# Analyzing and Predicting Success in Music

**Inwon Kang** ( kangi@rpi.edu )
Rensselaer Polytechnic Institute
**Michael Manduluk**
Rensselaer Polytechnic Institute
**Boleslaw K. Szymanski**
Rensselaer Polytechnic Institute

**Article**

**Keywords:**

**Posted Date:** June 28th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1772541/v1

# Analyzing and Predicting Success in Music

**Inwon Kang**[1,2,*], **Michael Mandulak**[1,2], **and Boleslaw K. Szymanski**[1,2,3,*]

[1]Rensselaer Polytechnic Institute, Network Science and Technology Center, Troy, NY 12180 USA
[2]Rensselaer Polytechnic Institute, Department of Computer Science, Troy, NY 12180 USA
[3]Społeczna Akademia Nauk, Łódź, Poland

## ABSTRACT

The emergence of streaming services, e.g., Spotify, has changed the way people listen to music and the way musicians achieve fame and success. Classical music has been the backbone of Western media for a long time, but Spotify has introduced the public to a much wider variety of music, also opening a new venue for musicians to gain exposure. In this paper, we use open-source data from Spotify and Musicbrainz databases to construct collaboration-based and genre-based networks. Our goal is to find the correlation between various features of each artist, the current stage of their career, and the level of their success in the music field. We build regression models using XGBoost to first analyze correlation between features provided by Spotify. We then analyze the correlation between the digital music world of Spotify and the more traditional world of Billboard charts. We find that within certain bounds, machine learning techniques such as decision tree classifiers and Q-based models perform quite well on predicting success of musicians from the data on their early careers. We also find features that are highly predictive of the artist's success. The most prominent among them are artists' collaboration counts and the span of their career. Our findings also show that classical musicians are still very centrally placed in the general, genre agnostic network of musicians. Using these models and success metrics, we make suggestions that aspiring musicians can rely on considering which aspects of their career should be improved to increase their success measures in both Spotify and Billboard charts.

## Introduction

At every age and genre of music, there is a small fraction of musicians who achieve critical success. Although the definition of success often varies, many research efforts have been focused on discovering measures that act as proxies for success and on understanding the reasons behind the validity of these measures. Past works have taken this approach in various subject domains, such as art[1], movies[2] and in research[3], and have presented models for predicting the success of a given individual.

Considering the music field specifically, past works utilize Spotify's database to analyze the popularity of individual songs[4] or to construct networks of musicians and analyze the relationship between popularity and the network rank[5]. The latter use these network models to capture the interactions of nodes, most often from a collaborative or relational perspective.

This work aims to contribute to this line of research on analysis and predictions of the success of musicians by using together the Spotify, MusicBrainz and Billboard's datasets. We analyze the careers of musicians using several features, such as music genres, productivity (quantified as the number of releases), release trends, and collaborations with other musicians. We show that the success of a musician defined by the follower count and popularity score can be determined with some precision by a complex relationship between multiple features defined in a collaboration network. We also find that traditional success measures such as appearing in Billboard's Hot 100 list can be predicted accurately with a simple classifier using the collaboration features that provide guidance for aspiring musicians on how to advance their networking profile. Using this knowledge, we then try to find what features and to what extent a musician should improve in order to become more successful in both digital and analog domains.

These predictions are further refined through the relation of genre labels to popularity values under a modified $Q$-model[3], showing a relationship between success, genre, and several logistic properties of top artists. In both cases, we apply a variety of classifier models and show that the accuracy raises up to 90% with tree-based classifiers using a combination of these properties, demonstrating these models' ability to predict success among top artists under Spotify-generated data. We then show that the task of predicting whether an artist will be featured in Billboard's Hot 100 list only by looking at the features of their collaborators can be learned with a tree based classifier with accuracy reaching up to 92%. We also find that collaboration with high-profile musicians is not always necessary for a musician to achieve success of appearing on a Billboard's Hot 100 list.

## Background

As a preface to our work exploring trends of success among modern music artists, we review similar efforts in quantifying the features to a variety of disciplines. Within the field of science of science, Wang et al. propose a temporal model for citation

patterns within research publications, showing similar release trends among successful papers[6]. Sinatra et al. extend this into a success-citation model, showing promising results towards quantifying and predicting scientific impact[3]. Success can also be studied in a focused manner, such as through trends on Nobel Prize recipients[7]. Apart from success, similar properties are evaluated using publication data, such as regularity within research interest growth over a career[8] and the quantification of such growth relative to success in a competitive setting[9].

Outside of the science of science field, similar methods and efforts are applied towards the quantification of success through specific properties in the fields of art[1,10], show business[2] one-hit wonders in creative fields[11] and within the music industry, both for success distribution and collaboration[12,13]. Within all of these efforts, generalized models are developed and trends are noted, commonly specific to the field and the properties studied, allowing for some basic prediction. Our work aims to contribute to these efforts by studying modern music within Spotify and applying prediction methods, similar to the work by South et al.[5]. Unlike the former, our work focuses on top Spotify artists and the application of classification models for prediction using separated sets of genre and collaboration data.

## Preliminaries

### Data

We first detail our collection of data. Our work is based on data from two datasets derived from a combination of data queried from the MusicBrainz database[14] and data collected using Spotify's API through the Spotipy Python package[15]. MusicBrainz is an open source music database that contains release information on roughly 1.9 million artists and around 3 million releases as of August 2021. Using these two databases, we were able to gather data on musicians active and music created between 1890 and 2020, tagged with 3031 genre tags supplied by the community. To focus on the release patterns within a specific market, we restrict the region of releases to the United States of America for both the collaboration network and the genre network.

### Artist Collaboration Network

First, we consider a dataset as the source of a collaboration network of musicians similar to the collaboration network constructed by South et al.[5] This dataset is initially seeded with a list of the top 1,000 artists by their stream count, provided by `Chartmasters.org`. Pulling data from Musicbrainz, we associate these artists with other artists based on their releases to develop a collaboration network. Data collection continued until the network had no more new edges, yielding 114,955 artists and 27,359,052 releases. Artists that did not have a Spotify profile were filtered out from the data. The resulting dataset contains 22,517 artists and 595,849 unique releases, disregarding duplicate releases outside of the US. For each release, we gather the release date, associated label and region from MusicBrainz and collect the release artists' popularity scores and numbers of followers from the Spotify API.

The counts of numbers of releases for musicians in the collaboration network are mostly low. As seen in Fig. 1(a), 90% of the musicians in the dataset have no releases. This is to be expected, since our data collection method focused on the collaboration history of an individual musician regardless of their release counts. About 40% of musicians have at least one and at most three official releases. This means that in our dataset only 6% of musicians have four or more releases.

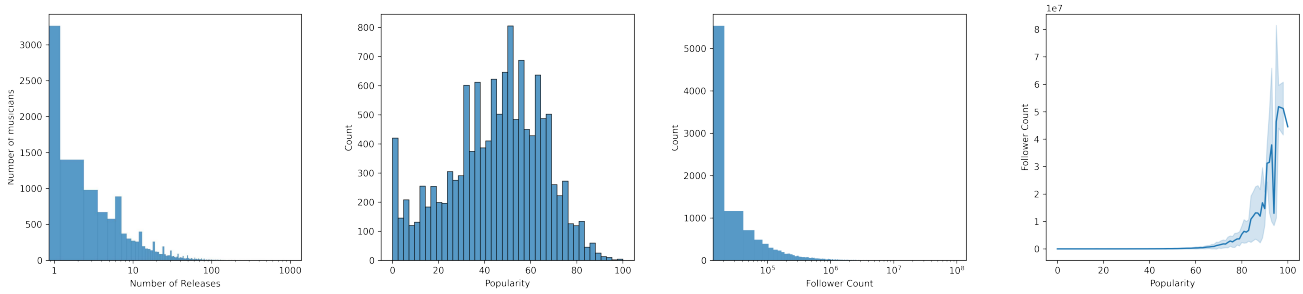


**Figure 1.** **a**. Release Count distribution of musicians. **b**. Popularity score distribution of musicians. **c**. Follower Count distribution of musicians. **d**. Follower count associated with popularity

#### *Popularity Score*

The popularity score is a metric provided by the Spotify API, with values ranging from 0 to 100. It is a measure used and calculated internally to rank musicians by their perceived popularity at a point in time. This score is affected by the number

of plays an artist has, as well as how recent those plays were. This metric provides an insight into the musicians' success at the time of measurement. Unfortunately, the historical data of musicians' popularity scores are not publicly available. Thus, we cannot track the change of popularity of a musician. Despite this, we use this score as a metric for viewing how popular a musician currently is due to Spotify's usage of the measure as a ranking baseline.

#### *The Number of Followers*

Spotify also provides the number of followers an artist has. This value can be any non-negative number, ranging from 0 to several millions. These numbers are defined by an internal function of the Spotify platform that allows users to create follow links to musicians that trigger Spotify notifications to such users about new releases by so followed artists. This measure is a relative indicator of how 'popular' an artist is and provides a baseline comparison metric between multiple artists. Furthermore, since this value is not affected by time, we can view the follower count as a more temporally robust measure of popularity compared to the Spotify popularity value. While this value is subject to change as users follow and unfollow artists, such updates cause minor changes to overall rankings among top artists.

#### *Billboard's Hot 100*

While Spotify's API provides users with two metrics to gauge an artist's success, high values of these metrics do not necessarily translate into traditional success. Billboard, an American music magazine, weekly publishes a list called Billboard's Hot 100, which is a ranking of songs that are considered to be the most successful in terms of two features, albums sales and radio plays. Spotify does not consider these two features when calculating their metrics. Since Spotify currently dominates in digital streaming of music, it can be argued that the features from Spotify are sufficient measures of an artist's success. However, digital streaming suffers from a wide variety of manipulations from its users, such as fans of a particular artist boosting their idol's stream counts in order to place higher in Spotify's rankings. For this reason, we also use the Billboard's Hot 100 ranking and its correlation to Spotify popularity and other features of an artist to find a more reliable measures of success in the music business.

### Release Genre

To separately consider genre-popularity relations among top artists, we extract data based on shared genres using methods similar to the first dataset to generate a secondary dataset. Pulling data from Spotify's listing of the daily top 200 songs from January of 2017 to November of 2021, we collected 2,363 artists with primary releases in the United States of America. From there, we reference and generalize Spotify's genre labels for each artist to associate each artist with a list of related release genres. It is important to note that an artist can be listed as part of multiple generalized genres. We consider each artist relative to 27 pre-determined genres, chosen based on frequency in the dataset such that the overflow label of "other" is not dominating. Note that an artist is assigned the label "other" when at least one of the genres in which the artist is listed under does not belong to any of the other 26 generalized buckets. These genres are associated with an artist as a binary vector of length 27. The distribution of artists over these generalized genres is shown in Fig. 2. This distribution was developed using data from both the

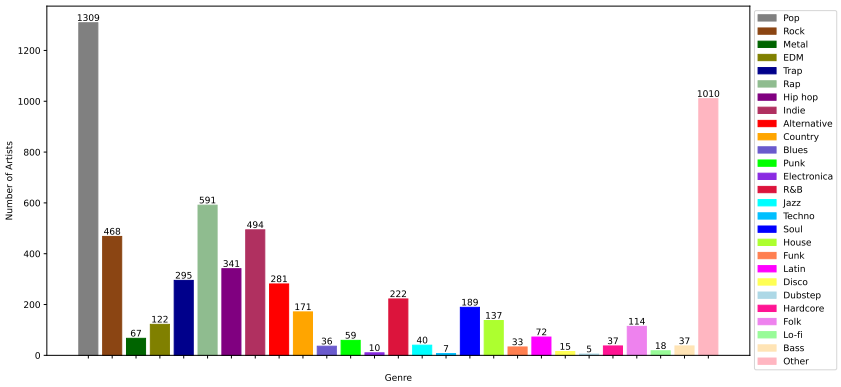


**Figure 2.** Distribution of the genre-focused dataset (2,363 total artists)

MusicBrainz database and the statistics within Spotify's API. Specifically, we associate an artist with their top 10 songs during November of 2021, their respective popularity measures, and their productivity, defined as their all-time total releases as of November of 2021.

# Methods

## Musicians Analysis

### *Collaboration Network*

The music industry is heavily dependent on collaboration. A relatively unknown musician can achieve fame quickly by collaborating with a well known artist. The form of collaboration may not always be mutual, as in some cases, a musician can re-interpret or be inspired by music created decades, or even centuries before. The information about collaborations between musicians can contain many nuanced details about their success. Hence, we construct a network of musicians linked by collaboration and extract features of such networks to be used in predictions.

More formally, a collaboration network $G_M$ between musicians is a directed weighted graph whose nodes represent a set of musicians, $M$, while edges $E$, each edge drawn between a pair of musicians $M_A, M_B$ with the weight $w(M_A, M_B)$ when musician $M_A$ has released records featuring music of musician $M_B$. Our musicians network is composed of 22,509 nodes and 595,849 edges. For the success metrics of each musician, the follower count ranges from 0 up to 90 million while the popularity score is limited to the range $[0, 100]$.

### *Network Rank*

To test the correlation between musician's position in the network and their various measures of success, we use a network rank measure based on eigenvector centrality. Since the popularity score ranges between $(0, 100)$, we simply group the musicians of equal popularity scores together and measure the standard deviation and mean of their network rank scores. Although the popularity score is convenient, its drawback is being a single value. In contrast, the follower count of each musician is a sequence of values measured at the same time for all artists. To consider the network rank score at various follower counts, we create 100 synthetic bins to group the musicians together, since the values do not have a set bound. The follower count has an exponential distribution, with most musicians on the lower end and only a few at the top. This follows the trend noted by J.A. Davies[12] in that the distribution of success among musicians tends to be spread exponentially. To normalize this data, we apply a logarithmic function to the number of followers and then categorize them at 100 intervals evenly divided between the minimum and maximum value.

In both comparisons for popularity and follower count with the musicians' network rank, we observe that the upper-mid range musicians tend to have the highest network rank as shown in Fig. S1 in Supplementary Mat Materials.

Interestingly, the artists who have the highest network rank scores are classical musicians. The top three musicians are Wolfgang Amadeus Mozart, Ludwig van Beethoven and Johannes Brahms, names that are familiar even to those who are not avid classical music listeners. After filtering out classic musicians by using the genre tags, we found that most musicians with high network rank scores were instrumental musicians, such as the London Symphony Orchestra or Philharmonia orchestra. In fact, when examining the genre tag mentioned at the top 100 network ranked musicians, we find that every genre is either classical or instrumental. The fact that musicians who produced/composed music over 200 years ago are still dominating the collaboration network can appear strange at first. Yet, this is understandable considering that sampling, which is an act of borrowing a piece of music from someone else while giving them credit, is now a prevalent method of digitally producing music. At the time of writing, neither Spotify nor Musicbrainz offered information on the type of collaboration, so we were not able to construct a more detailed network with different types of edges.

### *Musician Classification using Spotify*

After observing that there indeed is a pattern among popular/widely followed artists and their releases, we build a prediction model to identify artists who can be considered successful.

To each musician $M$, we assign 10 features derived from the Popularity Score, and the Number of Followers, as well as the same 10 and another 9 features derived from the Billboard appearances. All the assigned features are listed in Table S1 in Supplementary Mat Materials.

For predicting the popularity score and follower count, we leave out the features related to other Spotify metrics in order to remove bias in our classifiers. Features used for each target variable is marked with a V. Given these features, we set to build three classifiers that can predict the popularity score($M_{pop}$), follower count ($M_{fol}$) and the number of Billboard appearances. South et al.[5] use a model called Social Group Centrality, which groups musicians into different bins, such as celebrity, community leaders and masses based on their popularity and network rank. In our case, we are more interested in finding a model to predict the success metrics given the node features. Thus, we consider some well-known shallow machine learning models to learn the problem of predicting success. We only chose to look at shallow models because of their transparency, so that we can further understand model learning patterns and the relationships between features.

We first build two classifiers to predict the popularity and follower count of each artist. After confirming that we are able to learn a classifier for each problem, we then build a classifier to predict whether the artist will appear in a Billboard's Hot 100 list.

## Predicting success measures

### *Predicting Popularity Score*

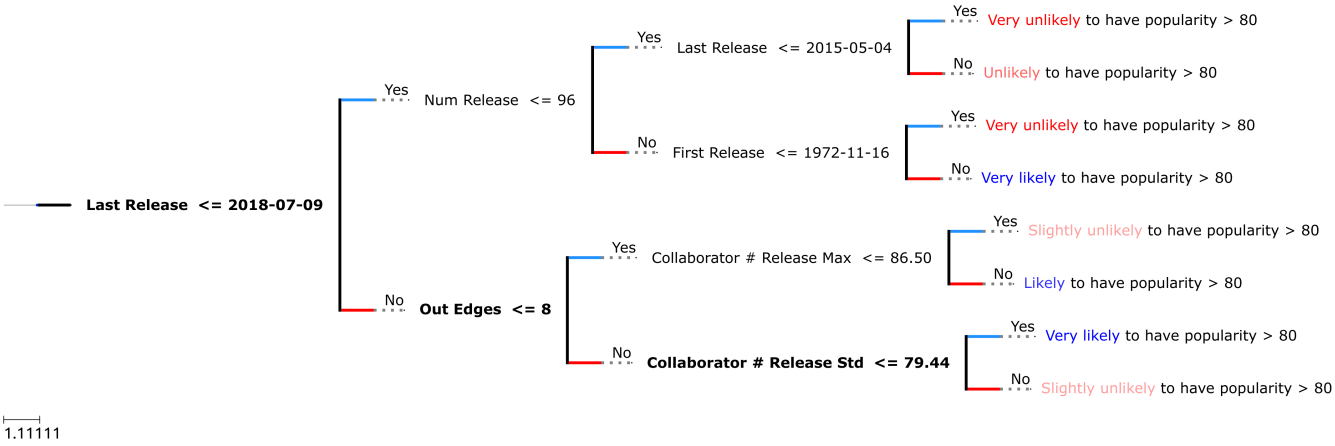


**Figure 3.** A Visualization of XGBoost tree classifying popularity $> 80$

We first look into building a classifier for the popularity score using our collected features. To limit the scope of the problem, we consider finding the threshold $\sigma_p$ of the popularity score that we can accurately predict. For a musician $M$, the label is 'true' if $M_{pop} > \sigma_p$ and 'false' otherwise.

In order to test different types of shallow ML models, we considered logistic regression, SVM, Random Forest and XGBoost classifiers. We chose these models because they are either simple or tree-based structures. This allows us to better understand each model's ability to identify the success of musicians.

In practice, we found that using linear models does not work with our data. This is in agreement with our observations that there are no linear correlations between the features. The ensemble models (Random Forest) also were not able to learn from the data, producing high training accuracy but falling below 50% in the testing step, suggesting high overfitting. However, we found that a single XGBoost tree appears to be very well suited for our data. Even at $\sigma_p = 80$ where the positive samples are only 2% of the data, the XGBoost classifier was able to classify the musicians with a high accuracy score that was constantly above 80%, while the other models classified performed worse than 50%. In order to compensate for the fact that the number of artists who had a popularity score above this threshold was very small compared to those who did not, we re-sampled our dataset by adding repeated entries of such artists to balance our dataset. We confirmed that this process did not result in an overfitting classifier by using five-fold cross validation and observing that the performance of the model was steady throughout the different folds.

Because XGBoost outperformed every other classifier we considered, we looked more into using this classifier to examine the features of a popular artist. XGBoost can be used like an ensemble model, training $N$ trees and aggregating their predictions. We found that reducing the number of trees in the classifier improves the performance, suggesting that there is an underlying simple structure that is being estimated by the many trees. In fact, XGBoost continued to work very well after we reduced it down to a single tree. Using this single tree as a classifier, we are able to get up to 83% accuracy in our testing step. A visualization of the tree is shown in Fig. 3, with the top three features marked in bold. Note that the feature values may not reflect the real life values, as they were normalized ahead of time. We also constrained the max depth of the tree in order to produce a model that can be easily viewed and understood. We found that when restricting the tree to the depth of at most three yielded the best result.

Unpacking the tree, we see that the follower count is the first feature considered. This is expected, as we noted a high correlation between follower counts and popularity score. In addition to these features, we see that the number of collaboration features also plays a role in the decision step. Specifically, we see that the collaboration features are considered for musicians in the mid-range of follower counts.

According to the results, we note two specific observations:

If a musician received **fewer** collaborations than a certain threshold, they are **highly likely** to have a popularity score **smaller than** 80. But if not, they are **slightly likely** to have a popularity score **greater than** 80.

If a musician gave out **fewer** collaboration than a certain threshold, they are **highly likely** to have a popularity score **smaller than** 80. But if not, they are **slightly likely** to have a popularity score **greater than** 80.

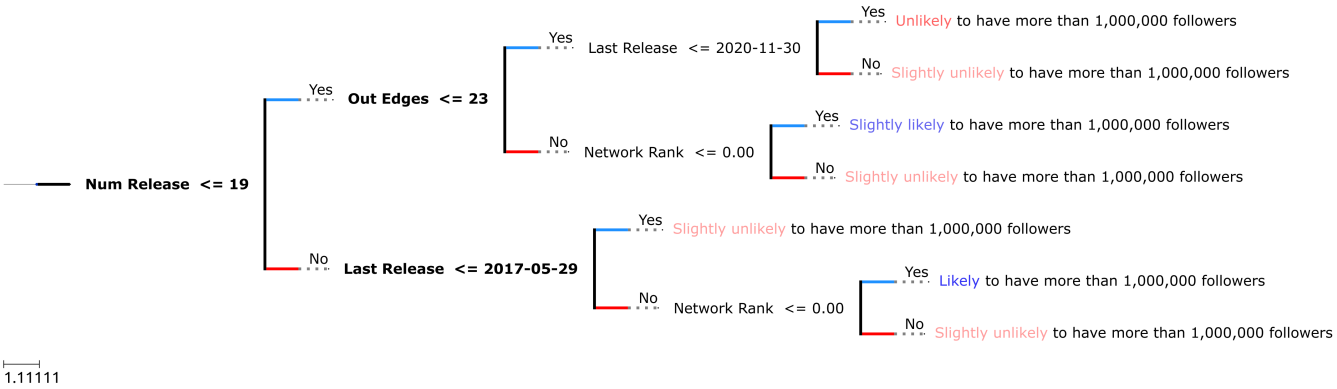


**Figure 4.** A Visualization of XGBoost tree classifying follower count $>$ 1,000,000

### *Predicting Follower Count*

Since the popularity score is highly related to the follower count, we trained another XGBoost tree to predict the follower count given the remaining features. Following the same procedure as training for the popularity score, we first picked a threshold $\sigma_f$ to divide the data into two sets. We used $\sigma_f = 1,000,000$ for this metric, with which 0.9% of musicians had a number of followers greater than this threshold. We found that the same model parameters as the popularity model performs just as well on the follower count task. Our classifier for the follower count with $\sigma_f = 1,000,000$ showed steady performance throughout the cross validation step, resulting in the final accuracy score of 88%. Fig. 4 shows a visualization of this classifier tree with the top three features marked in bold.

Unlike the classifier for popularity, we found that the classifier for follower count included features such as first release or a release count, while not considering collaboration features such as collaboration counts. Analyzing the nodes in Fig. 4, we observe that:

If a musician has a release count **less than** a certain threshold, the musician is **highly likely** to have **less than** 1,000,000 followers. But if not, they are **slightly likely** to have **more than** 1,000,000 followers

If a musician has career length **less than** a certain threshold, the musician is **highly likely** to have **less than** 1,000,000 followers. But if not, they are **highly likely** to have **more than** 1,000,000 followers

We observe that the follower count classifier does not follow the same decision process as the popularity score classifier. We found that while the collaboration of musicians has a large impact on the popularity score, the collaboration features are not considered nearly as much when predicting the follower count. This suggests that while collaborating with many other artists may temporarily boost one's popularity, the collaborations do not mean nearly as much when it comes to their overall success in the grand scheme.

### *Predicting appearance on Billboard's Hot 100*

After building a classifier for each of Spotify's provided metrics, we build a final classifier on predicting whether an artist will appear on a Billboard's Hot 100 list using the same collaboration network. We used the Billboard's Hot 100 list because it is widely accepted as a benchmark of success in a musician's career. While Spotify's metrics provide an insight of trends within the streaming business, we are interested in finding out if we can translate this to a more traditional metric of success which is the Billboard's rankings. Our data was gathered from September of 2021, so we only consider Billboard rankings after this date. We used the dataset provided by Kaggle[16] that contains songs that appeared in at least one of the Billboard's Hot 100 lists from 1958 until November of 2021. Following the same steps as the previous two classifiers, we first re-sample our dataset so that it contains equal numbers of artists who have and those who have not been listed. In order to focus on the collaboration features of the musicians rather than their inherent features, we excluded the artists' popularity and follower count in the considered feature set. Instead, we added extra features that described the collaboration history of the artist, such as the standard deviation, average and maximum of their collaborators' popularity and follower count. Using the same settings as previous classifiers, we were able to learn this task with a test accuracy of 90%. The learned decision tree can be seen in Fig. 5 with the top three features marked in bold.

It is interesting to note that we are able to predict the artist's appearance on a Billboard's Hot 100 list without looking directly at their Spotify popularity and follower count. Looking at the resulting classifier shown in Fig. 5, we draw another two conclusions:

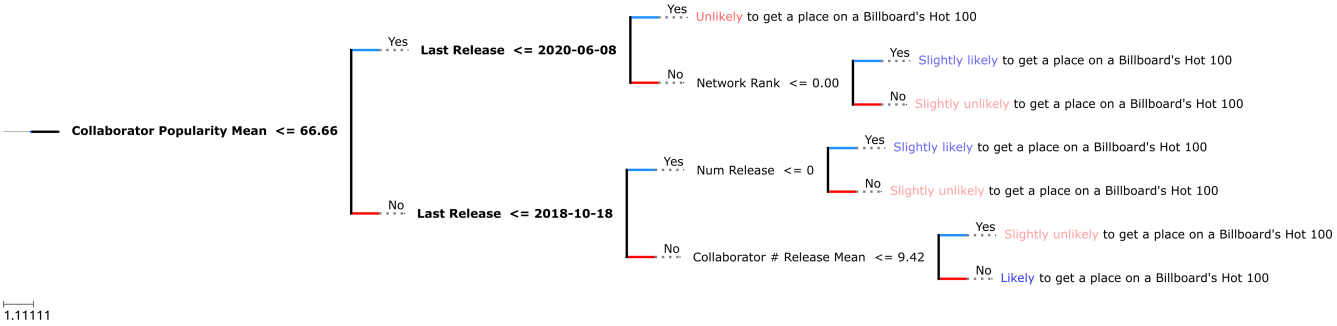


**Figure 5.** A visualization of XGBoost tree predicting artist's appearance on a Billboard's Hot 100 list

A musician who collaborates with **more popular** artists with last release **before** 2020, is **unlikely** to appear on a Billboard's Hot 100 list. But if not, they are **slightly likely** to appear.

A musician who collaborates with **less popular** artists whose collaborators have **less than** 10 releases is **slightly unlikely** to appear on a Billboard's Hot 100 list. But if not, they are **slightly likely** to appear.

## Applying learned models

Once we build decision trees for each success metric and identify the patterns for them, we test our observation by changing the feature values of synthetic musicians who are initially classified as 'not successful' to see what features affect each metric the most. In order to simulate real musicians while reducing our search space, we use K-Means clustering to generate 300 clusters of musicians who were labeled as **False** for each success measure. We then use the value of each cluster's center as our synthetic data points to change. After generating the clusters, we then apply 5000 random permutations to the feature values and test the resulting feature values on the XGBoost classifiers presented above.

For each metric, we show two examples of the same synthetic artist whose prediction changed with the least amount of perturbation applied (shortest Euclidean distance between the original and the perturbed features) and the most perturbed features (furthest Euclidean distance) out of the samples that were classified as positive.

### *Boosting Spotify popularity*

Out of the 300 synthetic musicians who were predicted to have a popularity score below 80, we are able to successfully transform 167 of them to be predicted to have a popularity score greater than 80. Out of the three metrics we examined, the popularity score is the metric, with which the change of the prediction outcome occurred for the largest number of perturbed samples. This suggests that the popularity score may be the easiest metric for an artist to change in a short amount of time among the three metrics considered in this work. In the examples shown in Figs 6(a) and 6(b), we find that the features that had the largest impact towards boosting one's popularity score are their collaboration features such as incoming/outgoing edges in the collaboration network and the release count statistics of their collaborators.

### *Boosting Spotify follower count*

Out of the 300 synthetic musicians generated to have follower count $\leq$ 1,000,000, we are able to convert 39 instances to be predicted to have follower count $>$ 1,000,000. In the example shown in Figs 6(c) and 6(d), we observe that the feature that matters the most for success is the artist's number of releases, in agreement with the observation made analyzing the corresponding decision tree.

### *Boosting Billboard appearance probability*

Perturbing the synthetic data points to be predicted to place on Billboard's Hot 100 list proved to be the hardest task among the three metrics we examine. This most likely also represents the reality of appearing on this list, because it is limited to 100 artists per week, whereas the Spotify metrics do not have a set limit. Out of the 300 synthetic artists, we are able to change the prediction outcome of 35 artists. As seen in the decision tree, we find that both the artists' intrinsic features and collaboration features need to be improved to be placed in the Billboard list. Interestingly, we find that the quality of collaborations, such as the average follower count of the collaborators, matters more than the sheer number of collaborations, as can be seen in the example shown in Fig. 6(f).

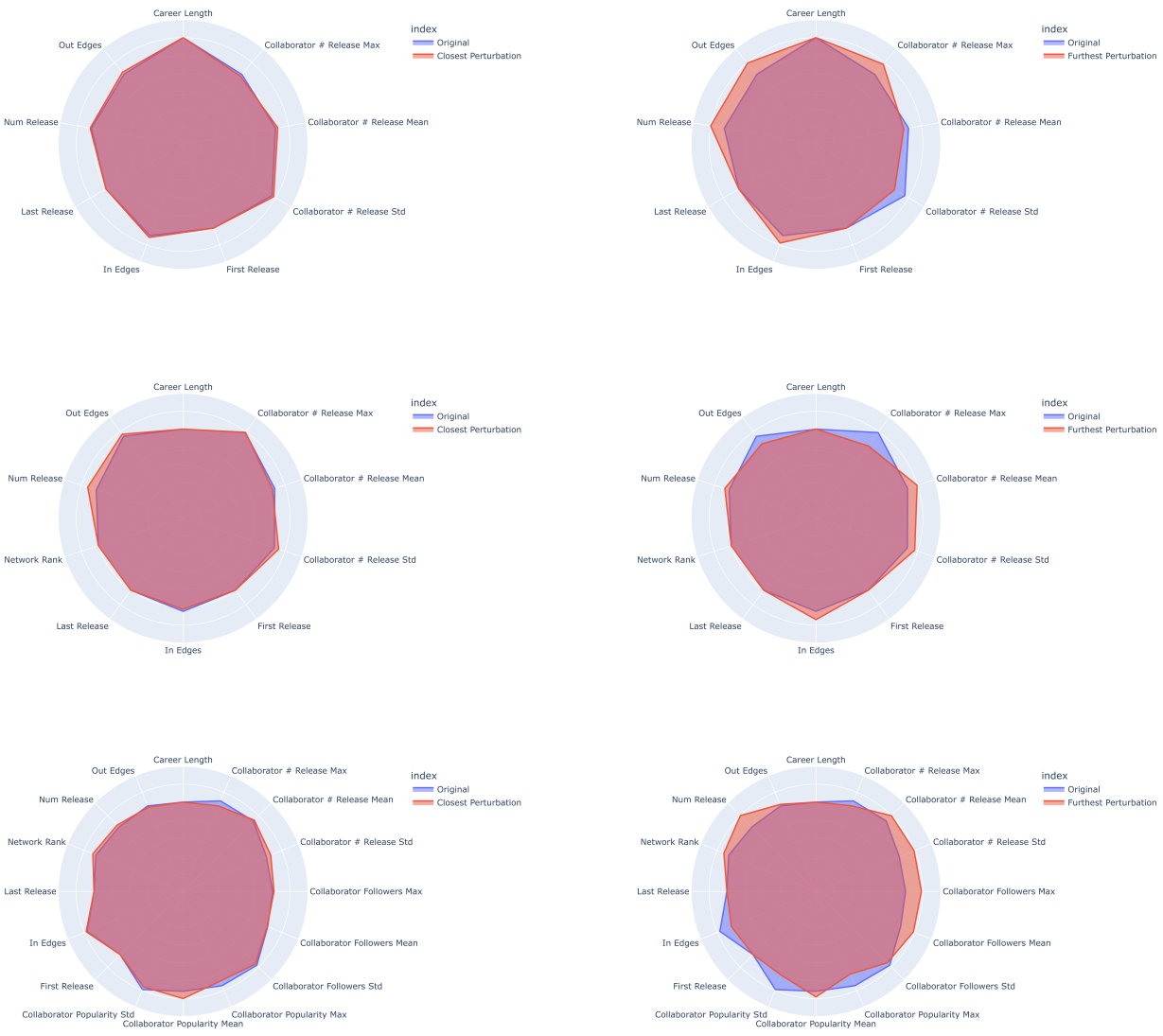


**Figure 6.** Results of applying perturbations to synthetic artists. Figs **a**,**c**,**e** show the least amount of perturbation applied and Figs **b**, **d**, **f** show the most perturbation applied that results in a prediction of (**a**,**b**) popularity $> 80$, (**c**,**d**) follower count $>$ 1,000,000 and (**e**,**f**) appearing on a Billboard's Hot 100 list

## Genre Analysis

Utilizing the genre-based dataset, we develop two network views focusing on the relation of top Spotify artists and their genre associations. These network views provide insight into the genre-popularity relation within Spotify's data and generalize a correlative consideration of success relative to genre association.

### *Q-Model*

Sinatra et al.[3] consider the quantification of success within academia by tracking citation numbers generated by each published paper within a set time period under a specific topic classification. Within this consideration, a number of models were considered to parameterize success relative to citation numbers and productivity throughout a researcher's career. The most successful model, titled the $Q$-model, considers a $Q$ parameter to act as a "luck" factor within a researcher's career. This parameter is noted to be intrinsic and relatively unchanging throughout careers. The $Q$ value is represented in the equation: $Q_i = e^{\langle log c_{10,i} \rangle - \mu_p}$. For our application, we extend such a definition to the field of music by using Spotify's popularity value in place of citation numbers ($c_{10,i}$) and by pulling a productivity value from an artist's total release from MusicBrainz. Note that

(a) Popularity-banded network view of $\pm 10$ (b) Genre similarity network view

**Figure 7.** Popularity-banded and Genre Similarity Network Views for genre-artist relation analysis

even though our $Q$ parameter is more restricted due to the bounded Spotify popularity value, the $Q$ value can be considered with a scaling parameter to compare with other fields or can be used by itself as a local measure relative to the dataset.

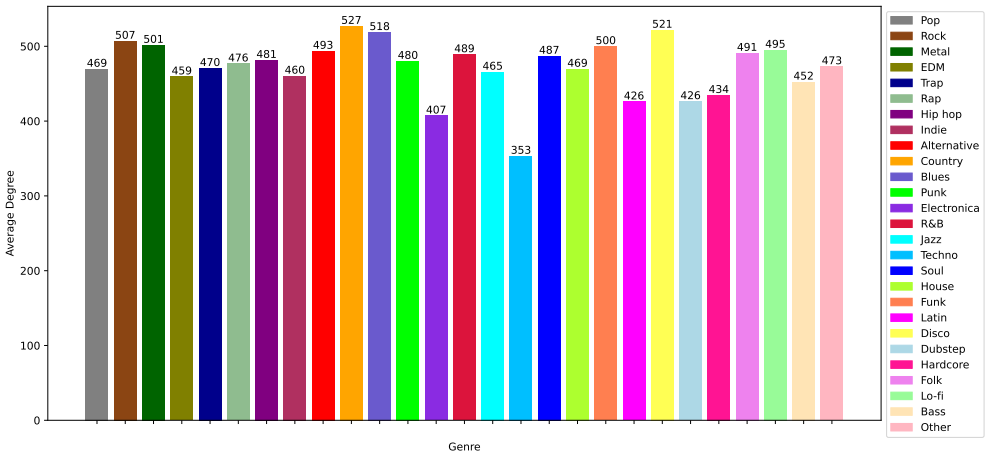


**Figure 8.** Average degree distribution for popularity-banded network

***Genre Similarity Network***

The first network developed relates artists based on the sharing of at least one genre label. Specifically, let $G_{gs}$ be an undirected graph with no multi-edges defined by the set of artists $N = \{a_1, a_2 ... a_{2363}\}$ where an edge exists between $u$ and $v$ if and only if the genre-inclusion set of $a_u$ and $a_v$, $GI_{a_u}$ and $GI_{a_v}$ are related such that $GI_{a_u} \subseteq GI_{a_v}$ or $GI_{a_v} \subseteq GI_{a_u}$. Each node $a_n$ is assigned two properties: a weight ranging from 0 to 100 dependent on the artist's average popularity and a color associated as a mix of

the artist's genre vector of releases. This network view is shown in Fig. 7(b). Notably, artist-genre relations tend to be diverse yet cluster around a significant group of artists lacking genre diversity. These clusters are connected to each other through a couple of key artists that fulfill a niche genre pattern.

#### *Popularity-banding Network*

The second network focuses on the relation of artists within averaged "bands" of popularity. We define this network, $G_{pb}$, with the same definition as $G_{gs}$ with the exception that an edge exists between $u$ and $v$ if and only if $a_u$'s average popularity value is within 10 popularity points of $a_v$'s. This results in a network of dense layers of artists separated into banded regions of Spotify popularity. Given the coloration of the artists relative to their release genres, such a network view provides insight into the distribution of genres among Spotify's popularity values given a listing of top artists and their data. This network view is shown in Fig. 7(a). An interesting feature of this network view is the apparent diversity of genres within each popularity band with a few lone nodes among the maximized popularity values. We further shed light on this diversity with the observation of average degree distributions within this network view, shown in Fig. 8. The average degree distribution is notably consistent across all genres with some variance in some of the niches, such as "newer" genres like Techno and Electronica. This consistency initially suggests a lack of relation between popularity metrics and production genre within this network view.

(a) Raw view of the Musicians Network

(b) Musicians with popularity $> 90$

**Figure 9.** Musician network views for collaboration analysis

## Results

Utilizing the aforementioned data and network structures, we draw conclusions regarding the correlation between musicians, their collaborators and their genre of music.

### Collaboration between musicians can be used to predict success measure

We define a network of musicians connected by their collaboration and use the network rank in combination with other features to build a predictor model for follower count and popularity. Instead of grouping musicians to different bins as done in[5], we use a regression approach to estimate whether the musician will meet the success criteria or not. We found that XGBoost models perform much better than other linear models, and that it can offer insights into the reasons behind musician's success.

Using the popularity score and follower count as two different success measures, we find that there is a high correlation between popularity scores and collaboration history - especially so for the popularity score. However, we find that follower count is also influenced by the standalone features, such as release count or the timestamp of the first or last release.

We find an interesting observation that classical artists have massive amounts of collaboration compared to other artists due to their music being sampled.

We also find that the more widely accepted standard of success, appearing on a Billboard's Hot 100 list, can be accurately predicted using the musician's release pattern and the features of their collaborators. Our findings show that aspiring musicians

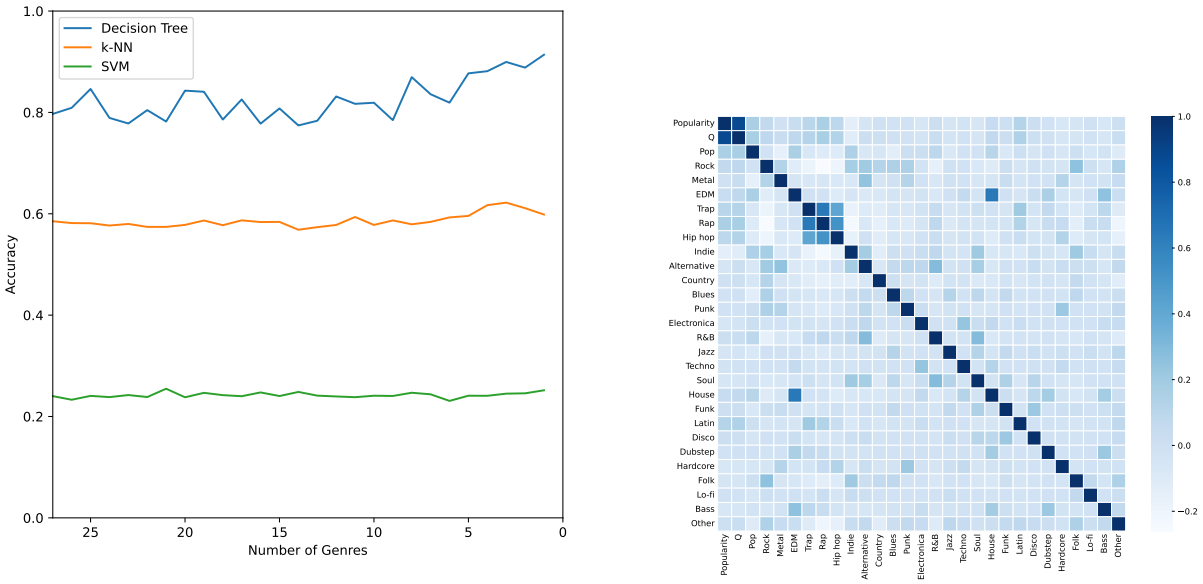


**Figure 10.** **a**. Genre prediction from $Q$ values, popularity measures and release counts, **b**. Correlation Matrix of Genre-focused Data

do not necessarily have to collaborate with the top stars of the industry to reach fame. In fact, we find that if the artist themselves and their collaborators are consistently releasing music, they are more likely to reach the Billboard charts rather than those who do not.

### ML models can be used to help musicians improve their visibility

While any of the three considered thresholds (follower count $>$ 1,000,000, popularity score $>$ 80, appearing on Billboard's Hot 100 list) cannot be said to be direct definitions of success, meeting these thresholds certainly help an artist to gain more visibility. Using the three decision trees that we are able to train from our dataset, we find that by improving certain aspects of their features, an artist can raise their probability of making themselves more successful. Our examples show that as suggested by the decision trees, one should focus the most on their collaboration for popularity and overall release count for number of followers. To appear on the Billboard's list, a musician should work on both aspects of their career, releasing more music while also collaborating with other prolific/well known artists.

### Primary release genres show no correlation to Spotify popularity measures under a $Q$-model

Seeking to draw conclusions relating music genres to the Spotify popularity score, we utilize the genre-focused data by applying the $Q$-model (modified for Spotify popularity ranges) to find a correlative measure. Thus, we compute this measure using all of the genre-parsed artists and their computed $Q$ values from their Spotify popularity measure along with a binary vector denoting genre inclusion. The results are shown in Fig. 10(b). We attribute the high correlation between popularity measure and the $Q$ value to the dependence within our modified $Q$-model. Notably, we see some strong correlations between specific genres such as Trap, Rap and Hip Hop, likely due to the commonality of their pairings for a number of artists and their general cultural overlap, which is to be expected. Unexpectedly, we observe close to no correlation between any of the genres, the Spotify popularity measure or the $Q$ parameter. This suggests that the popularity distribution within our genre dataset is relatively even across all genres, but from a $Q$-model perspective, there is no indication from the Spotify data that a specific genre will increase the likelihood of success in music relative to Spotify's popularity measure.

Finally, we examine the popularity distribution across our genre-parsed dataset. We present results in Fig. S3 in the Supplementary Mat Materials observing the mean and variability for each song and their respective genre with genre overlaps included. Thus, we see a relatively consistent midpoint around fifty and sixty popularity points for each genre, which is understandably above average given the usage of only top 100 artists in the dataset. These results are consistent with our correlation analysis, showing with some variance, that average Spotify popularity scores are consistent across the included genres. We note the interesting data point of a maximized popularity score among the Pop and Soul genres within our dataset,

demonstrating the top song on Spotify at the time as being of these genres. A further look confirms this observation since the song "Easy on Me" by Adele, which topped the Spotify charts soon after its release in late 2021. This observation prompts the interesting question of whether a classification of Top 100 songs is strict enough to see such correlation between genre and popularity. Thus, further analysis into strictly maximized popularity metric songs and their respective artists could show favored genres within a single popularity band of "hit songs."

### Genre prediction through machine learning methods shows potential for prediction properties

To further observe potential for correlation between artist average popularity values and genres, we employ machine learning methods to predict genres from our given data. We employ three types of classifiers for our genre-focused dataset: decision trees, k-nearest neighbors (k-NN) and support vector machines (SVM). For each model, we use a dataset consisting of each artist, their average popularity score from their top ten songs, their associated $Q$ value and their productivity. For training, we include a binary vector denoting genre inclusion for each artist, noting the possibility of multiple genres per artist. For accuracy, we measure the average accuracy of each model with a preset number of genres, picked uniformly at random. We show the results of each model in Fig. 10(a). Notably, we achieve a high accuracy of approximately 80% in correct genre prediction using decision trees, suggesting a complex, but strong relationship between popularity, $Q$ values and openness to genre inclusion. Furthermore, the consistently high accuracy values up to the drop around five genres indicates the importance of approximately five "broad" genres, as our dataset includes a number of "niche" genres in the scope of music culture. Thus, despite the lack of direct correlation between individual properties of each artist relative to Spotify measures, there exists some complex relationship and potentially predictability of the studied top artists from their genre to their success.

## Discussion

In this work, we seek to explore various ways we can represent the music business in a network structure, as well as the correlation between success metrics with the network features in combination with artist features collected from the Musicbrainz database and the Spotify API. We find that success, which is represented by Spotify's popularity score and number of followers, can be predicted using simple classifiers that are human-readable. We also find that there is not a strictly linear correlation between the artists' genres and their success. However, using tree-based classifiers, we observe that there exists a complex relationship between an artist's success and their genres. We find that classical musicians still permeate the music business hundreds of years after their time of release, and that success measures of artists can be predicted using publicly available data. We also find that while collaboration between artists may improve one's popularity score on Spotify, it does not impact the artists' overall follower count nearly as much. We also find that artists who collaborate with less known artists can also get a place on the Billboard, given their collaborators are prolific enough.

Our approach can be generalized and applied to any type of network data that has a target metric. It can also be extended to create a recommendation system for aspiring musicians, providing them with hints on how to raise their visibility on a given platform, such as recommended collaborations or number of releases. Future work can extend this by adding more data to the network or considering different classifier models. Spotify does not currently provide an 'all-time' measure for popularity or the number of streams each song has. Having access to this set of more detailed data may enable researchers to gain more interesting insight about what makes a musician more successful than their peers.

## References


1. Fraiberger, S. P., Sinatra, R., Resch, M., Riedl, C. & Barabási, A.-L. Quantifying reputation and success in art. *Science* **362**, 825–829 (2018).
2. Williams, O. E., Lacasa, L. & Latora, V. Quantifying and predicting success in show business. *Nat. communications* **10**, 1–8 (2019).
3. Sinatra, R., Wang, D., Deville, P., Song, C. & Barabási, A.-L. Quantifying the evolution of individual scientific impact. *Science* **354** (2016).
4. Ochi, V., Estrada, R., Gaji, T., Gadea, W. & Duong, E. Spotify danceability and popularity analysis using sap. *arXiv preprint arXiv:2108.02370* (2021).
5. South, T., Roughan, M. & Mitchell, L. Popularity and centrality in spotify networks: critical transitions in eigenvector centrality. *J. Complex Networks* **8**, cnaa050 (2020).
6. Wang, D., Song, C. & Barabási, A.-L. Quantifying long-term scientific impact. *Science* **342**, 127–132, DOI: 10.1126/science.1237825 (2013). https://www.science.org/doi/pdf/10.1126/science.1237825.
7. Li, J., Yin, Y., Fortunato, S. & Wang, D. Scientific elite revisited: Patterns of productivity, collaboration, authorship and impact. *J. The Royal Soc. Interface* **17**, 20200135, DOI: 10.1098/rsif.2020.0135 (2020).

8. Jia, T., Wang, D. & Szymanski, B. K. Quantifying patterns of research-interest evolution. *Nat. Hum. Behav.* **1**, DOI: 10.1038/s41562-017-0078 (2017).
9. Yu, X., Szymanski, B. K. & Jia, T. Become a better you: Correlation between the change of research direction and the change of scientific performance. *J. Informetrics* **15**, 101193, DOI: https://doi.org/10.1016/j.joi.2021.101193 (2021).
10. Franceschet, M. Art for space. *J. Comput. Cult. Herit.* **13**, DOI: 10.1145/3402443 (2020).
11. Berg, J. M. One-hit wonders versus hit makers: Sustaining success in creative industries. *Adm. Sci. Q.* **0**, 00018392221083650, DOI: 10.1177/00018392221083650 (0). https://doi.org/10.1177/00018392221083650.
12. Davies, J. The individual success of musicians, like that of physicists, follows a stretched exponential distribution. *The Eur. Phys. J. B* **27**, 445–447 (2002).
13. Janosov, M., Musciotto, F., Battiston, F. & Iñiguez, G. Elites, communities and the limited benefits of mentorship in electronic music. *Sci. Reports* (2019).
14. Swartz, A. Musicbrainz: a semantic web service. *IEEE Intell. Syst.* **17**, 76–77, DOI: 10.1109/5254.988466 (2002).
15. Lamere, P. Spotipy (2019).
16. Dave, D. Billboard "the hot 100" songs, DOI: 10.34740/KAGGLE/DS/1211465 (2021).

## Acknowledgements

We thank the Network Science and Technology Center at RPI for partial support of this work.

## Author contributions statement

I.K, M.M. and B.K.S. designed the study and wrote the paper with input from all authors.

## Competing interests

Authors declare no conflict of interests.

## Additional information

The data is available by sending a request to I.K. E-mail: inwon.kang04@gmail.com

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- suppmat.pdf