$

$$\mathrm{dist}(A_i, A_j) = \sqrt{\alpha^2 \cdot \frac{\left|\mathrm{W}_{char}(\mathbf{c}_{avg_{A_i}} - \mathbf{c}_{avg_{A_j}})\right|^2}{7} + \beta^2 \cdot \frac{\left|\mathrm{W}_{char}(\mathbf{v}_{avg_{A_i}} - \mathbf{v}_{avg_{A_j}})\right|^2}{5}}$$

Using above equations, we now compute $D_{follow}(A_i)$ and $D_{nonFollow}(A_i)$ for each artist A_i. In order to better illustrate the general picture, we then take the average of both values over all artists. We get

$$\overline{D_{follow}(A_i)} = 0.2927 < 0.4079 = \overline{D_{nonFollow}(A_i)}$$

This confirms the hypothesis. We thus conclude that, based on our definition of distance (and music similarity), the 'influencers' do affect the music created by the followers.

8.2 Identifying Contagious Characteristics

We interpret **contagiousness** in the following ways:

- Qualitatively, a characteristic (e.g., valence) is said to be *contagious* if this characteristic of influencers carries over into this characteristic of their followers. (e.g., if an influencer and his followers tend to demonstrate more similar levels of valence compared to others characteristics, then valence is highly contagious.)
- Quantitatively, the more contagious a characteristic c is (e.g., valence), the closer followers should be to their influencers in terms of this characteristic. We measure closeness using the absolute difference in this characteristic between the influencers and their followers (e.g., if artists A_i and A_j have valence 0.4 and 0.65 respectively, then the closeness is 0.25).

We thus devise the following measure of contagiousness:

- Let $\mathrm{Char}_c(A_i)$ denote the value of the characteristic c of the artist A_i, where
$$c \in C = \{\text{danceability}, \text{energy}, \dots, \text{mode}, \text{key}\}$$
- Then contagiousness index of characteristic c is defined as a function:
$$\mathrm{cont} : S \to \mathbb{R}, \qquad \mathrm{cont}(c) = \frac{\mathbb{E}[\,|\mathrm{Char}_c(A_i) - \mathrm{Char}_c(A_j)|\,]}{\mathbb{E}[\,|\mathrm{Char}_c(A_m) - \mathrm{Char}_c(A_n)|\,]}$$

where A_i, A_j are arbitrary artists
A_m, A_n are artists s.t. $\sigma_{mn} \leq l$
l is the maximum *separation* between artists under consideration

Interpretation of this contagiousness function:

- Numerator: absolute difference in this characteristic, averaged over **arbitrary artists pairs**;
- Denominator: absolute difference in this characteristic, averaged over **artists pairs that influence each other**, directly or indirectly, within l layers of influence (*separation* $\leq l$).
- For **highly contagious** characteristic c, followers are expected to be significantly **close** to their influencers. This fact is well demonstrated in the contagious index: closeness results in a smaller denominator and thus a **greater** contagiousness index.

We now compute the contagiousness index $\mathrm{cont}(c)$ for each characteristic c, where $c \in \{\text{danceability}, \text{energy}, \dots, \text{mode}, \text{key}\}$. We set $l = 1$, because the influence of an artist can be carried to the followers of his followers through his followers; however, l can take other values.

Table 6: Contagiousness of each Characteristic

Name	energy	danceability	valence	mode	loudness	tempo	key
Contagiousness	1.4704	1.4105	1.3646	1.3293	1.2976	1.2961	1.2778

The results shown in Table 6 indicate *energy* and *danceability* are the most contagious characteristics while *tempo* and *key* are the least. This is consistent with our intuition: the energy of a song leaves the strongest impression of the music in its listeners.

9 Revolutions in the Evolution

9.1 Quantitatively Identifying Revolutions

The overall style of music is constantly evolving, but not every change is created equal. Some are slow, minor, and incremental; others, rapid, major, and profound. In this section, we explore the revolutions (major leaps) that took place in the history of music.

Our interpretation of **revolution**:

- Qualitatively, we consider a period to be a revolution if and only if it **differs** from previous years **to a significant extent**.
- Quantitatively, year *t* is said to be a revolution if and only if there is a large **distance** between the (average) characteristics of year *t* and those of the *n* previous years.

With the normalized full music data (grouped by year), we employ our **distance** function $\text{dist}(\mathbf{u}, \mathbf{v})$ to devise the following function to **measure revolution**. Let $\text{Years} = \{1926, 1927, \ldots, 2020\}$ be the set of years, and

$$\text{rev}: \text{Years} \to \mathbb{R}; \qquad \text{rev}(t) = \text{dist}(\boldsymbol{\mu}_t, \boldsymbol{\mu}_{t-n})$$

where $\boldsymbol{\mu}_t$ denotes the mean characteristics of all music produced in year t, and $n \in \mathbb{Z}^+$.

In Fig. 10, we take $n = 3$ and plot the function $\text{rev}(t)$ against t for $t \in \text{Years}$.

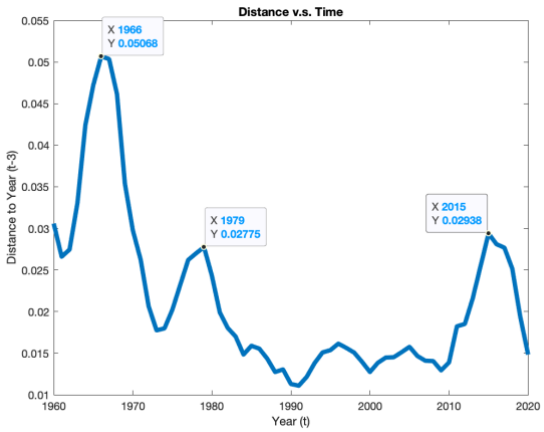

Figure 10: Difference between year (t) and year (t-3) over time

From Fig. 10, we can see that the rev function peaks at t = 1966, 1979, and 2015, respectively. This implies that the (average) music released in 1966, 1979, and 2015 are **notably different** from the (average) music released 3 years before. Thus, the data signifies that musical

revolutions might have taken place in 1966, 1979, and 2015. We now proceed to discuss the events (and potential revolutions) that happened in these years.

Note: the trend before 1960 is not analyzed here, as very few data were available for those years, which resulted in unacceptably high variance and noise.

9.2 The Engine of Revolution: Revolutionary Artists

After determining the three revolutions (around 1966, 1978, and 2015, respectively), we progress to identify the influential artists that drove the revolution. We divide the process into 4 steps:

1. Modelling the artist's directional influence on the industry

Each artist A_i deviates from the average of their contemporaries by a **direction vector**, given by

$$\mathbf{\Delta a = a_i - m}$$

where $\mathbf{a_i}$ (or $\mathbf{m}$, resp.) denote mean characteristics of all music produced by artist A_i (or by their contemporaries, resp.) during the period of revolution.

In a qualitative sense, the artist A_i influences their contemporaries by **pulling** other creators to a direction specified by the characteristics of A_i's music. We thus interpret the deviation vector $\mathbf{\Delta a}$ as the **directional influence** that the artist A_i exerts on the industry during this particular revolution. Note that $\mathbf{\Delta a}$ is dependent on time and so varies from revolution to revolution.

2. Modelling the direction of the revolutions

Each revolution has a **direction vector** (e.g., 0.23 increase in valence, 0.16 decrease in energy, etc.), given by

$$\mathbf{\Delta \mu = \mu_{end} - \mu_{start}}$$

where $\mathbf{\mu_{after}}$ (or $\mathbf{\mu_{before}}$, resp.) denote mean characteristics of all music produced in the year immediately after (or before, resp.) the period of revolution. We interpret the difference vector $\mathbf{\Delta \mu}$ as the direction to which the revolution progressed.

3. Comparing the directions

For each revolution, the artist A_i is a revolutionary if and only if the influence A_i exerts on the music industry has a direction that matches the way where the revolution heads toward.

Cosine distance captures angle between two direction vectors. For two vectors to have similar directions, **we expect their angle to be small**, which corresponds to a large cosine value.

Hence, we employ the cosine distance to measure how close the two direction vectors $\mathbf{\Delta a}$, $\mathbf{\Delta \mu}$. Let $< \mathbf{\Delta a}, \mathbf{\Delta \mu} >$ denote the angle between $\mathbf{\Delta a}, \mathbf{\Delta \mu}$

$$\Rightarrow \quad \cos < \mathbf{\Delta a}, \mathbf{\Delta \mu} > = \frac{\mathbf{\Delta a} \cdot \mathbf{\Delta \mu}}{|\mathbf{\Delta a}||\mathbf{\Delta \mu}|}$$

$$\Rightarrow \quad < \mathbf{\Delta a}, \mathbf{\Delta \mu} > = \arccos\left(\frac{\mathbf{\Delta a} \cdot \mathbf{\Delta \mu}}{|\mathbf{\Delta a}||\mathbf{\Delta \mu}|}\right)$$

4. **Identifying Revolutionary Artists**

Combining previous results, we specify the following criteria artist A_i to be revolutionary:

- A_i must be prolific during the revolution period and **released at least 3 works**;
- **Δa**, the deviation vector of A_i, must be close to **Δμ,** and the angle between them must be within a specified angle:
- $$< \mathbf{\Delta a}, \mathbf{\Delta \mu} > \ \leq \theta_0$$
  In this problem we choose θ_0 to be $\frac{\pi}{4}$.
- A_i must be influential. This is measured using PageRank.

Encapsulating these criteria, for each revolution (around 1966, 1978, and 2015, respectively), we search for artists A_i such that

$$\begin{cases} |S_i| \geq 3 \\ \dfrac{\mathbf{\Delta a} \cdot \mathbf{\Delta \mu}}{|\mathbf{\Delta a}||\mathbf{\Delta \mu}|} = \cos < \mathbf{\Delta a}, \mathbf{\Delta \mu} > \ \geq \dfrac{\pi}{4} \end{cases}$$

where S_i is the set of music produced by artist A_i during the revolution.

And then we sort the selected artists by their influence (i.e., PR(A_i)) and get Table 7.

Table 7: Revolutionists during 3 Major Musical Revolutions

Stage of Revolution	1963-1969	1976-1982	2012-2018
Revolutionary Artist 1	The Beatles	The Clash	Usher
Revolutionary Artist 2	The Rolling Stones	Michael Jackson	Pharrell Williams
Revolutionary Artist 3	Jimi Hendrix	Blondie	Britney Spears
Revolutionary Artist 4	The Kinks	Donna Summe	Beyoncé
Revolutionary Artist 5	Led Zeppelin	The Jam	Rihanna
Revolutionary Artist 6	The Who	The Police	Miguel

Our findings reflect the major revolutions in music history. For example, in the 1960s, music saw the evolution of rock genre into a **more electric variant**. The rock genre also diversified into subgenres such as **beat and blues rock**. As a result, music styles became more varied in the 1960s. The Beatles, for example, were known for popularizing psychedelic rock, while The Rolling Stones notably assimilated numerous musical genres into their work.

In addition, music was heavily influence by the multitude of current events in the US, such as the Vietnam war, the Cold War and the Civil Rights Movement. Music became an outlet for expressing hope for social changes, such as in John Lennon's "Imagine". This trend may have resulted in a **great change of valence and energy** in music.

Our model highlighted another period of great change in 1976-1982. This likely reflects the rise of disco music into prominence in the 1970s, followed quickly by its fall in popularity in the 1980s.[11] On the other hand, the 1980s revolutionized the music industry with the rising use of synthesizers, with electronic genres such as Eurodance increasing in popularity. The rapid change in the dominant genre, combined with the incorporation of more electronic elements into music, corroborates with the period of **higher music variation** identified by our model. The development of punk rock also matches the revolutionaries we identified in this period, with **Blondie and The Clash** being one of the first acts to achieve fame in both the US and the UK.

The third music revolution is largely representative of the explosion of the Internet and social media. The rise of streaming platforms such as Spotify has **lowered entry costs** into the music industry, enabling independent artists' music to reach a larger audience; widespread instantaneous sharing on social media also garnered supporters of smaller artists' work. These enabled a larger number of artists to thrive and thus produce a **greater variability** of music.

Compared to previous revolutions where a new genre rose into prominence, this revolution is marked by the **lack of homogeneity**, which fulfils the increasing music variation criterion in our model. Our revolutionaries also match the real-world most influential artists, with Pharrell Williams' "Happy" as the most-played song and Rihanna as the artist with the most number-one singles in the decade.

10 Interplay between Music and Society

In the previous section, we explored how our model can identify revolutions in the music history, and investigated the driving sociopolitical factors behind each revolution. In this section, we will take a brief look at how music and society can influence and change each other.

We plot the number of songs in *full_music_data* for each year against year in Figure 11. As can be observed, there is a boom in the number of music produced in the late 1940s, which plateaued in around the 1960s. Music production can be a good indicator of quality of life: as the basic needs of life have been catered to, there can be more resources diverted to leisure and artistic production, such as music.

This sharp rise in music production thus corresponds to a large rise in world welfare, which likely reflects the post-World War II **economic expansion**, lasting roughly from 1950 to 1970. As economies boomed, the demand for music rose rapidly.

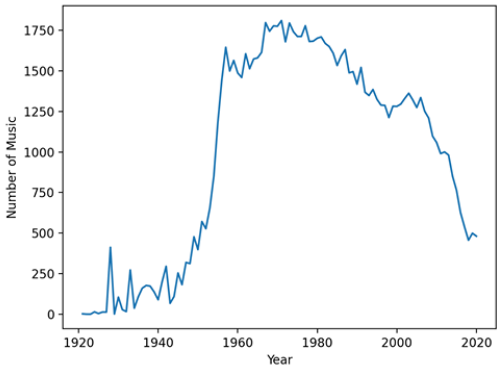

Figure 11: Number of Music over Time

Additionally, our analysis in the previous section also demonstrates how major societal changes, such as the advent of social media, can be reflected in the data. This confirms our hypothesis that changes in music is closely related to socioeconomic factors.

Interestingly, the number of music plummeted shortly after it peaked; this may be explained by an incomplete inclusion of data in the given data set, with proportionally much less music recorded in the later years. This is the most likely explanation, as it is highly unlikely that the number of music produced each year dropped to 1950s levels in the 2010s.

11 Sensitivity Analysis

We now test the sensitivity of the damping factor d in the PageRank algorithm. As shown in Fig. 12, when we change d from **0.1 to 0.9**, the rank of The Beatles (highest rank when $d = 0.5$), Trinidad Cardona (lowest rank when $d = 0.5$), Jackie DeShannon, and It's a Beautiful Day (two arbitrarily chosen artists) don't vary significantly; there is a steady decrease in the ranking of Vikki Carr (who ranked $5568/2 = 2784$th place when $d = 0.5$). Vikki Carr has been active since 1962. Therefore, as d grows larger, her influence is less "dampened," and her PageRank increases accordingly.

There is also steady decrease in the ranking of Vikki Carr (ranked 2784th when $d = 0.5$). This is because Vikki Carr has been active since 1962. As d grows larger, her influence is less "dampened," and her PageRank increases accordingly.

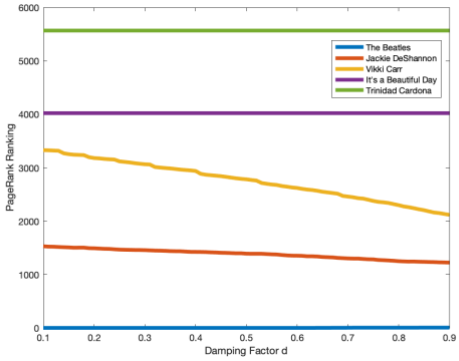

Figure 12: Changes of PageRank Rankings of 5 Artists against Damping Factor d

12 Conclusion & Evaluation

Conclusion

- The directed graph and PageRank established The Beatles, Cab Calloway, Bob Dylan, Louis Jordan, and Billie Holiday as the top five influencers in the musical evolution.
- The designed distance function demonstrates that influencers do have effects on their followers.
- Some characteristics are more contagious than others. Energy and danceability are the most contagious characteristics while tempo and key are the least.
- Major leaps in the music industry took place around 1966, 1979, and 2015. Revolutionary artists include The Beatles & The Rolling Stones (1966); The Clash & Michael Jackson (1979); Usher & Pharrell Williams (2015).
- Major music revolutions occurred in the 1960s, 1970s and 2010s, driven by the evolution of rock music, popularization of electronic music and the advent of the Internet respectively.

Strengths

- Cleaned the data by reshaping the distribution, which reduces sensitivity and prevents overfitting to a small number of characteristics.
- Optimized the weight parameters instead of assigning weights subjectively.
- Designed highly generalizable models applicable to other industries.
- Impressive and informative visualizations: radar chart, heat map, directed graph, etc.
- Adopted interdisciplinary approach: performed quantitative analysis and contextualized findings in social, economic, and historical backgrounds.

Weaknesses

- Limited scope: our analysis mainly focuses on music pieces after 1960
- Subjectivity of parameters: not every parameter is optimized.
- Biased representation: Music in other regions are underrepresented in our analysis.
- The assumption that an artist produced similar music may not be well justified.

References

1. Augmented dickey–fuller TEST. (2020, July 03). Retrieved February 08, 2021, from https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test
2. Clustering distance measures. (2018, October 20). Retrieved February 08, 2021, from https://www.datanovia.com/en/lessons/clustering-distance-measures/
3. Cosine similarity. (n.d.). Retrieved February 08, 2021, from https://www.sciencedirect.com/topics/computer-science/cosine-similarity
4. Knees, P., & Schedl, M. (n.d.). Music Similarity and Retrieval. Retrieved February 8, 2021, from http://www.cp.jku.at/people/schedl/Research/Publications/pdf/knees_sigir_2013_tutorial.pdf
5. A large-scale evaluation of acoustic and subjective music ... (n.d.). Retrieved February 8, 2021, from https://www.cs.swarthmore.edu/~turnbull/cs97/f09/paper/berenzweig04.pdf
6. Neapolitan, R., & Jiang, X. (2007, September 05). Collaborative filtering. Retrieved February 08, 2021, from https://www.sciencedirect.com/science/article/pii/B9780123704771500281
7. Peter Knees, & Markus Schedl. (2013, December 01). A survey of music similarity and recommendation from music context data. Retrieved February 08, 2021, from https://dl.acm.org/doi/10.1145/2542205.2542206
8. Ross, E. (2017, April 18). How songs with explicit lyrics came to dominate the Billboard Hot 100. Retrieved February 08, 2021, from https://www.newsweek.com/songs-explicit-lyrics-popular-increase-billboard-spotify-583551
9. Saaty, R. (2002, May 02). The analytic hierarchy process-what it is and how it is used. Retrieved February 08, 2021, from https://www.sciencedirect.com/science/article/pii/0270025587904738
10. Social network analysis in Python. (n.d.). Retrieved February 08, 2021, from https://www.datacamp.com/community/tutorials/social-network-analysis-python
11. Solutions, R. (n.d.). Disco. Retrieved February 08, 2021, from https://web.archive.org/web/20101025103728/http://allmusic.com/explore/essay/disco-t2151
12. Stephanie. (2020, September 21). Box cox transformation. Retrieved February 08, 2021, from https://www.statisticshowto.com/box-cox-transformation/

Document to the ICM Society

Dear ICM President,

Greetings. We are writing to report on our methodologies and findings in our research on the factors influencing music creation, based on the data provided to us by the ICM.

To establish the most prominent influencers in the data given, we implemented the *PageRank algorithm* to rank the artists, based on the number of followers they have, and the influence their followers have achieved. By considering the artists' genres, we investigated how much artists within any given genre influence all artists – we found that artists are most influenced by artists of the same genre, and that pop/rock and jazz artists are widely influential.

In our following work, we measured how similar music are by their characteristics. To meaningfully do this, we had to attach relative importance to each of the characteristics, and we used the *Analytical Hierarchical Process* to ensure that we are consistent in this process.

We based our measurement of music similarity on the assumption that an artist will produce similar music. With a notion of music similarity, we could then investigate more complex problems, such as if and how artists emulate their influencer's styles in their own work.We found that artists are most similar to their influencers in terms of the energy of the music they produce.

To identify major revolutions in music development, we looked into years in which music styles became much less similar to music some years ago. From this, we identified three revolutions in the 1960s, late 1970s and the 2010s, led by the likes of The Beatles, The Clash and Usher respectively. Smaller revolutions can also happen on the genre level: we found interesting changes in R&B music, which we linked to historical development of the genre.

In the end, we were able to establish interesting links between our data and actual socioeconomic development. For example, we discovered that the spread of the Internet has encouraged significant variability in music, and that the post-World War II economic boom fostered a time of extensive music creation. This supports our intuition that music development is closely related to major social, economic and political changes, and may even mutually influence each other.

However, we were constrained by the datasets with limited genres. This prevented us from measuring more accurately the influence of some novel genres such as trap music; some of the genres could also be split into subgenres as their development diverges over time, for example EDM and vaporwave are unique genres that branched from the electronic genre. With a richer data, we could see how subgenres branched, diverged, and influenced each other.

Going from here, we can investigate further into global music development. By building on our present methodologies, we can then study how music spreads across linguistic and geographical boundaries, and their interplay with each other and socioeconomic developments.

Yours Sincerely,
Team 2104738