# Analysis of Stylistic Similarity for Music Recommendation

**Group 3: Angad Singh, Abhishek Anand, Ameya Swar, Haitao Liu**

## Executive Summary

The project is inspired by one of the research papers by James Hughes titled as “Quantitative Patterns of Stylistic Influence in the Evolution of Literature”. In this paper, the author tries to understand the stylistic similarity of the authors of books ranging hundreds of years. The main research of our project is to find the quantitative stylistic similarity between the lyrics corpus of artists that span over 50 years from 1960-2010. The business motivation behind the project is to be able to provide an artist recommendation based on the similarity score calculated in the first part of the project. Kullback Leibler Divergence was implemented to compute the similarity scores of the artists. A number of graphs were generated showing the temporal distance vs similarity based on the different genres of songs available and an overall comparison. Networks were constructed using an artist as the central code to link it with similar artists according to the similarity scores.

# Introduction

In today's world, you are surrounded by songs wherever you go or wherever you are. Supermarkets, movie, theaters, malls, bathrooms are all playing songs over the speakers. There are numerous concerts taking place around the world, there are billions of people listening to these songs every day. When something is so abundantly found in the environment, it is bound to have an impact on the culture of the society or is bound to be influenced by the culture of the society as the people who write these lyrics may try to express their general mood through these songs.

Music and songs have been part of human culture for centuries, even in today's world, music plays a huge role. Wherever you go, you are surrounded by songs everywhere, the supermarket, the movie theater, the mall, there are songs playing over the speakers. There are numerous concerts taking place around the world, there are billions of people listening to music every day. This study is based on trying to find a pattern of similarity between the songs and lyrics of these songs over the decades. It is inspired by one of the papers by James Hughes which was titled, "Quantitative Patterns of Stylistic Influence in the Evolution of Literature". This paper attempts to study the stylistic influence in the evolution of literature over a period of four centuries. They generated a matrix of similarity to compare the similarity between the works of different authors and greater magnitude suggested greater influence on an author. [1] Similarly, we computed a similarity matrix to test the similarity between the lyrics of each artist.

The project started by collection of raw data from Kaggle and there was intensive work done to clean the data to make it suitable for running our models on the same. A lyrics corpus was generated for the artists we needed to conduct our analysis. After using Kullback Leibler Divergence to calculate the similarity score between the lyrics of each artist, analysis of the results were performed. These results were documented visually using graphs and networks were created to provide artist recommendations.

# Data Preparation

The data, consisting of 380,000+ song lyrics from metrolyrics, was obtained from kaggle.com. [2] The data consisted of the fields such as song index, song, year, artist, genre, and lyrics. However, there were lot of lyrics that were empty. Also we restricted our study to lyrics that were in English. Therefore records with empty values and those which were written in other languages were discarded to restrict the comparison matrix of artists. Below is a snapshot of a sample dataset

| index | song | year | artist | genre | lyrics |
|---|---|---|---|---|---|
| 0 | ego-remix | 2009 | beyonce-knowles | Pop | You know I'm gonna cut right to the chase<br>Some women were made but me, myself<br>I like to think that I was created for a<br>special purpose |
| 1 | then-tell- | 2009 | beyonce-knowles | Pop | playin' everything so easy,<br>it's like you seem so sure.<br>still your ways, you dont see<br>i'm not sure if they're for me. |
| 2 | honesty | 2009 | beyonce-knowles | Pop | If you search |

**Figure 1: Snapshot of Sample Data**

Moreover, the data was restricted only to those artists who had 75 or more songs to their names. The lyrics corpus of such artists were concatenated and one record per artist was created. The dataset, thus created, consisted of artist, concatenated lyrics corpus, genre, and weighted year (or year or prominence). The weighted year of the artist is calculated using the formula:

$$Y_i = \text{floor}(\Sigma S_i N_i / \Sigma N_i)$$

Where $S_i$ is the year of the song release of the artist and $N_i$ is the number of songs in that year.

| | artist | lyrics_corp | Genre | Weighted_year |
|---|---|---|---|---|
| 6185 | calvin-harris | the sun your heartbeat of solid gold<br>i love you youll never know when<br>the daylight comes you feel so cold<br>you know im too afraid of my heart<br>to let you go waiting for the fire to<br>light feeling like we could do right<br>be the one that makes tonight | Electronic | 2013 |

**Figure 2: Snapshot of concatenated data**

Suppose an artist sings releases 20 songs in 2008 and 15 songs in 2012, the weighted year or the year of prominence of the artist will be floor((2008 X 20 + 2012 X 15) / (20+ 15)) = 2009. The rationale behind using weighted year will be clear when comparison between artists is

generated. These operations eventually narrowed down the number of artists from 16000+ to 955. Once the concatenated lyrics corpus dataset is generated, the lyrics of the artists are compared using Kullback Leibler divergence.

# Why content-free words ?

In our study, we have used a list of content-free words to determine the stylistic similarity between the artists. Content-free words are those words which don't convey much on their own, but are used as bridge between words to convey the entire meaning. In essence, the quantitative analysis of content-free words sheds light on the style of the artists under consideration. Had content word such as action verbs been used, the analyses would have shifted to the similarity of intent of the lyrics or the artists. [1]

# Similarity Score Calculation

Once the lyrics corpus dataset is generated, each artist is compared with all the other artists to check similarity between their lyrics. The following steps are performed to generate a dataset that tells about similarities between artists.

- The occurrence of each content-free word of each artist is calculated and the count is aggregated.
- The count is normalized so that the components sums to 1.
- All the pair of artists are subjected to symmetrized Kullback-Leibler divergence to calculate divergence between the pair. The below formula is used:

$$D_{\mathrm{KL}}(P_i, P_j) = \frac{1}{2} \sum_{\omega \in \Omega} \left( P_i(\omega) \log \frac{P_i(\omega)}{P_j(\omega)} \right) + \left( P_j(\omega) \log \frac{P_j(\omega)}{P_i(\omega)} \right)$$

where where Ω is the set of content-free words and P(ω) is the corresponding unitized content-free frequency vector for artist i. Using the divergence thus calculated, similarity score is generated, which is given by $S_{ij} = \exp(-d_{ij}/0.5)$. [1] The higher the similarity score between the artists, the higher will be the similarity between the styles of their lyrics determined by occurence of content-free words in their lyrics.
The new dataset has the fields Artist 1, Artist 2, Similarity score, and temporal distance.

Temporal distance is used to study the pattern of stylistic influence over the period of study. Temporal distance is the difference between the years of prominence (weighted years) between two artists.

Temporal Distance between artists i and j = $Y_i - Y_j$

| Artist1 | Artist2 | Similarity | TemporalDistance |
|---|---|---|---|
| 10-cc | 2-chainz | 0.850487 | 9 |
| 10-cc | 2-live-crew | 0.899488 | 0 |
| 10-cc | 2pac | 0.897795 | 2 |
| 10-cc | 50-cent | 0.888563 | 4 |
| 10-cc | 5th-dimension | 0.88639 | 7 |
| 10-cc | 7-seconds | 0.877736 | 3 |
| 10-cc | a-ha | 0.928389 | 0 |
| 10-cc | aaliyah | 0.862246 | 1 |

**Figure 3: Snapshot of similarity and temporal distance**

# Why do we consider Kullback Leibler divergence?

KL Divergence has its origins in information theory. The primary goal of information theory is to quantify how much information is in data. The most important metric in information theory is called Entropy, typically denoted as H. The definition of Entropy for a probability distribution is:

$$H = -\sum_{i=1}^{N} p(x_i) \cdot \log p(x_i)$$

We can interpret entropy as "the minimum number of bits it would take us to encode our information". The key thing with Entropy is that, simply knowing the theoretical lower bound on the number of bits we need, we have a way to quantify exactly how much information is in our data. Now that we have this, we want to quantify how much information is lost when we substitute our observed distribution for a parameterized approximation. [3]

Kullback-Leibler Divergence is just a slight modification of our formula for entropy. Rather than just having our probability distribution pp we add in our approximating distribution qq. Then we look at the difference of the log values for each. Essentially, what we're looking at with the KL divergence is the expectation of the log difference between the probability of data in the original distribution with the approximating distribution. With KL divergence we can calculate exactly how much information is lost when we approximate one distribution with another.

It may be tempting to think of KL Divergence as a distance metric, however we cannot use KL Divergence to measure the distance between two distributions. The reason for this is that KL Divergence is not symmetric. Hence we consider the temporal distance between the two artists to measure the distance between two artists.

# Results

## Distribution of temporal distances

The histogram reveals that most pairs of artists have low temporal distance, signifying temporal localization since most artists are related to contemporary artists. The result is a right-skewed histogram.

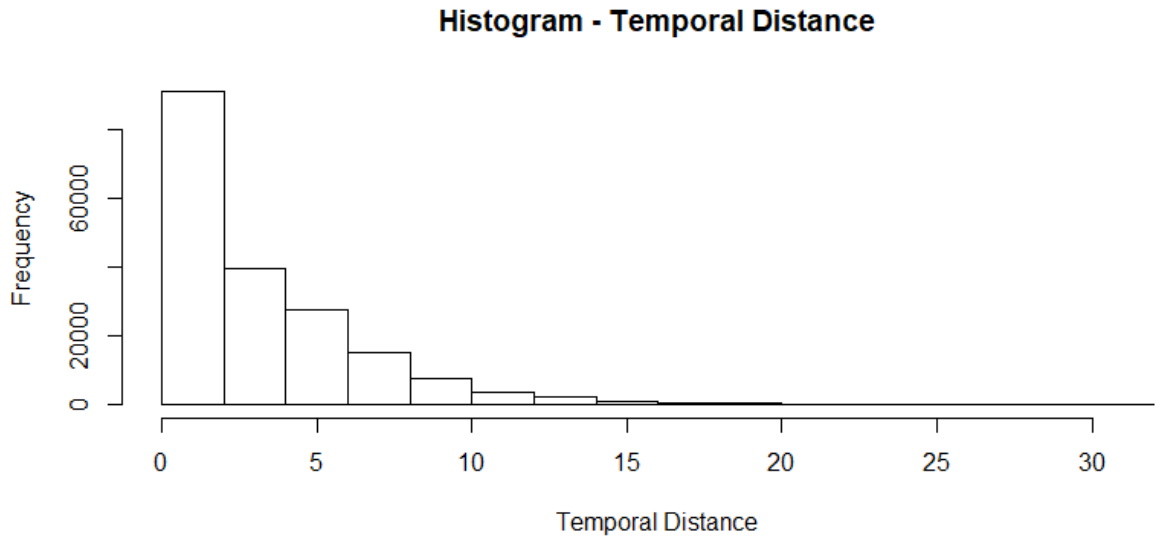


**Figure 4: Snapshot of Temporal Distance distribution**

## Similarity Graphs

We calculated the similarity and temporal distance of all the artists in our dataset and tried to find the relation between the similarity score and the temporal distance. The 2 line charts below show the similarity score with the increasement of the temporal distance for all the artists and pop artists:

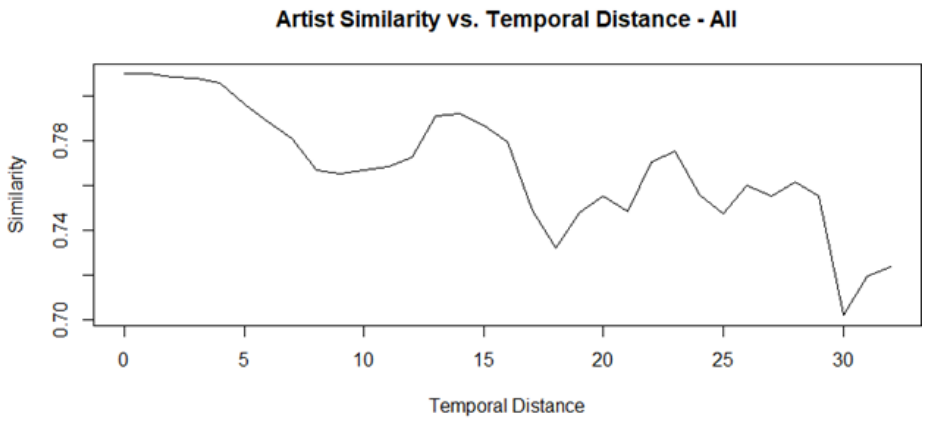


**Figure 5: Graph for Artistic Similarity vs Temporal Distance (all artist)**

It can be seen from the line chart that the similarity of artists has reduced as the temporal distance between the artists has increased, which means that artists are more influenced by their contemporaries. The lyrics created in the same year are more likely to use the same expression and words. For example, some words that are trendy in one period will be more likely mentioned in some lyrics of the song in that period. It is understandable that the similarity decreased as the temporal distance increasing. Each contemporary has their own language pattern that might recognize from the lyric.

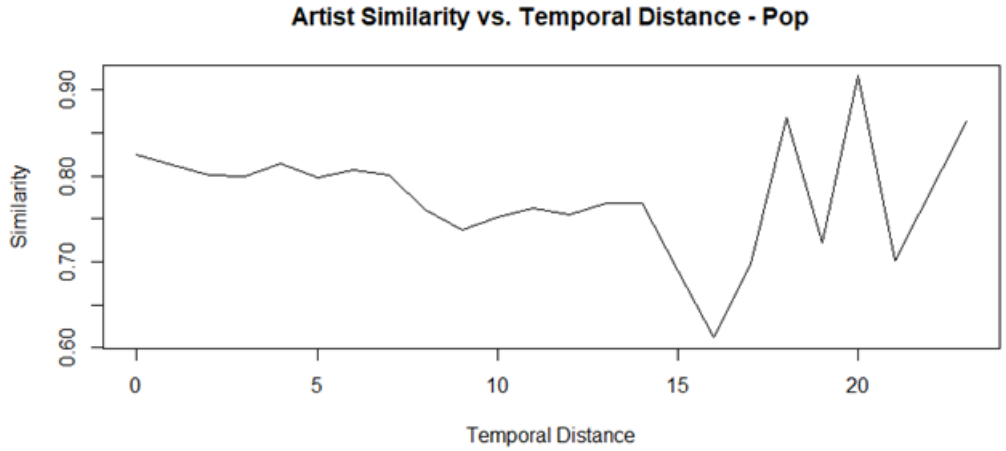


**Figure 6: Graph for Artistic Similarity vs Temporal Distance (pop artists)**

We also plot the line chart only for pop artist, the result obtained are different from the previous one. It can be seen that the similarity score reduced to around 0.6 when the temporal distance reaches 15 with a decreasing slope as the temporal distance kept increasing. After that value of 15, there was a rise again in the similarity scores of the artists which could possible mean that these artists picked up styles from the artists that were prominent 15+ years before then and there was an influence on their style. These could be them doing some songs with the retro theme and giving tribute to their favorite artists from when they were growing up.

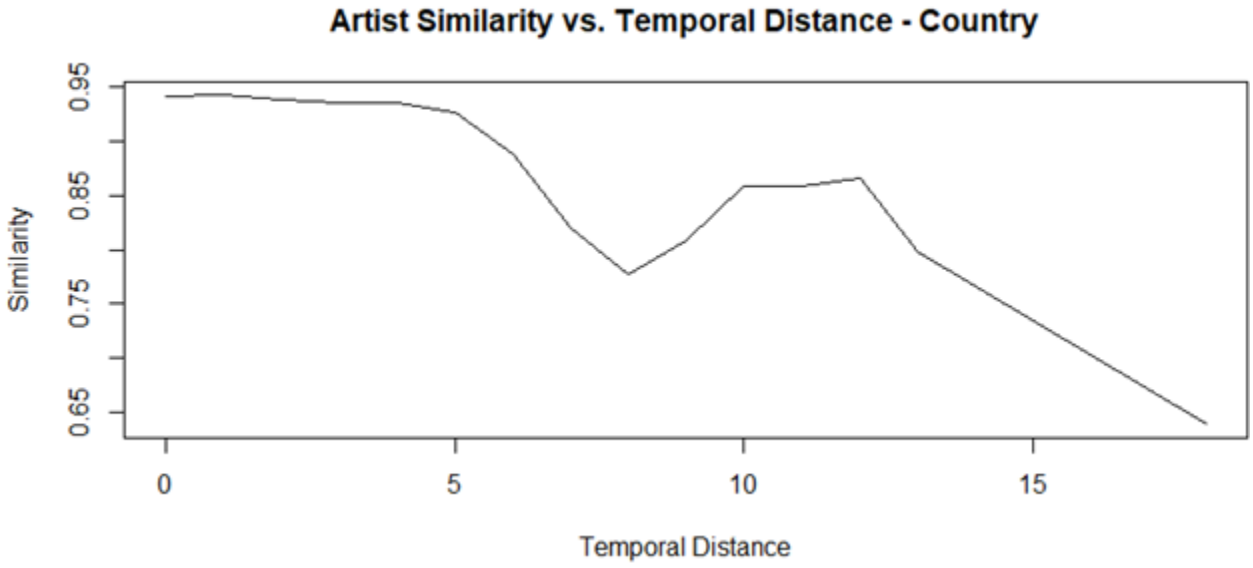


**Figure 7: Graph for Artistic Similarity vs Temporal Distance (country artists)**

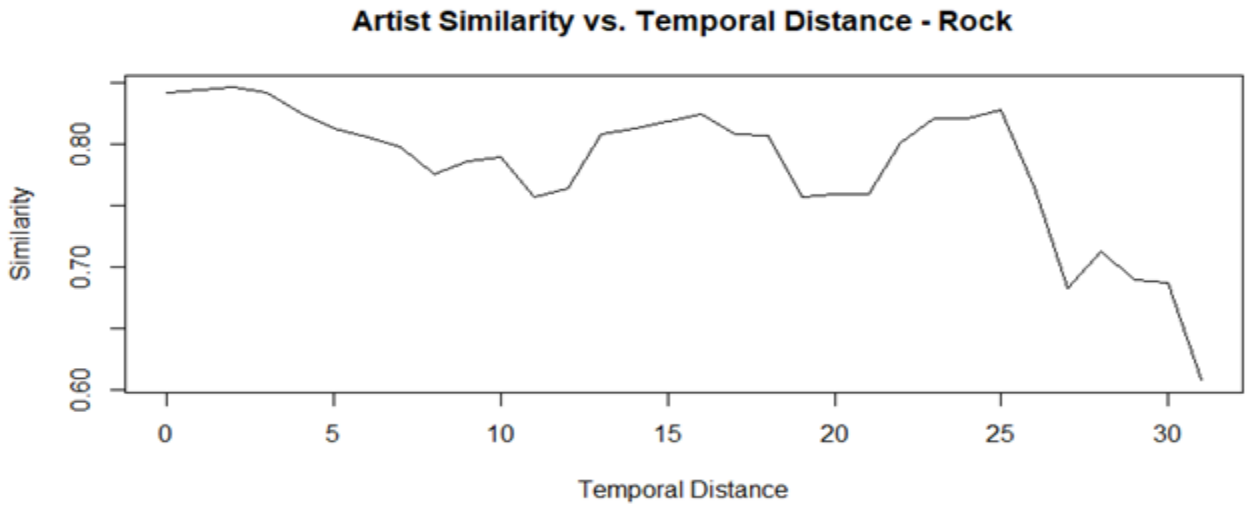


**Figure 8: Graph for Artistic Similarity vs Temporal Distance (rock artists)**

Likewise, the graphs for rock and country artists show similar trends where average similarity diminishes as temporal distance increases. The spikes can be explained, again, due to influences of other artists from previous generations.

## Visualization Networks

In this project, we built the recommendation network for each artist. The nodes in the network represent the name of the artists and the weight of the edges represent the similarity score. The higher the similarity score, the thicker the edges. All the recommended artists are connected to the central node which is the artist as the subject of this recommendation. There was a very huge list of artists that could be connected to each other therefore only the best 20 artists for each were taken for the recommendation system to build a network on. There are a few examples shown below.

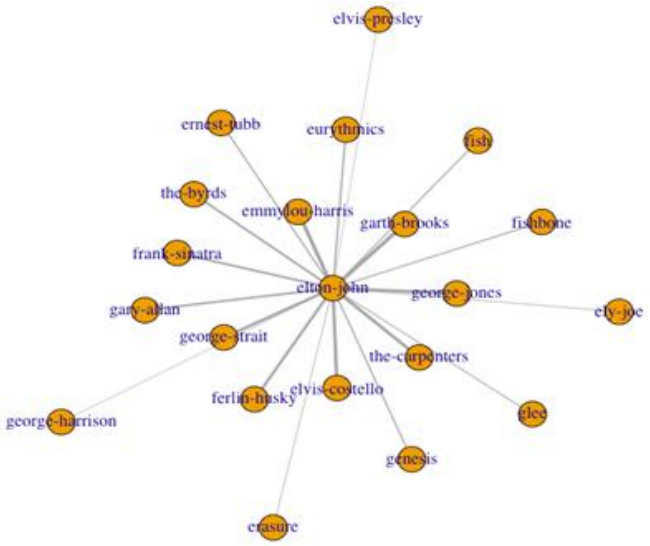


**Figure 9: Elton John Network**

This network shows the 20 artists that have a similar style of the lyrics with Elton John. It can be seen from the graph that Garth Brooks , Emmylou Harris and The carpenters have the highest similarity to Elton John in their lyrics. So we can recommend these artists to the user who like to listen Elton John's music.

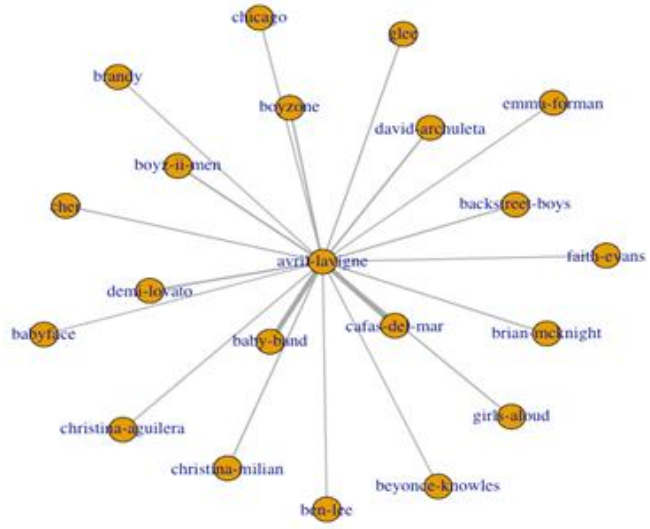


**Figure 9: Avril Lavigne Network**

The graph above presents the 20 artists that have a similar style of the lyrics with Avril Lavigne. You can see some artists have the same genre with Avril such as Chicago, baby band, backstreet boys, Boyzone. There is a stylistic similarity between these artists when compared to work of Avril Lavigne.

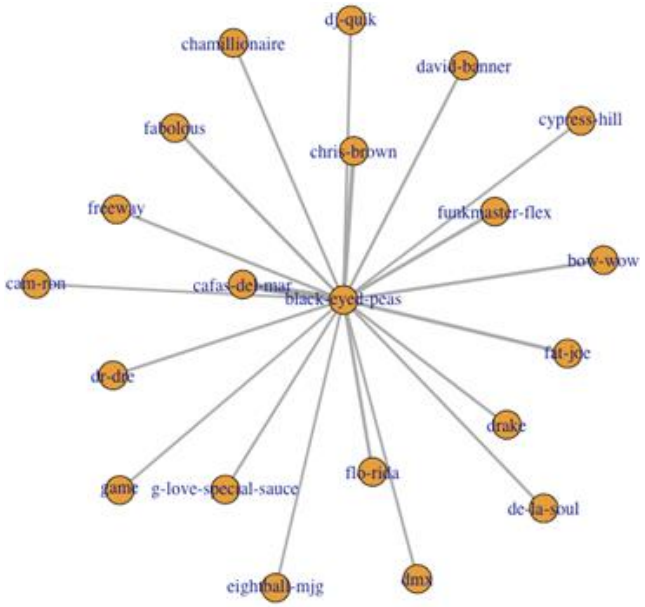


**Figure 10: Black Eyed Peas Network**

As is shown above, this graph presents the network of the similarity for black eyed peas. This music band was a hip pop band originally and then changed to pop and electronic music. It can be seen that the recommendations fall under in the similar category.

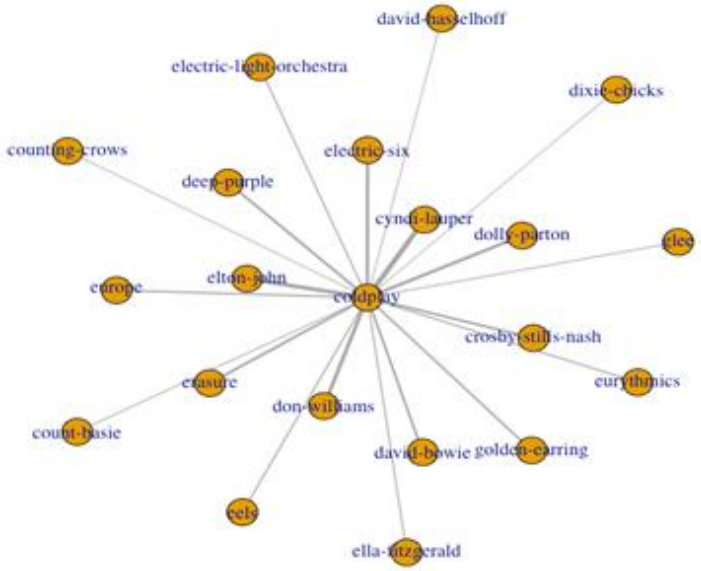


**Figure 11: Coldplay Network**

It can be seen from this graph that the artists that have the high similarity with Coldplay are Cyndi Lauper, Don Williams, David Bowie and Electric Six.

# Conclusions

The main objective of this study was to compare and analyze the stylistic similarities between the artists from a given period of time. There was a total separation of around 50 years between the first and the last artist, ranging from 1960 to 2010. There were a number of graphs constructed to understand the nature of the relationship between the stylistic similarities of artists with respect to the temporal distance. It was observed that as the temporal distance between artists increased, the stylistic similarity between them went down. This meant that an artist who is writing music in 2010 will have a similar style to someone who wrote a song in 2005 than an artist from 1980s. It is a result which was expected as you can yourself understand that songs written close to each other are temporally bound to be more similar compared to those that have a number of years between them. This was proved and attested by the results obtained in the study for the comparison of the stylistic similarity between these artists.

It is understandable that peers are bound to be more influenced by their contemporaries especially in the field of arts and the case similar in this situation. Although, in some instances, for example the pop artists it was observed that after a certain increase in temporal distance, there was rise in similarity scores of the artists. The possible reason for this is that these artists were writing pieces in their work which were tributes to those artists from the past which inspired and influenced them to become artists in the first place.

Another part of the study was to be able to construct networks using a particular artist and their similarity scores to provide song recommendations. People who are more inclined to listening to certain style of lyrics can be targeted for this recommendation system. We constructed networks with a few artists to see how the system was performing. It was observed that artists similar to each other were being recommended, for example, people listening to 50 cent would be recommended artists like Eminem, Big Sean, Dr Dre, Akon etc. Similarly, recommendations were provided for other artists like Elton John, Coldplay, Avril etc. Even these recommendations were in line with what would usually be associated with their type of artists in the industry.

# References

[1] James M. Hughes, Nicholas J. Foti, David C. Krakauer, Daniel N. Rockmore, Quantitative patterns of stylistic influence in the evolution of literature. Proceedings of the National Academy of Sciences of the United States of America.

[2] Kaggle, “380,000+ lyrics from MetroLyrics”, Available: https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics

[3] Wikipedia, Kullback Leibler Divergence, Available: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence