UM–SJTU Joint Institute

Computational Methods for Statistics and Data Science (STAT4060J)

# Final Draft

## The Influence of Music

Group 25


Name: Baole Fang ID: 518370910056
Name: Kaiyang Chen ID: 518021910962
Name: Yumeng Bai ID: 518370910161


Date: August 4, 2021

# 1 Introduction

Music has become one of the most important parts in human society. There have been countless arguments about who is the most influential musician. In order to answer that question and to have a better understanding about how music evolves over time, we can start by analyzing the influence of musicians [3]. With data describing the influence network of musicians and the musical features of songs, our research focuses on how to rank musicians by their influence and how to verify it. Therefore, we need to answer these questions in this paper:

- How to rank musicians by their influence?

  As mentioned above, ranking musicians by their influence can not only help settle the arguments about who is the most influential musician from a mathematical view, but also will help people understand music evolution.

- What do musical features tell us about songs?

  Is the energy positively correlated to loudness? In other words, if a song is more energetic, is it louder? Answering this question can help people understand the characteristics of music better. If someone wants to choose some songs suitable for dancing, then he may want to choose some energetic songs. Another possible question that worth discussing is whether a more acoustic song (without technology enhancements) is more popular. If so, then it will help musicians write songs that are more popular.

- How can we use musical features to develop a measure of musical similarity?

  Musical similarity is useful because people can use it to classify songs into different genres. Besides, it plays an important role in music recommendation system. If a user likes a particular song, then he is more likely to enjoy another song that is similar to that song. More importantly, the measure of musical similarity serves as the foundation for the next question.

- Are the songs of more influential musicians and their followers more similar than those of less influential musicians and their followers?

  Although people assume that musicians that have been influenced tend to produce songs similar to those of the influencers. However, is it actually true? If not, then it means that followers are trying to differentiate themselves from their influencers. This may be one of the reasons why musics can evolve.

First, we propose two methods to rank musicians by their influence and evaluate their relevance using permutation tests. The first method is based on the number of followers that a musician have (quantity of influences). If one musician has more followers, then it means that he is more influential. The second method is based on a page rank algorithm. It not only considers the quantity of influences, but also takes the quality of influences into account. There is a parameter $\alpha$ in the second method. Permutation tests on ranks are carried out to determine whether these two methods give dependent results under different $\alpha$.

Second, we analyze the dependence between different features. For those dependent features, further experiments, such as shift/scaled independence test and linear regres-

sion, are carried out to find out how they are dependent. Further implication about the characteristics of musical features can be deduced from those test results.

Third, we propose a method to calculate the similarity of different songs. One possible way to verify its validness is to show that musicians within the same genre are more similar than those between different genres. Therefore, a permutation test is carried out to prove it. With the measure of musical similarity, we are prepared for the final question.

Fourth, we verify whether more influential musicians have a stronger impact on their followers in terms of similarity. All musicians are divided into two groups, high-influential group and low-influential group, by their influence rankings obtained from the first method. Then, stratified bootstrapping is used to test whether the songs between more influential musicians and their followers are more similar than those of less influential ones.

With these four methods [4], we are able to solve the questions mentioned above and help people understand the influence of music better.

# 2 Data

For the purpose of developing a model that measures musical influence and examining evolutionary trend of artists and genres, we scraped our data from AllMusic.com [1] and Spotify's API [2]. The obtained data are described in following files:

- "***influence_data***" (from ALLMusic.com) represents musical influencers and followers, as reported by the artists themselves, as well as opinions of industry experts. These data contains influencers and followers for 5,854 artists in the last 90 years. It provides the information about the active year of the influencers and followers as well as their belonged genres. e.g.

| influencer_name | influencer_main_genre | influencer_active_start | follower_id | follower_name | follower_main_genre | follower_active_start |
|---|---|---|---|---|---|---|
| The Exploited | Pop/Rock | 1980 | 74 | Special Duties | Pop/Rock | 1980 |

- "***full_music_data***" (from Spotify's API) provides 16 variable entries, including musical features such as *danceablity*, *tempo*, *loudness*, and *key*, along with *artist_name* and *artist_id* for each of 98,340 songs. It provides information about the basic information and feature statistics for each song. e.g.

| artist_names | artists_id | danceability | energy | valence | ... | release_date | song_title (censored) |
|---|---|---|---|---|---|---|---|
| Fat Freddy's Drop | 178301 | 0.6 | 0.365 | 0.131 | ... | 2005 | Ernie |

Among all the feature statistics that we have (artist_names, artists_id, danceability, energy, valence, tempo, loudness, mode, key, acousticness, instrumentalness, liveness, speechiness, explicit, duration_ms, popularity, year, release_date, song_title (censored)), some are hard to understand simply from a single discription word, i will pick several confusing feature and explain.

- **valence**: A measure describing the musical positiveness conveyed by a track. A value of 0.0 is most negative and 1.0 is most positive. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry). (float)

- **tempo**: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. (float)
- **mode**: An indication of modality (major or minor), the type of scale from which its melodic content is derived, of a track. Major is represented by 1 and minor is 0.
- **speechiness**: Detects the presence of spoken words in a track. The more exclusively speech- like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks. (float)

In order to conduct our research in a more convenient manner, we cleaned up the raw data (integration by year or artist and normalization of feature statistics) and generated the following directly usable files:

- "***data_by_artist***" is concluded from "*full_music_data*". It provides the mean values of features from every recorded songs of certain artist. e.g.

| artist_names | artists_id | danceability | energy | valence | ... | popularity | count |
|---|---|---|---|---|---|---|---|
| Frank Sinatra | 792507 | 0.384 | 0.238 | 0.364 | ... | 26.004 | 1369 |

- "***data_by_year***" is concluded from "*full_music_data*". It provides the mean values of features from every recorded songs of certain year. e.g.

| year | danceability | energy | valence | tempo | ... | duration_ms | popularity |
|---|---|---|---|---|---|---|---|
| 1921 | 0.426 | 0.237 | 0.425 | 100.4 | ... | 229912 | 0.35 |

- "***music_genre***" attached every song in "*full_music_data*" with their according genre. There are 20 genres in our dataset in total, however, we only use 18 out of it (children and unknown are excluded because of insufficient sample size). e.g.

| artist_names | artists_id | danceability | energy | valence | ... | release_date | song_title (censored) | Genre |
|---|---|---|---|---|---|---|---|---|
| Fat Freddy's Drop | 178301 | 0.6 | 0.365 | 0.131 | ... | 2005 | Ernie | electronic |

Below is the visualization of music features with respect to popularity:

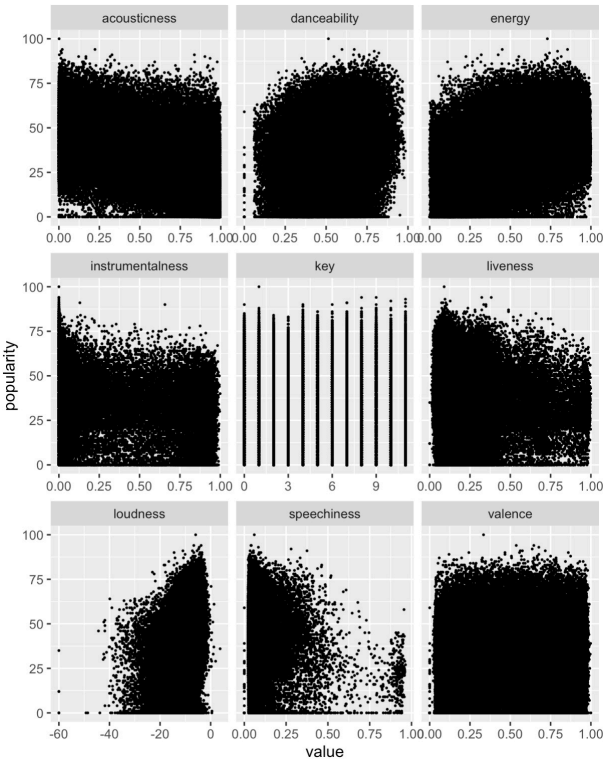


Figure 1: Music Features with Respect to Popularity

The statistics in above data is not well-prepared for our further manipulation, we perform a Min-Max Normalization for every data points in our dataset and made their range between [0,1] with the equation:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

From the normalized data, we can freely explore the possible relationship between musics of different genres or musicians and get a glimpse of the evolution in music. For example, we can easily visualize the influence network of depth 10 which root is *Coldplay*.

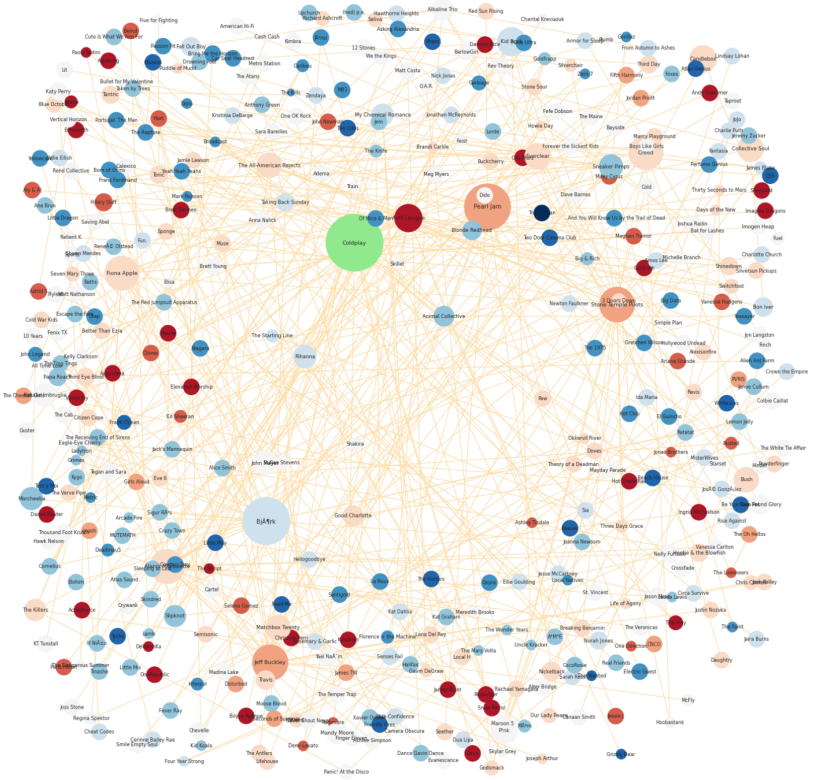


Figure 2: Influence Network of Depth 10 (Coldplay)

More and precise usage of data will be explained further in the following sections.

# 3 Methods

## 3.1 Ranking algorithms and their relevance

In order to better understand the influence of musicians as well as the evolution of music, we first try to rank the influence of musicians. To start with, two different ranking methods are carried out.

First, we quantify the influence in the dataset "***influence_data***" by simply counting the number of followers that a musician have. This method is intuitive because if one musician has more followers, we tend to think that he is more influential.

In the second method, we aim to take the quality as well as the quantity of influence into consideration. This method is based on the PageRank algorithm [5] developed by Google, which is a link analysis algorithm that is used to rank web pages in the search engine results. The underlying assumption of PageRank is that more important websites are likely to receive more links from other websites. In our case, the assumption holds true that more influential musicians have larger chances to have more followers. Hence, we can apply PageRank algorithm to rank the influence of musicians as follows.

Let $u$ be a musician. Then let $I_u$ be the set of influencers of $u$ and $F_u$ be the set of followers of $u$. Let $N_u = |I_u|$ be the number of influencers of $u$ and let $\alpha$ be a factor used for normalization. Let $N$ denote the total number of musicians. Then, we can obtain the

rank $R$ by the general PageRank algorithm as:

$$R(u) = \alpha \sum_{v \in F_u} \frac{R(v)}{N_v} + (1 - \alpha)\frac{1}{N}$$

Next, we perform permutation tests for the above two influence ranking methods to determine whether they give dependent results under different value of $\alpha$ in the second method.

Permutation test is a non-parametric test, in which we calculate a large number of possible values of the desired test statistic under all possible rearrangements of the observed data points to obtain an empirical distribution of the test statistic for the null hypothesis. One of the main advantages of permutation test is that it can be applied for any test statistic, regardless of whether its distribution is known.

To analyze the relevance between the simple ranking and page ranking with specific parameter for the influence of musicians, we conceptualize the results of two rankings of total $n$ musicians as the following two samples:

$$\text{Simple Ranking: } X_1, X_2, \ldots, X_n$$
$$\text{Page Ranking: } Y_1, Y_2, \ldots, Y_n$$

In order to test whether the two samples are dependent, we first assume that:

$$Y_i = \beta_0 + \beta_1 X_i + R_i$$

with $\beta_0$ and $\beta_1$ as main variables, and $R_i$ as an unobserved, latent variable, which is assumed to be independent of $X_i$. Then we set the null hypothesis as: $\beta_1 = 0$. Hence, if the null hypothesis is rejected, we can conclude that the two ranking methods are dependent, otherwise, they are independent.

## 3.2 Feature analysis

With the purpose of exploring the evolution of music and researching the ranking and relationship among musicians, it is important to know the correlation of the music features in order to carry out further analysis.

First, in our data "***full_music_data***", we have feature statistics for every recorded songs. We can easily perform permutation test that used the correlation between two features to test the null hypothesis that these features were independent. By comparing every feature pairs, an independence table between features are generated.

Then, we filter out those independent pairs and try to explore how is the remnant pair dependent. We will try shift/scaled independence test or linear-regression and found possible explanation for their correlation.

## 3.3 Measure of musical similarity

To prepare for the fourth question, a measure of musical similarity is purposed in this method. We use the reciprocal of the Euclidean distance of two songs/musicians in feature space as their similarity:

$$similarity(x, y) = \frac{1}{\epsilon + \sum_i \sqrt{(x_i - y_i)^2}}$$

where $x$ and $y$ are two songs/musicians, $x_i$ and $y_i$ indicate the $i^{th}$ feature of $x$ and $y$, and $\epsilon$ is a small number that avoid zero denominator (0.01 in this case). It range is $(0, 100]$.

In order to show that this measurement is valid, we made an assumption that songs/-musicians within the same genre are more similar than those between different genres. One implication of this assumption is that music genres are dependent of musical features. Therefore, if we can reject the null hypothesis that musical similarity is independent of genres, the validness of this measure of similarity is verified.

To begin with, a K-Means clustering is used to group all musicians into 18 groups, which is the number of genres. The reason we choose K-Means clustering is that both K-Means and the measure of musical similarity are based on the Euclidean distance of two songs/musicians. A permutation test based on K-Means is carried out to test whether musicians within the same genre are more similar than those between different genres. Here, we introduce a test statistic called the maximum clustering coefficient (MCC). MCC is determined by the cluster that has the most musicians divided by the total number of musicians in that genre:

$$MCC(g) = \frac{|C_g|}{|X_g|}$$

where $g$ is genre, $X_g$ is the set of musicians in genre $g$, $C_g$ is the largest cluster in $X_g$. It is derived as follows:

1. Run K-Means to musicians of 18 genres into 18 groups
2. For each genre $g$, find the set of musicians ($X_g$) belong to genre $g$
3. In the set $X_g$, find the largest cluster $C_g$, and treat it as the predictor of $X_g$

MCC is a coefficient between 0 and 1. It describes the proportion of musicians in one gnere explained by the clustering. The higher MCC is, the better the model describes the genres.

The null hypothesis is that the similarity of two songs is independent of their genres. So, the null distribution of MCC can be obtained by permuting the genre label of all songs. If the observed MCC of each genre lies above the confidence interval of the generated MCC, then it means that songs are similar within the same genre than between different genres. If we can reject the null hypothesis, it means that the measure of musical similarity is reasonable.

## 3.4 Influence and similarity

Knowing how to measure the similarity between different musicians, we can test whether more influential musicians have a stronger impact on their followers in terms of musical similarity. In other words, we want to test whether the songs of high-influential musicians and their followers have a higher similarity than those of less-influential ones. The high-influential musicians are composed of the top 50% musicians in the influence ranking and the low-influential musicians are composed of the bottom 50% musicians in the influence ranking. In order to test this statement, we use a hypothesis testing.

The statistic we use is called the similarity score (SS). The similarity score (SS) of a

musician is calculated as

$$SS(x) = \frac{1}{|S(x)|} \sum_{y \in S(x)} similarity(x, y)$$

where $S(x)$ is the set of followers of $x$. Since the range of $similarity(x, y)$ is $(0, 100]$, the range of $SS(x)$ is also $(0, 100]$.

The null hypothesis is that there is no difference in similarity score (SS) between the high-influential musicians and low-influential musicians. We can use a stratified bootstrap to generate a 95% confidence interval for the difference of means for the similarity score (SS) in these two groups. If it contains 0, then we can accept the null hypothesis and argues that influence and similarity do not have a close causal relation. If the confidence interval is bigger than 0, then it means that high-influential musicians tend to leave a larger impact on their followers.

# 4 Simulations

## 4.1 Ranking algorithms and their relevance

### 4.1.1 Two methods of Ranking

We propose two different methods to rank the influence of musicians. For the first one, we simply count the number of followers of a musician as the index of his/her influence and rank. In the second method, we use the PageRank algorithm, which not only considers the influence of the musicians to their followers but also the influence of a musicians' followers to others. Several different values of the parameter $\alpha$ in PageRank algorithm are simulated.

Take $\alpha = 0.5$ as example, the top ten musicians calculated by the PageRank algorithm are shown as follows in Table 1. In addition, the out degree listed in the table represents the number of followers, which is the index of influence in the first method. Also, the influence score is the rank value $R$ mentioned in Section 3.1.

| Name | Rank | Influence Score | Out degree (Followers) | In degree (Influencers) |
|---|---|---|---|---|
| The Beatles | 1 | 1 | 615 | 31 |
| Cab Calloway | 2 | 0.6027 | 27 | 2 |
| Bob Dylan | 3 | 0.6023 | 389 | 29 |
| Louis Jordan | 4 | 0.5924 | 53 | 2 |
| Billie Holiday | 5 | 0.5680 | 106 | 2 |
| Lester Young | 6 | 0.4918 | 51 | 1 |
| The Rolling Stones | 7 | 0.430610887 | 319 | 39 |
| Roy Acuff | 8 | 0.424844878 | 50 | 0 |
| Hank Williams | 9 | 0.422921312 | 184 | 3 |
| Woody Guthrie | 10 | 0.383834005 | 80 | 0 |

Table 1: Top ten musicians in PageRank Algorithm with $\alpha = 0.5$.

According to the table, we can directly see that the PageRank algorithm is not very similar to the first simple ranking method. For example, for Cab Calloway only has 27

direct followers, but ranks the second among thousands of musicians. Further analysis will be conduct in Section 5.1.

### 4.1.2 Permutation Test for dependence of two ranking methods

We use the permutation test to find out whether the two ranking method is relevant to each other. As mentioned in Section 3.1, we have the following assumptions.

$$\text{Simple Ranking: } X_1, X_2, \ldots, X_n$$
$$\text{Page Ranking: } Y_1, Y_2, \ldots, Y_n$$
$$Y_i = \beta_0 + \beta_1 X_i + R_i$$

And the null hypothesis for the test is that the two rankings are independent of each other, i.e, $H_0 : \beta_1 = 0$ with p-value as 0.05. The test statistic selected for the permutation test is the Pearson correlation coefficient as follows:

$$\rho_{xy} = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

with $Cov(x, y)$ as the covariance of the two variables, and $\sigma$ as the standard deviation of each variable.

Then the permutation test is conducted, and a two-tailed p-value for the test statistic is computed, which results in 0. Take the situation when the parameter of PageRank method $\alpha = 0.5$ as example, the corresponding distribution is shown below in Figure 3.

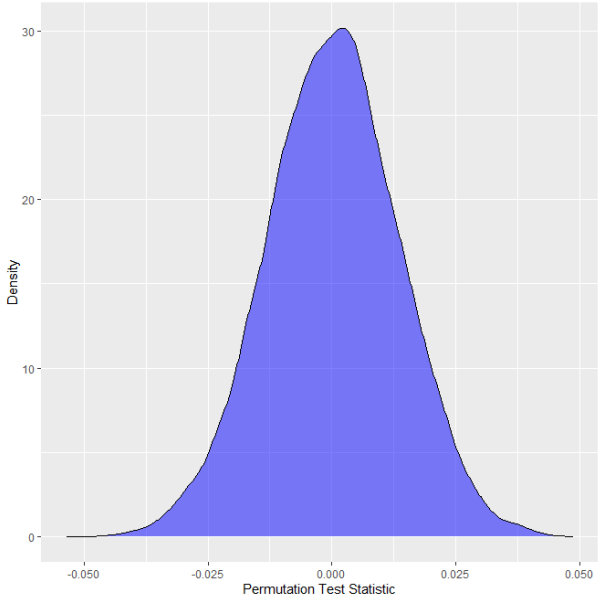


Figure 3: Permutation test with $\alpha = 0.5$

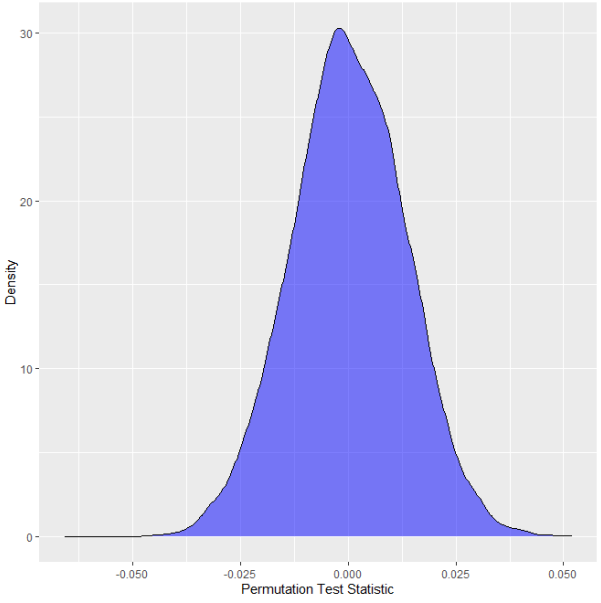


Figure 4: Permutation test with $\alpha = 0.9$

With enough evident, we can reject the null hypothesis that $H_0 : \beta_1 = 0$. Hence, we can draw the conclusion that the two ranking methods are not independent.

We also simulate different values of the parameter $\alpha$ of PageRank method, and find that we can always get a zero for two-tailed p-value, and reject the null hypothesis. The

distribution obtained for $\alpha = 0.9$ is shown in Figure 4. We can see that although the general shape slightly changes, the final result is not affected by the change in $\alpha$. In addition, the permutation test has a good performance in generated data, which results in a big power.

## 4.2 Feature analysis

First, in order to determine whether two features are independent, instead of two data samples, we use a single sample of pairs $(W_i, Z_i)$, i = 1, ..., n, IID where W and Z are one of the symbols. Then we perform permutation test on these pairs under the hypothesis that $F_{WZ} = F_W F_Z$ (independent), we arbitrarily permute all the Z values and any statistic T(W,Z) would have the same distribution. The test statistic here is the correlation between two feature samples because sample correlation is a way to summarize the linear relationship between two variables and we compute a 2-tailed p-value for the test statistic. The null hypothesis is that two features are independent of each other, and if the 2-tailed p-value is below 0.05, we have strong confidence to reject the null hypothesis, that is two feature might be dependent. The result p-value between each pair is shown below:

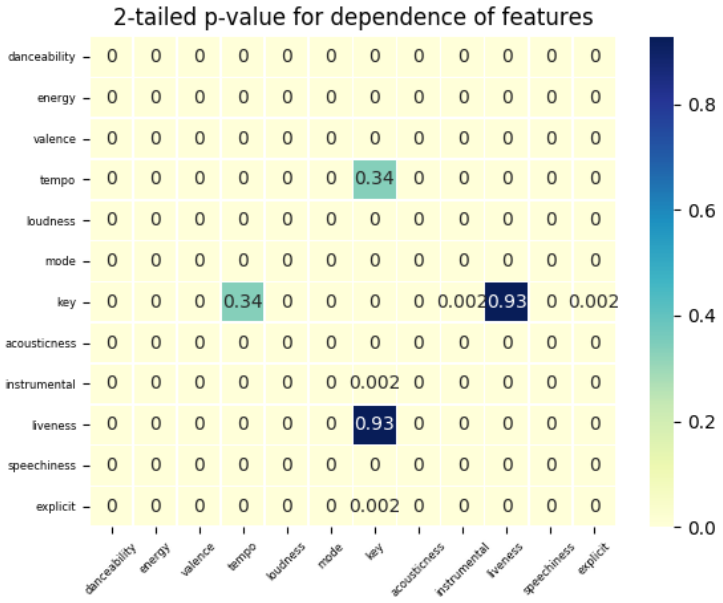


Figure 5: P-value for Feature Pairs

Then, among all the dependent feature pairs, we delve further to study their specific relation by picking the most representative pairs, that is "energy" and "loudness". We first run a linear regression on $(Energy_i, Loudness_i)$ data pairs and get the following relation:

$$\hat{E} = 2.613L - 1.484$$

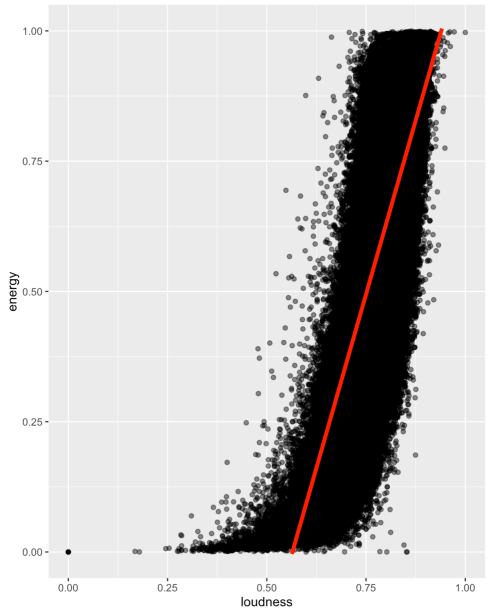


Figure 6: Linear Fit for Energy and Loudness

Although the R-square of the fit is only 0.617, we try to measure the similarity of Energy and $\hat{E} = 2.613L - 1.484$ using the two sample permutation test. It's convenient to think about a combined sample of the form:

$$(W_i, Z_i), i = 1, ..., 2n$$

where

$$W_i = \begin{cases} Energy_i, i <= n \\ Loudness_i, i > n \end{cases}$$

and

$$Z_i = I(i \leq n)$$

The null hypothesis is that $H_0 : E = \hat{E}$, and the test statistic is

$$T(Z_1, ..., Z_{2n})$$

where T(Z) is the difference of means(Welch's permutation t-test). Under the null hypothesis, group labels are uninformative, a permutation test will permute the $Z_i$ to get a conditional distribution for T. Let's consider

$$H_1 : E \neq \hat{E}$$

such that the distribution of T would be shifted if $H_1$ is true. And we compute the two-tailed p-value for it in order to determine whether we can reject $H_1$. The result shows that p-value is 0.

## 4.3 Measure of musical similarity

As mentioned before, musical similarity is calculated as:

$$similarity(x, y) = \frac{1}{\epsilon + \sum_i \sqrt{(x_i - y_i)^2}}$$

where $x$ and $y$ are two songs/musicians, $x_i$ and $y_i$ indicate the $i^{th}$ feature of $x$ and $y$, and $\epsilon$ is a small number that avoid zero denominator (0.01 in this case).

To verify its validness, we can test whether musicians within the same genre are more similar than those between different genres. The null hypothesis is that musical similarity is independent of genres. The test statistic we use is MCC. If the observed MCC of each genre lies above the confidence interval of the generated MCC, we can reject the null hypothesis and prove the validness of our musical similarity measurement.

The observed MCC of each genre is plotted in Figure 7.

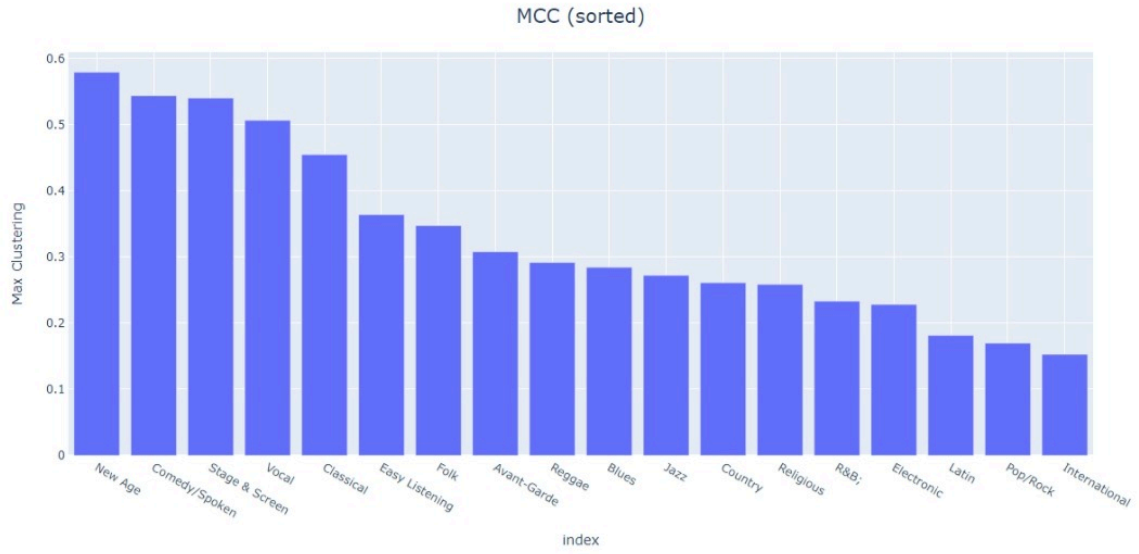


Figure 7: Observed MCC

Then, genre labels are permuted. The generated MCCs and their confidence interval are plotted together with the observed MCC in Figure 8.

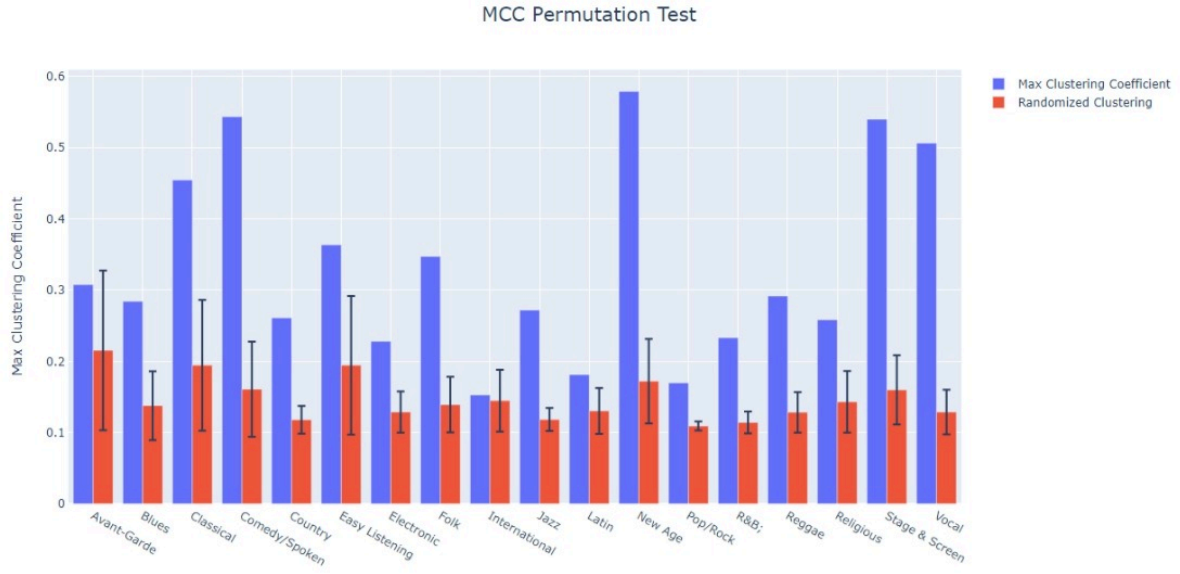


Figure 8: MCC permutation test

As we can see, the MCCs of all genres except Avant-Garde lie above the 95% confidence interval of the generated MCCs. Therefore, we can reject the null hypothesis and conclude that our measure of musical similarity is valid.

## 4.4 Influence and similarity

In order to test whether more influential musicians have a stronger impact on their followers in terms of musical similarity, we use a stratified bootstrap to generate a 95% confidence interval for the difference of means for high-influential musicians and low-influential musicians. Their similarity scores (SS) are plotted in Figure 9. The high-influential musicians are composed of the top 50% musicians in the influence ranking and the low-influential musicians are composed of the bottom 50% musicians in the influence ranking.

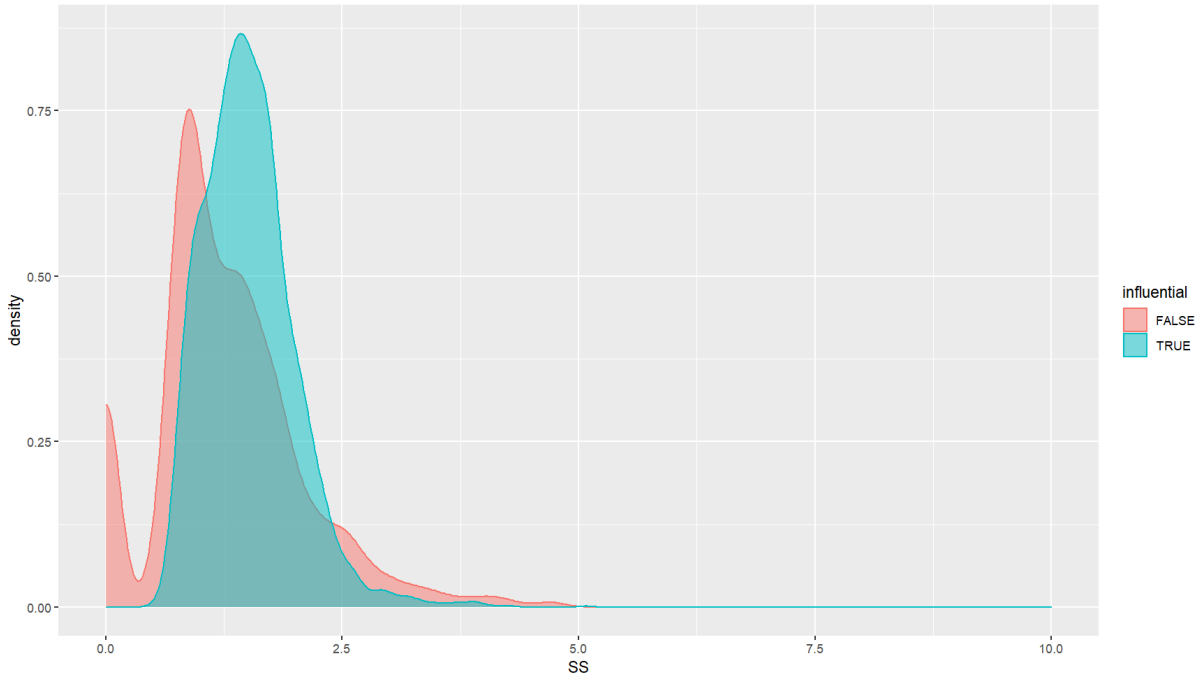


Figure 9: SS of high and low influential musicians

By running the stratified bootstrap, the 95% confidence interval is $(0.1827, 0.2880)$, which is high above zero. Therefore, it means that high-influential musicians indeed have a stronger impact on their followers.

# 5 Analysis

## 5.1 Ranking algorithms and their relevance

### 5.1.1 Two methods of Ranking

According to the ranking results, we have found that some of the ranking results of the two methods seem to be quite different from each other. Take Cab Calloway as example, he only has 27 direct followers, but ranks the second among thousands of musicians. To find out the reason, we visualize the direct followers and the followers of followers of Cab

Calloway as follows in Figure 10 and Figure 11. In the plots, we represent Cab Calloway himself with a red scatter point in the center of the graph, his direct followers with a purple scatter point. Then, all direct followers are linked with Cab Calloway with red lines, are other followers are linked with their influencers with purple lines.

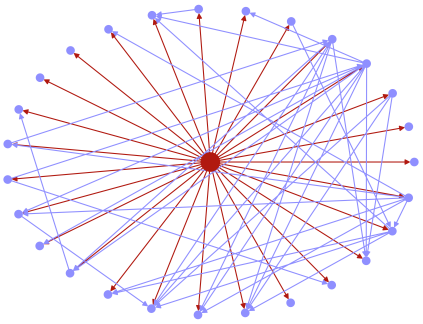

Figure 10: Direct Followers of Calloway

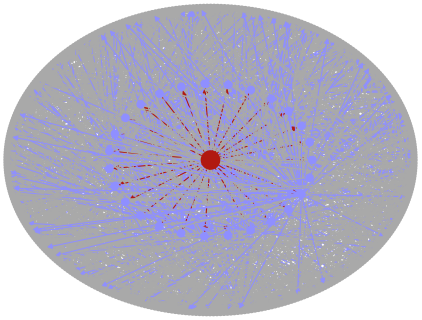

Figure 11: Indirect Followers of Calloway

We can find that also the number of direct red lines is not very large, however, there are a tremendous amount of purple lines, which indicates the huge indirect influence of Cab Calloway. Hence, the PageRank algorithm take both direct and indirect influence into consideration to give him a high ranking. If we simply use the number of direct followers to measure the influence of the musician as the first simple ranking method does, the rank of Cab Calloway will be much lower, which is not reasonable since he lies in the center of a highly-connected influential network as shown in Figure 9. Therefore, we can conclude that the PageRank algorithm is superior to the first simple ranking method under this circumstance.

### 5.1.2 Permutation Test for dependence of two ranking methods

From the previous results, we have found that the result that the two ranking methods are not independent does not depends on the value of PageRank parameter $\alpha$. Think about the similarities and differences between the two ranking methods, we think the result is reasonable, since in the PageRank algorithm, the number of out degrees, which are the followers, has an important contribution to the ranking score. Hence, the two methods are always dependent. In addition, we plot the data as scatter dots to see the relationship between the ranking score, which represents the ranking of PageRank method, and the number of out degrees, which represents the simple ranking. We choose two relatively extreme values 0.1 and 0.9 for $\alpha$. Then plots we got are shown below in Figure 12 and Figure 13.

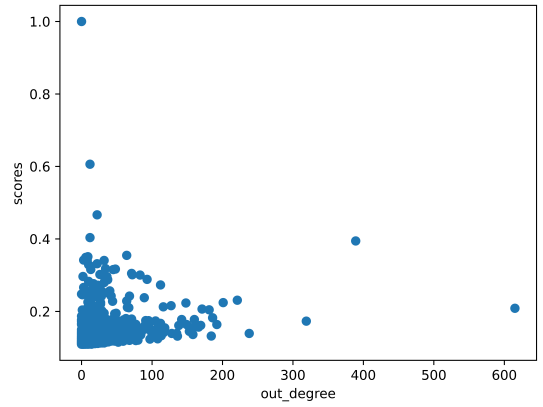


Figure 12: Ranking Score v.s Out Degrees with $\alpha = 0.1$

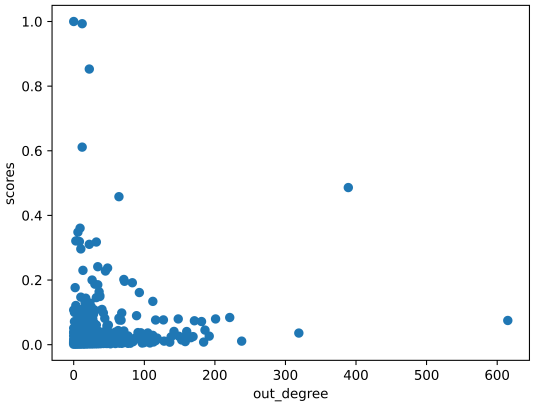


Figure 13: Ranking Score v.s Out Degrees with $\alpha = 0.9$

We can observe that most of the scatter gather in the left bottom of the plots. Hence, within the valid range for $\alpha$, there are always relationships between the two methods, which is in line with the results of our permutation tests.

## 5.2 Feature analysis

From the p-value results we get in simulation process, we found that most of the features are dependent to each others. The result is reasonable because the features of a specific song is dependent intuitively, for example, "energy" and "loudness", a song with bigger loudness is sure to convey more energy to their audience. However, it is noticeable that two of the feature pairs seems to be independent (because their p-value is bigger than 0.05 and we can not reject the null hypothesis that they are independent): "tempo" and "key", "liveness" and "key". Here, key is the estimated overall key of the track, integers map to pitches using standard Pitch Class notation. E.g. $0 = \mathrm{C}$, $1 = C\#/Db$, $2 = \mathrm{D}$, and so on. And tempo is the overall estimated tempo of a track in beats per minute (BPM). The independence between these two might be interpreted as the tempo of the song can be fast or slow regardless of the key, it i s without doubt that we have both fast and slow songs for the same key. As for the liveness and key pair, it is also understandable because liveness might be influenced by music genre, but there are songs of both high key and low key in one genre.

Then, we choose the dependent feature pairs "energy" and "loudness" to explore how are they related. While running linear regression give us the model $\hat{E} = 2.613L - 1.484$, the R-square of it is only around 0.62. Thus, we perform a two sample permutation test. The test statistic is the difference of mean and the hypothesis is $E \neq \hat{E}$. The resulting 2-tail p-value is 0. That means we can reject the hypothesis that two distribution is different (T is shifted). Accordingly, we can conclude that the distribution of Energy is closed to the distribution of loudness scaled by a factor of 2.613 and shifted -1.484.

What's more, we directly do linear regression for single feature with popularity. The results show that the popularity is not simply linearly related to any features (R-square is small). Also, from the relation scatter graph we shown in the data part, there is no linear relationship between popularity and features intuitively. The potential relationship

between popularity and features need to be further explored.

## 5.3 Measure of musical similarity

Musical similarity is measured as

$$similarity(x, y) = \frac{1}{\epsilon + \sum_i \sqrt{(x_i - y_i)^2}}$$

where $x$ and $y$ are two songs/musicians, $x_i$ and $y_i$ indicate the $i^{th}$ feature of $x$ and $y$, and $\epsilon$ is a small number that avoid zero denominator (0.01 in this case).

It comes from the idea of clustering. In clustering, we are able to group data points by their locations in feature space. The closer two samples are, the more similar they are. Therefore, we use the reciprocal of the Euclidean distance of two songs/musicians in feature space as their similarity.

Before verifying the validness of this measurement, we assume that songs/musicians within the same genre are more similar than those between different genres. One implication of this assumption is that music genres are dependent of musical features.

To verify its validness, we carry out a permutation test. Since the null hypothesis is that music genres are independent of musical features, we permute music genre labels to get the null distribution. From Figure 8, we can reject the null hypothesis that music genres are dependent of musical features. Therefore, the validness of our proposed musical similarity measure is verified.

From Figure 8, we can further tell that the gaps of observed MCC and generated MCC in some genres (eg. New Age, Stage & Screen, and Vocal) are larger than those of other genres. It implies that those genres (eg. New Age, Stage & Screen, and Vocal) have a stronger characteristics because songs/musicians are more similar within those genres. Therefore, when it comes to music recommendation, if a user already likes the songs/-musicians in those genres with high MCC, it is much safer and effective to recommend another song/musician in the same genre because they are highly similar to each other.

## 5.4 Influence and similarity

In this part, we analyze the relationship between musical influence and similarity. We want to answer whether high-influential musicians have a stronger impact on their followers. If there is a strong impact, then a high similarity between these two musicians should be high. Thus, the problem is transformed into whether the similarity score (SS) of high-influential musicians is higher than that of low-influential musicians. Therefore, we want to compute the confidence interval of the difference of means of SS between high-influential and low-influential musicians.

The method we use is called stratified bootstrap. Bootstrap is a resampling method that uses random sampling with replacement. The advantage of bootstrap is that it can compute measures of accuracy (eg. bias, variance, confidence interval, etc.) without knowing any prior information. It is done by computing $n$ statistics with $n$ resampling. With $n$ statistics, the measure of accuracy can be easily computed. Stratified bootstrap is a special bootstrap method that allows certain conditions. In our case, the condition is that we sample from two groups of musicians and take the difference of SS mean. It

is useful because we do not need to figure out the distribution of the difference to obtain the confidence interval, which is required if we simply compute the difference of SS mean on the original data.

From the simulation results that the 95% confidence interval $(0.1827, 0.2880)$ lies above zero, we can conclude that high-influential musicians indeed have a stronger impact on their followers. Another conclusion we can deduce is that although musicians are impacted by their influencers, they are still trying to create original musics. As we can tell from Figure 9, nearly no musicians have a SS higher than 5 (SS ranges from 0 to 100). It means that most musicians are trying to differentiate themselves from their influencers produce something new.

# 6 Discussion

Our project starts with an interesting question: Who is the most influential musician throughout history? To answer that question, we aim to analyze the influence of musicians.

With data describing the influence network of musicians and the musical features of songs, we mainly focus on the following four questions in our project:

- How to rank musicians by their influence?
- What do musical features tell us about songs?
- How can we use musical features to develop a measure of musical similarity?
- Are the songs of more influential musicians and their followers more similar than those of less influential musicians and their followers?

To start with, we first propose two different methods to rank musicians, one is to simply count the number of followers, the other one utilizes the famous PageRank algorithm developed by Google, which can jointly consider direct and indirect influence at the same time. Then we perform permutation tests, with correlation as the test statistic, on the two rankings to test whether they are independent to each other. We also change the value of PageRank parameter $\alpha$ to see whether it can affect the result.

To answer the second question, we perform permutation tests on different features of music to find the dependence among features. Then, with dependent feature pairs obtained, we perform shift/scaled independence tests to see how they are dependent.

To answer the third question, we propose the reciprocal of the Euclidean distance of two songs/musicians as a measure for musical similarity. In order to show that this measurement is valid, we made an assumption that songs/musicians within the same genre are more similar than those between different genres. One implication of this assumption is that music genres are dependent of musical features. Therefore, if we can reject the null hypothesis that musical similarity is independent of genres through a permutation test on a test statistic called MCC, the validness of this measure of similarity can verified.

To answer the last question, we use stratified bootstrap to compute the confidence interval of the difference of SS mean between high-influential and low-influential musicians. By investigating where 0 lies relative to the confidence interval, we can answer the last question.

After the simulation and analysis described above, our results and implications are summarized as follows.

- For the ranking of musicians, we successfully propose two ranking methods, and find that the PageRank ranking method is much superior to the simple ranking method under this circumstance. According to the permutation tests, the two methods are dependent to each other, regardless of the choice of parameter $\alpha$.
- For the dependencies among musical features, through the permutation tests, we conclude that most of the features are dependent to each other. Then we further study the relationship between dependent pair "energy" and "loudness", and find that they remain dependent to each other after the shift/scaled independence tests.
- In the third part, we manage to propose a measure of musical similarity, and verify its validness through MCC. We also find that some of the genres are highly similar to each other, which may be valuable for the recommendation system.
- For the last part, through stratified bootstrap, we can conclude that high-influential musicians have a stronger impact on their followers in terms of musical similarity. In addition, we find that most of the similarity scores(SS) are not high, this may implicate the importance of originality for musicians.

Despite the above results and contributions, there are still problems we have not solved yet. The followings are some interesting unsolved questions and implications about them based on our project.

- How to further verify the results of PageRank ranking? Maybe we can find some authoritative rankings and compare our ranking results with them.
- Since that most of the musical features are dependent with each other, why are several feature pairs being independent? To answer this question may require deeper understanding in music.
- We have found that originality plays an important role in music, so how can we define and identify innovators among musicians? Will they have a greater influence on others?
- Among all musical features, which one is most likely to be inherited or influenced? Why? To answer this question, stronger insights and knowledge in music are also needed.
- Whether a more acoustic song (without technology enhancements) or songs with other specific feature is more likely to become popular? This question is quite valuable for musicians, but we have not yet found the way to analyze the concrete relationship between popularity and features.

# References


[1] Allmusic. Available at https://www.allmusic.com/.

[2] Spotify. Available at https://www.spotify.com/us/.

[3] COMAP. 2021 ICM problem D: The influence of music. 2021. Available at https://www.comap-math.com/mcm/2021_ICM_Problem_D.pdf.

[4] Mark Fredrickson. STAT4060J slides. 2021.

[5] Larry Page, Sergey Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1998.

# Appendix

Please see detailed implementation on https://github.com/Kaiyang-Chen/stats-406-project.