Research on Intelligent Evaluation System of Influence Model Using Cosine Similarity

Zhitao Li¹, Xin Wang², *, Xiaoling Li¹
¹College of Mathematics and Statistics, Shenzhen University, Shenzhen, Guangdong, 518060
²WeBank Institute of FinTech, Shenzhen University, Shenzhen, Guangdong, 518060
*Corresponding author: 2018132124@email.szu.edu.cn

Abstract. Nowadays, music plays an increasingly important role in social development. As time goes by, music influences music, artists, and the times. The purpose of this paper is to create a model to quantify the influence of previously created music on new music and music artists. Firstly, this paper constructs two music influence networks based on artists and genres and defines and explains some meaningful parameters of music influence networks. The parameters of this paper describe the influence of communication between artists, and then this paper calculates and explains the parameters in the subnet. Next, this paper uses cosine similarity to measure similarity between two music samples and innovatively uses variance as the weight of music features to identify differences between music. Then the similarity measure is constructed based on matrix norm. Finally, this paper concludes that artists within schools are more similar than artists between schools.

1. Introduction.
Music is the essence of the long history. It has been heard and appreciated by millions of ears, and many souls have been baptized. After many rounds of polishing, mutual influence, and mutual achievement, music has become a treasure in human cultural heritage. As time goes by, music influences music, musicians, and the times. At the same time, they change because of them to achieve the sublimation of evolution.

To better understand the charm of music, this paper hopes to build a model to measure the influence of music and then explore the influence and development of music.

2. Music influence-oriented network
Music influence is mainly manifested in its influence on artists or music schools. In order to simplify the model, this paper divides music images into the following two types:

a) Influence between artists.

b) The influence of music schools.

As for the influence between artists, this paper constructs the communication network of artists' music's influence on artists.

Obviously, it is a directional network. In this network, each node represents an artist, and each directed edge represents the relationship between two connected artists. Through data preprocessing,
this paper finds that 251 artists are independent among 5854 pieces of data. That is to say, and they have neither influenced others nor been influenced by others. The goal of this paper is to explore the influence of artists. This small part of data does not contribute to the network model, so it is ignored. This paper constructs an artist influence network through a series of data processing and analysis, as shown in Figure 1, which has 5603 nodes and 42770 directed edges.

**Figure 1. Artist Influence Network**

In Fig 1, the size of a node indicates a degree—the greater his exposure, the more music artists he influenced. Through data cleaning, this paper found that every artist’s genre has not changed. Therefore, we use nodes of different colors to represent different genres. There are 20 music genres in the data set, but some music genres account for a relatively small proportion of the data due to the imbalance of data. To simplify the directional network and make it easier to identify, this paper classifies music genres that account for less than 2% of the data into one category and names them as other categories.

At the same time, this paper establishes an influence network among music schools.

**Figure 2. genre influences network**
In Fig 2, the color of nodes represents genres, the size of nodes represents the number of artists in genres, and the thickness of edges from genre A to genre B represents the appearance degree from genre A to genre B.

By consulting relevant literature, this paper finds that there are many indexes to capture the parameters of "music influence" based on network metrics, such as degree, centrality, clustering coefficient, density, intermediate degree, etc. [1]. After many attempts and comparisons, we have chosen the following five important indicators that best reflect the "music influence".

- **Degree Centrality $C_d$**
  Degree Centrality [2] is the most direct measurement of node centrality in network analysis. In the music influence network, we use the out-degree nodes to measure the centrality rather than the total degree. Because we define the influence of music as the influence on others, that is, the out-degree. In-degree means being influenced by how many music artists, which is not within the scope of our research. Hence we define that $d(i)$ is the degree of node i, $C_d$ is the degree centrality of the node. The formula is as follows:

$$C_d = d(i) \quad (1)$$

- **Closeness Centrality $C_c$**
  Closeness centrality [3] is the reciprocal of the shortest distance from a node to all other reachable nodes, which is normalized after accumulation. Closeness Centrality can measure the distance between an artist and all other members in the music influence network. The closer the distance is, the more influential the artist is. The formula is as follows:

$$C_c(i) = \frac{1}{d(i)} \cdot \frac{N}{\sum_{j=1}^{N} D_{ij}} \quad (2)$$

where $D_{ij}$ represents the shortest cell path from node i to node j, N is the total number of points, $d(i)$ is the average distance from node i to all nodes in the network.

- **Eigenvector Centrality $C_e$**
  The basic idea of Eigenvector Centrality [4] is that the centrality of a node is a function of the centrality of adjacent nodes. In our network, the centrality of eigenvectors can represent the transmission influence of artists on others, that is, the more important an artist’s influencers are in the network, the more important the artist is. The Google Pagerank Algorithm [5] is also a variety of feature vector centrality.

$$C_e(i) = x_i = c \sum_{j=1}^{n} a_{ij} x_j \quad (3)$$

where, $x_i$ is the important measure of node i, c is a scale constant, $a_{ij}$ is the adjacency matrix of the network.

- **Network Density $D_n$**
  Network Density [6] refers to the ratio of the number of actual connections to the number of possible connections in an informal network. Network density can be used to measure the closeness of music influence among network members. The density of a network with N nodes and L actual connected edges can be expressed as:

$$D_n = \frac{2L}{N(N-1)} \quad (4)$$

- **Second-order Degree $S_d$**
  To measure the transitive influence of a single node, we use the total output of a musician’s influencer set plus the size of the musician’s output. The formula is as follows:
\[ S_d(i) = d(i) + \sum_{j \in \text{link}(i)} d(j). \]

Where \( d(i) \) is the degree of node \( i \), and \( \text{link}(i) \) is the connection set of node \( i \).

We define a subnet as an influence network starting from an artist. Take the second-order transfer network with ID 145543 as an example, and its transfer subnet is shown in Figure 3.

![Figure 3. Second-order Transitive Network with ID: 145543](image)

We can get some important characteristics about the node and the whole sub-network from the parameters defined above. For example, we can get the node's relative importance by calculating three important different centralities and then rank the importance of the nodes in the sub-network. The density of the network can show the closeness of the nodes. After calculation, the density \( D(g) \) of the sub-network is 0.2619, while the overall density of the network is 0.001, indicating that the sub-network is closely connected. Second-order Degree can get the transitivity influence degree of nodes. The calculation indexes of some nodes are shown in Table 1.

| ID     | \( C_c \)  | \( C_e \)  | \( S_d \) |
|--------|------------|------------|-----------|
| 145543 | 0.153836   | 0          | 6         |
| 494537 | 0.181765   | 0.000743   | 25        |
| 275832 | 0.142647   | 0.013173   | 27        |

### 3. Similarity index system

#### 3.1. Songs Similarity

Music similarity can be expressed as the similarity between samples. There are many traditional measures, such as the Jaccard similarity coefficient, cosine similarity, Pearson similarity coefficient, etc. [7]

Cosine similarity measures the similarity between two vectors by measuring the cosine value of the angle between them. The cosine value of 0° is 1, while the cosine value of any other angle is not greater than 1, and its minimum value is −1. So the cosine value between the two vectors determines whether they are pointing in roughly the same direction. The cosine similarity can be used to synthesize the differences of each feature, and the normalized similarity value ranges from 0 to 1, which simplifies the analysis process.
Here, we use the variance of music features to measure the importance of features based on cosine similarity. The greater the variance of music features, the greater the weight it occupies. This is reasonable because after enlarging the weight of the volatile features, we can find the core of the problem more accurately and quickly when comparing the similarities and influences among artists and the differences within and between genres. The formula of variance weight is as follows:

\[ v = \sum_{n=1}^{n} a_j \]  
\[ \delta^2 = \frac{(x-v)^2}{N} \]  
\[ \omega_i = \frac{\delta_i^2}{\Sigma \delta_i^2} \]  

Then, we use the improved cosine similarity to measure the similarity of songs. The formula of cosine similarity is as follows:

\[
\text{Similarity} (A, B) = \frac{\sum_{i=1}^{n} \omega_i a_i b_i}{\sqrt{\sum_{i=1}^{n} \omega_i a_i^2} \times \sqrt{\sum_{i=1}^{n} \omega_i b_i^2}}
\]

where A and B are the feature vectors of any two songs, A = [a_1, a_2, ..., a_n], B = [b_1, b_2, ..., b_n].

3.2. Artist Similarity

An artist belongs to only one genre, the songs he sings are specific, and the similarity between songs has been obtained from the previous section. So, how to measure the similarity between artists? The practical idea is to compare the similarities of songs between two artists, get an average, and then compare them. However, there is a problem that only using the average method to measure the similarity between two artists will lose some important information.

We observed that the similarity matrix of two artists’ songs could be obtained by comparing the similarity of music songs between two artists, a symmetric matrix. The similarity of artists is the data difference degree of this matrix. Referring to the idea of matrix norm [8], we get three methods to measure the similarity between two artists. The formula is as follows:

\[ S_1 (P, Q) = \frac{\|M_{PQ}\|_\infty}{N_P} \]  
\[ S_2 (P, Q) = \frac{\|M_{PQ}\|_1}{N_Q} \]  
\[ S_3 (P, Q) = \frac{\|M_{PQ}\|_2}{\sqrt{N_P N_Q}} \]  

where M_{PQ} the songs similarity matrix between artists.

Their values range from 0 to 1. The closer to 1, the more similar the artists are. Due to a large amount of data, we only give one example: the similarity comparison between Ray Wylie Hubbard and Obituary. The song similarity matrix of the two artists is shown in Figure 4.
Based on this, we can calculate their similarity, as shown in Table 2. We can see from the picture that the three similarities of Ray Wylie and Freddie Hubbard are all around 0.55, while the average similarity of music is 0.6934, so we can know that the similarity of the two artists is relatively low. Moreover, we know that ray Wylie belongs to the country's genre, and Freddie Hubbard belongs to two different schools, which further verifies the accuracy of our results.

### Table 2. Artists Similarity

|        | $S_1(P, Q)$ | $S_2(P, Q)$ | $S_3(P, Q)$ |
|--------|-------------|-------------|-------------|
|        | 0.650843    | 0.553562    | 0.553562    |

3.3. Similarity of Artists within and between Genres

From section 3.2, we get the calculation method of similarity between artists. On this basis, we can calculate the average similarity of artists within the genre and the average similarity of artists between genres. By comparing the differences between the two, we can measure the influence of the attribute of the artist genre. Some of the results are shown in Figure 5.
We can see from the above that, in most cases, the similarity of artists within genres is more similar than that between genres. Only a few cases are different. For example, the similarity of artists within electronic is lower than that between electronic and jazz.

4. Conclusion
To understand the influence of music through networks, we build a music influence model. It includes the following parts: artist influence model, genre influence model, and music similarity model. Each part is connected to effectively identify and quantify the value of music influence in all aspects.

Our model can intuitively show the influence of music influence in different dimensions through the defined parameters. We give a precise and actual meaning and explanation for each definition, such as the centrality measurements of artist nodes and the influence degree within and between genres.

Our model has excellent practical value. When considering the indirect influence of the artist, we perfectly integrate the simplicity of the model and the practical significance of the problem and use the second-order degree of the node and the centrality of the eigenvector to measure the indirect influence of the artist. However, in quantifying the actual influence of artists, the indirect influence of artists is not considered, which simplifies the complexity of our model.

Culture is often associated with the development of music, which usually leads to some changes in culture. Our model has a good advantage in identifying the possible changes in music. By measuring the changes of parameters per unit time, we can quickly identify some possible cultural influences. Moreover, when we recognize these changes, we can quantify the possible cultural changes through the changes in musical characteristics.

References
[1] Duan-Bing Chen, Hui Gao, Linyuan Lü, and Tao Zhou. Identifying influential nodes in large-scale directed networks: The role of clustering. PLOS ONE, 8(10):1–10, 10 2013.
[2] P. Bródka, K. Skibicki, P. Kazienko, and K. Musial. A degree centrality in multi-layered social network. In 2011 International Conference on Computational Aspects of Social Networks (CASoN), pages 237–242, 2011.
[3] Junlong Zhang and Yu Luo. Degree centrality, betweenness centrality, and closeness centrality in social network. In Proceedings of the 2017 2nd International Conference on Modelling,
Simulation and Applied Mathematics (MSAM2017), pages 300–303. Atlantis Press, 2017/03.

[4] Phillip Bonacich. Some unique properties of eigenvector centrality. Social Networks, 29(4):555564, 2007.

[5] Ian Rogers. The google pagerank algorithm and how it works. https://cs.wmich.edu/gupta/teaching/cs3310/lectureNotes_cs3310/Pagerank%20Explained%20Correctly%20with%20Examples_website.cs.princeton.edu~chazelle_courses_BIB_pagerank.pdf.

[6] Peter V. Marsden. The reliability of network density and composition measures. Social Networks, 15(4):399–421, 1993.

[7] Peipei Xia, Li Zhang, and Fanzhang Li. Learning similarity with cosine similarity ensemble. Information Sciences, 307:39–52, 2015.

[8] Nicholas J. Higham. Experience with a matrix norm estimator. SIAM Journal on Scientific and Statistical Computing, 11(4):804–809, 1990.