Understanding the musical nexus in Spotify

Yan Rong Foo[a,b] **and Nelson Jia Wei Choo**[a,b]

This manuscript was compiled on December 6, 2023

This paper delves into the analysis of related artists on Spotify, exploring the mechanisms influencing listener preferences. Utilizing data from top artists in 2022, along with song lyrics from Kaggle, Spotify API, and Genius API, we investigate the formation of artist communities based on genre similarities. The sentiment analysis of lyrics further enhances our understanding of music preferences and contributes insights into Spotify's recommendation system. The study constructs a network of Related Artists, revealing a low-density graph and a power-law degree distribution, suggesting a scale-free network. Communities are identified through modularity analysis, indicating distinct clusters of related artists. Clustering coefficients highlight community cohesion, while betweenness centrality identifies central artists. Different communities show mixed popularity levels but distinct genre patterns. The degree distribution strongly correlates with artist popularity, suggesting that popular artists tend to have more related artists. Sentiment analysis using the labmt wordlist reveals artists with the happiest and saddest lyrics. TF-IDF analysis explores word importance in lyrics, generating word clouds for artists with positive and negative sentiments. The word clouds showcase positive sentiment with terms like ”loving” and ”sweet,” while negative sentiment word clouds feature words like ”fear” and offensive slurs. The nuanced nature of word clouds emphasizes the complexity of lyrical content and its varied expressions. This comprehensive analysis of Related Artists on Spotify contributes valuable insights into listener preferences, community structures, and sentiment patterns. The findings offer a foundation for enhancing music recommendation systems and understanding the intricate dynamics within the music industry.

social graph | network analysis | text analysis | sentiment analysis | tf-idf | music industry

Music is a constant presence in our daily lives, and it has undergone many layers of evolution especially in recent times, enabling the emergence of new genres and providing additional avenues for listeners to discover fresh artists and songs. This transformation has also facilitated the formation of communities and networks centered around music.

In the context of this paper, our objective is to conduct an analysis on artists' networks in the music industry. We aim to explore how related artists interact with one another due to certain similarities, which can give us insights on the underlying recommendation system from Spotify. Additionally, we seek to delve into the sentiment of lyrics to determine how it may affect the recommendation system.

Musical preferences can be identified through the user's listening history. This is what Spotify capitalizes on when developing their recommendation systems, which give rise to the idea of Related Artists. Related Artists are similar artists identified by Spotify based on the analysis of Spotify's community listening history. In order to achieve our objectives, we have selected a dataset from Kaggle which shows the top 1000 artists of 2022. We are interested in using this dataset as it allows us to use the Spotify Web API to extract the Related Artists of each artist, and create a network of related artists to conduct our analysis. We only included related artists that appeared in our dataset, so that we can see the interactions between related artists in our top 1000 artists. We also used the Genius API to extract the top three songs of each artist to conduct text analysis.

Results

Our study will rely on our dataset from Kaggle which shows the top 1000 artists of 2022. The artists are ranked in terms of popularity, which will be useful in comparison with the network's degree distribution later on. We converted all of the artist names to lower case to ensure standardization. The dataset includes the artist's name, lead number of streams, features, tracks and others. We will use the artist name on the Spotify API to retrieve the required columns such as genres and Related Artists. Given the search function in Spotify API that searches for

Significance Statement

This study on Related Artists in Spotify, analyzing top artists and their lyrics, holds significance in unraveling the intricate dynamics of listener preferences and community structures within the music industry. The identification of scale-free networks and cohesive artist communities sheds light on the underlying mechanisms of Spotify's recommendation system. Additionally, the sentiment analysis of lyrics provides nuanced insights into emotional expressions in music. These findings not only contribute to refining music recommendation algorithms but also offer a foundational understanding of how users engage with and form communities around music, thereby influencing the broader landscape of digital music consumption.

Author affiliations: [a]Technical University of Denmark; [b]National University of Singapore

Author Contributions

[1]Yan Rong Foo contributed to the text analysis, sentiment analysis and TF-IDF. Nelson Jia Wei Choo contributed to artist network analyis.

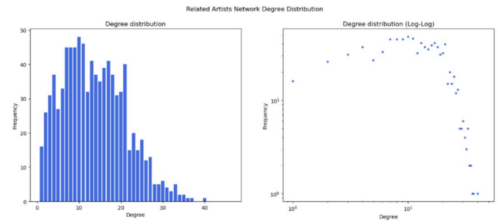
[2]To whom correspondence should be addressed. E-mail: s231657dtu.dk

the artist based on their name, it will give multiple search results, hence we have to filter and keep the most popular result to ensure that we are selecting the correct artist. For related artists, we only keep related artists that exist in our dataset of artists. This is to ensure our network will not have edges to nodes that do not exist.

Network and Graph Density. Each node represents a Spotify artist from our dataset. If two artists are Related Artists, there will be an edge between these two nodes. The network contains 893 nodes and 6068 edges. Afterwards, we did some analysis on the density of the graph, which represents the ratio between the edges present in the graph and the maximum number of edges that a graph can contain. This can give us an idea of how dense our artist graph is in terms of edge connectivity. For the density of the graph, we observed that it was low at around 0.015, which might imply that artists are not related to many artists. This makes sense because an artist is probably not able to be related to all artists as there would be a limit imposed by the Spotify algorithm, which would hence lead to a low artist network graph density.

Degree Distribution. Afterwards, we are interested in analysing the degree distribution. By examining the histogram of degree distribution as well as the Log-Log graph of degree distribution, we observe that most of the nodes have lower degrees, some have moderate degrees and a few nodes have higher degrees. This means that some artists have only a few related artists, while other artists have many related artists, some of which can be identified as hubs. We observe that the mean degree is around 13.6, which indicates that on average, each artist has about 13 related artists. We also observe the median degree, which shows artists having 13 related artists. The most prominent number of related artists is 8, meaning that most artists have about 8 related artists. We can also see that there is 1 artist with 48 related artists. We were interested in finding out who this artist is, and we found out that it is Bebe Rexha. This does make sense because she is an extremely popular artist that makes songs in multiple popular genres like Pop, Dance, EDM music. Finally, we see that the minimum number of related artists is 1, which is expected as we removed all the isolated nodes in our pre-processing.

From our Log-Log degree distribution graph, we can see that the degree distribution obeys a power law. Hence, we can check if the artist network is a scale-free network. A scale-free network is a network whose degree distribution follows a power law. This means that the logarithm of the probability of a node having a certain degree k depends linearly on the logarithm of degree k, and the slope of a line representing the relationship is the power law coefficient. When the power law coefficient is high, the number of nodes with high degree is smaller than the number of nodes with low degree. This means that the distribution of edges is more even. The power law coefficient we obtained from our analysis is about 15.122. Hence, the artist network follows the small world regime since degree exponent is greater than 3. This means that the variance of the degrees is finite, which shows us that the spread of related artists is quantifiable and does not diverge. The average distance between the artists also follows the small world result that's derived for random networks.



	Mean Degree	Median Degree	Mode Degree	Maximum Degree	Minimum Degree
0	13.590146	13.0	10	48	1

Fig. 1. The degree distribution of the network plotted on both a linear and logarithmic scale.

Communities and Modularity. Our goal here is to identify the optimal partition that separates the network into communities, to find out how related artists form clusters, and whether they form based on similar genres and popularity. Modularity is a network analysis measure that quantifies the extent to which a network can be divided into distinct, densely connected communities or modules. It aims to assess the quality of community structure within a network. The intuition behind modularity is to evaluate whether the connections within communities are significantly stronger than what one would expect in a random network. In other words, it helps us determine if there are meaningful and well-defined subgroups or clusters of nodes in a network. Since we are computing the modularity score of the best partition, we are obtaining the maximum modularity. We obtained a modularity score of 0.75 for the best partition which is considered high. As such, we can conclude that the artist network is well-separated into various distinct communities where each community is made up of different artists that are related to other artists in that same community.

Popularity and Genres within Communities. Since the Related Artists have formed communities, we want to investigate if the Related Artists in a community are similar in any way, which aligns to our initial objective. We will first check if artists in communities are generally of similar popularity, for which we will use a wordcloud representation to see if there are any extremely popular artists within communities.

From the wordcloud, we can see that within each community, there is a mixture of popular and less popular artists. From this, we can conclude that the related artists are not really similar in terms of popularity. However, we notice that a decent number of artists within each community have extremely similar genres. For example, in community 6, we see artists like David Guetta, Martix Garrix, Calvin Harris, The Chainsmokers, Avicii, all of which are electronic dance music (EDM) artists. Also in community 2, we see Miley Cyrus, Justin Bieber, Ed Sheeran, who are mostly pop artists. This led us to believe that there could be a relation in terms of genre instead.

After analysing the genres from each community, we can tell that there is definitely some similarity in genres within the community. For example, in community 11, we see genres appear frequently like Latino, Urbnano, Reggaeton, Latinocolombian which are all extremely similar to one another. This applied the same to community 0 which has genres like house, dance, electro, edm, pop, and community

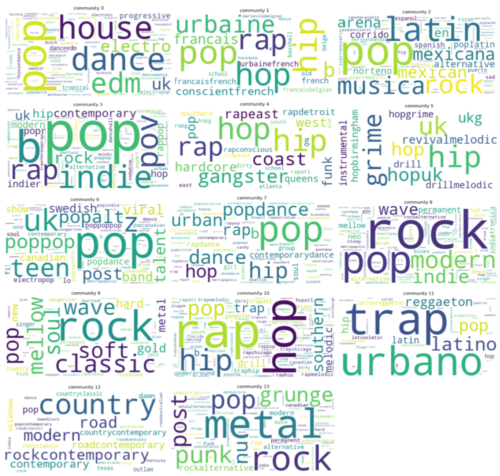


Fig. 2. Wordclouds which depict the genres within each community.

10 which has rap, hip hop, underground, rapmelodic. This shows that related artists do tend to have similar genres, which makes sense because listeners would have preferences for specific genres, hence related artists that are recommended would tend to be of that same genre as well.

Clustering Coefficient. The clustering coefficient captures the degree to which the neighbors of a given node link to each other. For our case, a node with high clustering coefficient suggests that most of the related artists of the node are also related artists with one another. From the graph, we observe that most of the artists have a clustering coefficient between 0.4 to 0.6. This means that the related artists of a specific artist are somewhat linked to each other. We do notice a small number of artists with clustering coefficients ranging from 0.8 to 1, which means that the related artists of a specific artist are mostly also related artists amongst each other.

From the summary statistics, we observe that the mean clustering coefficient is about 0.47. This means that related artists of a specific artists are normally somewhat related amongst themselves. This finding aligns with our hypothesis as artists can have more than 1 genres which makes them more possible to be similar with other artists.

Most of the communities have clustering coefficients between 0.5 to 0.7. This shows that artists in the communities tend to be fairly related. There are some communities who have clustering coefficients closer to 0.8, and some with a wide range of coefficients like community 8 and 13. This aligns with the clustering coefficient statistics of artists we generated previously. Also, the smaller communities such as community 3 and 11 tend to have higher clustering coefficients, which is expected since the size is smaller, the artists within that community tend to be more or less related.

We observe that there is a negative correlation between clustering coefficient and degree distribution, meaning if a node has a higher degree, its neighbors are less connected amongst each other. This is an interesting find because an artist could have many related artists because he or she could be known for multiple genres, hence associated with artists from different genres, but they themselves might only make music in one genre.

Betweenness Centrality. We observe that the betweenness centrality for most of the communities are generally low between 0 to 0.1. However, again we notice that the smallest communities which are community 3, 11, 13 have higher betweenness centrality compared to the other communities. This is due to their smaller size, hence artists within the community are more likely to occur within the shortest paths between other nodes.

There is a slight positive correlation of 0.25 between Centrality and the artist's popularity. This makes sense because a popular artist is expected to be more central and would occur amongst the shortest path of other artists.

We are interested to see which are the most central artists to see if our analysis on centrality is logical. We identified that the top three most central artists are Enrique Iglesias, Shakira and Jennifer Lopez. These are famous artists in the music industry and it is not surprising to see them as the most central artists in our artist network.

Artists Popularity. There is a strong positive correlation between degree distribution and artist's popularity. This shows that if an artist has more related artists, it is more likely the the artist is more popular. This is expected because artists who are more popular tend to collaborate with more artists, or have music in more genres, hence there is a higher chance of relation to more variety of artists.

Text Analysis. We analyzed the lyrics of each artist's top 3 songs to assess the prevalence of positive or negative sentiments. Using the LyricsGenius python package, we made API calls to extract the lyrics, sorted them by popularity, and concatenated the top 3 songs into a single text document for each artist. These documents were stored as text files and subjected to text preprocessing steps, including tokenization, lowercase conversion, punctuation removal, and elimination of English stopwords. Additional words, such as contextual terms like "[interlude]" were also removed. Other contextual terms like contributor names formatted as "contributorname of contributor," were also filtered out using regular expressions. One such regular expression used was: (contributor*) where * acts as a wildcard to capture the name of the contributor which can contain 0 or more characters. A limitation to this approach is that some lyrics may contain 'contributor' as a legitimate word which will be filtered out. However, removing these contextual words proved to be important when performing further analysis such as when using Term Frequency Inverse Document Frequency (TFIDF).

Lastly, to ensure uniformity, NLTK's WordNetLemmatizer was applied for lemmatization, grouping inflected forms of words together for more meaningful analysis. Once the tokens have been processed, the Counter class is used to count the occurrences of tokens in the corpus and is returned as a list of tuples (token, frequency) in the term frequency function.

Sentiment Analysis. Conducting sentiment analysis can help to gauge the degree of positivity or negativity present in the lyrics of each song artist. This analysis allows us to discern patterns within the lyrics, revealing nuanced sentiments. To perform sentiment analysis, we employ the labmt wordlist approach. The word list was created by combining and scoring the sentiments of 5000 words which appear the most

Fig. 3. Wordclouds which depict top 9 artists with happiest lyrics.

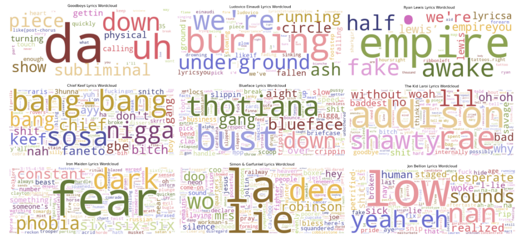


Fig. 4. Wordclouds which depict top 9 artists with saddest lyrics.

frequently in Twitter, the New York Times, Google Books, and music lyrics. Considering the specific context of musical lyrics addressed in this report, we assert that the labmt wordlist approach is well-suited for evaluating the sentiments of lyrics for various song artists. The sentiment value (sv) is calculated with the formula:

$$\mathrm{sv}_{text} = \frac{\sum_k v_k f_k}{\sum_k f_k}$$

where *vk* refers to the happiness average value of token k in the labmt wordlist and *fk* refers to the number of times token k appears in the text.

Using this method, the average sentiment value of the text can be calculated by taking into account the sentiment values of individual tokens and their occurrence frequencies in the text. The following histogram is produced.

The histogram displays a relatively normalized distribution. We observe that the majority of artists exhibit sentiment values in the range of 5.30 to 5.80, with an average sentiment value of approximately 5.55. Considering a sentiment score of 5 as neutral, it can be inferred that song artists generally incline towards incorporating positive language in their lyrics more often than not.

Afterwards, the top 9 artists with the happiest and saddest song lyrics were extracted. The top 9 artists with happiest song lyrics: Michael Silverman, Sigala, John Williams, Bing Crosby, Michael Bublé, Petit Biscuit, Lost Frequencies, Jack Ü and Stevie Wonder. The top 9 artists with saddest song lyrics: Goodboys, Ludovico Einaudi, Ryan Lewis, Chief Keef, Blueface, The Kid Laroi, Iron Maiden, Simon & Garfunkel and Jon Bellion.

TF-IDF. The TF-IDF value assigned to a word signifies its significance within an artist's song lyrics concerning the lyrics of other artists. Utilizing TF-IDF assists in identifying the most crucial words in a song's lyrics in relation to those of other artists. Examining the TF-IDF values of the top 10 artists with the happiest and saddest lyrics allows us to discern the specific words influencing the overall sentiment of lyrics and facilitates a comparison between happy and sad lyrical expressions. To enhance clarity, word clouds have been created, where the size of each word corresponds to its TF-IDF value.

Observations from the word clouds reveal distinct patterns in lyrics associated with positive sentiments. Key words such as "loving", "delight" and "sweet" are prominent, indicating a pervasive positive energy embedded within the lyrics, thereby conveying a positive sentiment. Additionally, the presence of terms like "merry" and "Christmas" in certain word clouds implies a cheerful and uplifting ambiance commonly found in Christmas-themed songs.

However, it is noteworthy that some word clouds diverge from immediate associations with positivity. An illustrative example is the word cloud generated from Lost Frequencies lyrics, wherein words related to Mexican culture, such as "Mexican," "Mariachi," and "Margarita," take precedence. This deviation suggests a potential thematic focus on praising or exploring Mexican culture within the lyrics of these particular songs.

In summary, the analysis of word clouds provides insights into the prevalent sentiments within song lyrics, showcasing the nuanced expression of positivity in some instances and the exploration of cultural themes in others. The diverse nature of these observations highlights the richness and complexity inherent in the lyrical content of various musical compositions.

On the other hand, in the case of lyrics conveying negative sentiments, easily identifiable words encompass "fear", "lie" and "phobia", along with offensive slurs. One interesting observation is that The Kid Laroi's word cloud stands out with terms like "Addison " and "Rae," seemingly alluding to the TikTok star Addison Rae. Upon closer examination of additional words in the word cloud, such as "shawty," "baddest" (slang for good), and "savage," it becomes apparent that the song likely extols Addison Rae rather than criticizes her. This admiration, however, is displayed in an unconventional manner, employing negatively connotated words for emphasis. This shows that even though some words have negative connotations, they can still be used to portray a positive image.

Discussion

Sentiments across artists. Incorporating principles from both graphical representation and textual analysis allows us to discern patterns within the dataset. A method employed for this purpose involved assigning the hue color of nodes in the graph to correspond with the captured sentiment values for each artist.

The resulting graph reveals two distinct components, with one primarily comprising artists exhibiting higher sentiment values and another predominantly featuring artists with lower sentiment values. Notably, these two components exhibit limited interconnection, suggesting that the sentiment value of lyrics might play a significant role in determining related artists within Spotify's algorithm. This implies that Spotify's recommendation system likely places a large emphasis on assessing the positivity or negativity of song lyrics listened to by users when suggesting new artists to those with similar lyrical sentiment preferences.

Sentiments across communities. An additional analytical approach involved investigating the sentiment values at the community level to discern potential links between sentiment and community formation. Upon breaking down the sentiment values of artists into community averages, it was found that communities exhibited average sentiment values ranging from 5.27 to 5.64. To further explore the variations in sentiment values among communities, we can try to establish a correlation between sentiment values and the genres of each community. This facilitated an examination of whether specific music genres within a community tend to manifest higher or lower sentiment values in song lyrics.

A focused analysis on the two communities with the lowest sentiment values, namely Community 2 and 11, revealed the prevalence of genres such as Latin, Urbano, Mexican, and Trap. Notably, these genres were also less prominent in other communities. This observation suggests a potential association between these music genres and negative lyrical content. Conversely, Communities 9 and 12 exhibited the highest sentiment values, with genres like Country and Classical being notably prominent. This association implies that these genres may have a correlation with more positive lyrical expressions within the respective communities.

Limitations. Data Limitations: The analysis heavily relies on data from 2022, which might not reflect the most recent trends or changes in the music industry. For example, there may be popular song artists who debut in 2023 but this analysis will not capture such data.

Changes in Recommendation System Algorithm: This report works with Spotify's algorithm, as of 22 Nov 2023, for generating related artists. It is possible that Spotify may have made changes to the algorithm for generating related artists, which may cause analysis results to differ.

Network Density Interpretation: Interpreting the low density of the artist network graph is challenging without a clear understanding of what constitutes a "related" artist. Based on initial analysis, there is no clear insight as to why some artists may have more related artists compared to others.

Lyric Sentiment Analysis: Sentiment analysis using the labmt word-list approach measures sentiments through word positivity and negativity, and thus may not cover the full spectrum of emotions.. This may oversimplify the complexity of emotions conveyed in lyrics. The nuances of lyrical content may not be accurately captured. Furthermore, our results have shown that words with negative connotations may not necessarily portray a negative image from the song. Validation against other sentiment analysis methods would strengthen the findings.

Future Work. Temporal Analysis: Understanding how artist networks and relatedness evolve over time can help provide insights into changes in trends and music preferences. This can be done by incorporating data from different years. By learning how the spotify recommendation system updates its "related artists" over time, it is possible to deduce how certain time-dependent features may affect the algorithm.

User Behavior Analysis: Consider incorporating user-specific data to understand how individual listening histories influence the formation of related artist communities.

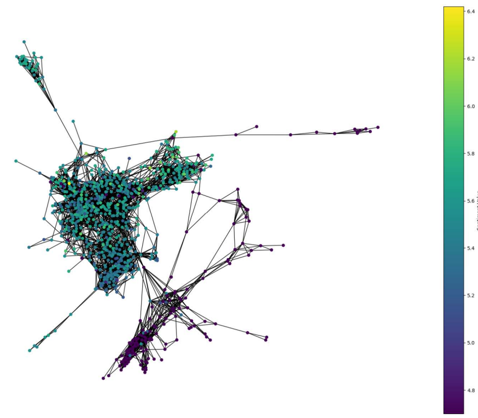


Fig. 5. The artist network with nodes colour coded based on their sentiment value.

Incorporate Additional Features: Include additional features such as artist collaborations, geographic location, or cultural influences to provide a more comprehensive understanding of artist interactions.

Enhance Sentiment Analysis: Utilize advanced sentiment analysis techniques, possibly incorporating machine learning models, to capture the complexity and context of lyrical emotions more accurately.

Materials and Methods

Modularity. Modularity is a measure used to assess the quality of community structure within a network. It quantifies the degree to which a network can be divided into distinct, densely connected communities or modules. The modularity Q for a given partition of a network is typically calculated using the following formula:

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

where Aij is the element in the adjacency matrix of the network, representing the connection between nodes i and j, ki is the degree of node, m is the total number of edges in the network, *(ci,cj)* is a function that equals 1 if nodes i and j are in the same community (partition), and 0 otherwise.

TF-IDF. The Term Frequency-Inverse Document Frequency (TF-IDF) reflects the importance of a word in a document relative to a collection of documents. Term Frequency (TF) measures how frequently a term appears in a specific document while Inverse Document Frequency (IDF) measures the rarity of a term across the entire document collection. An important word is characterized by a high TF value, indicating the term is important within that document, and a low IDF value, emphasizing their uniqueness. The TF-IDF formula is as shown:

$$\mathrm{TFIDF}_{i,j} = \mathrm{TF}_{i,j} \cdot \log\left(\frac{N}{\mathrm{DF}_i}\right)$$

where $TFIDFi,j$ represents the TF-IDF value of token i in document j, TFi,j represents the term frequency of token i in document j, N represents the total number of documents and dfi represents the number of documents containing token i

ACKNOWLEDGMENTS. We would like to thank Professor Jonas Lybker Juul and the teaching assistants for their feedback, which aided us in the development of this project.