UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Mariana de Oliveira Santos Silva

Collaboration-Aware Hit Song Analysis and Prediction

Belo Horizonte
2020

Mariana de Oliveira Santos Silva

Collaboration-Aware Hit Song Analysis and Prediction

Final Version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Mirella Moura Moro

Belo Horizonte
2020

Mariana de Oliveira Santos Silva

Collaboration-Aware Hit Song Analysis and Prediction

Versão Final

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientadora: Mirella Moura Moro

Belo Horizonte
2020

Silva, Mariana de Oliveira Santos.

S586c Collaboration-aware hit song analysis and prediction [manuscrito] / Mariana de Oliveira Santos Silva. - 2020.
xxv, 96 f. il.

Orientadora: Mirella Moura Moro

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f.89-96.
.
1. Computação – Teses. 2. Ciência de dados – Teses. 3. Aprendizado do Computador – Teses. 4. Redes complexas – Teses. 5. Mineração de dados (Computação) – Teses. 6. Hit song science – Teses. I. Moro, Mirella Moura. II. Universidade Federal de Minas Gerais; Instituto de Ciências Exatas, Departamento de Ciência da Computação. III. Título.

CDU 519.6*75(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6ª Região nº 1510

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Collaboration-Aware Hit Song Analysis and Prediction

MARIANA DE OLIVEIRA SANTOS SILVA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROFA. MIRELLA MOURA MORO - Orientadora
Departamento de Ciência da Computação - UFMG

PROF. ANISIO MENDES LACERDA
Departamento de Ciência da Computação - UFMG

Renato Vimieiro
PROF. RENATO VIMIEIRO
Departamento de Ciência da Computação - UFMG

Michele A. Brandão
PROFA. MICHELE AMARAL BRANDÃO
Tecnológico em Processos Gerenciais - IFMG

Belo Horizonte, 30 de Março de 2020.

I dedicate this work to my family and everyone who believed in me.

Acknowledgments

My special thanks to all who encouraged me and contributed in some way to the development of this Master's dissertation. In particular, I would like to thank my parents, Luciana and Valdivino, for their immeasurable love and, despite the difficulties, for all the sacrifices that they have made on my behalf. To my sister, Gabriela, for continuous support, understanding, and friendship. To my boyfriend, Matheus, for providing me with unfailing support and continuous encouragement. I am also grateful to my other family members and friends who supported me along the way. This accomplishment would have been impossible without them.

I would also wish to express my deep gratitude to my advisor, Professor Mirella M. Moro, for her continuous support, motivation, enthusiasm, and immense knowledge. Professor Mirella's office door was always open to discuss problems or questions about my research or writing. She consistently guided me in the right direction whenever she thought I needed it. She has raised many precious points in our meetings, and I hope I have managed to address several of them here. Your guidance has helped me throughout the research and writing of this dissertation. I could not imagine having a better advisor and mentor.

I would like to express my sincere thanks to Professor Anisio M. Lacerda for the opportunity for collaboration, his helpful and constructive recommendations, suggestions, availability, and assistance throughout the completion of Chapter 5 of this dissertation. Without his guidance and persistent help, this dissertation would not have materialized. My thanks are extended to everyone with whom I had the pleasure of working during this and other related projects. In particular, I would like to thank my labmate, Gabriel P. Oliveira, for his immense help and his collaboration and contribution to various projects related to this dissertation. I am also grateful to my former labmate, Lais Mota, for assessing me with the initial insights for this research.

Finally, I would like to express my deepest appreciation for the excellent support from UFMG and the staff of DCC for being constantly available to solve questions. Thanks are also due to the financial support of CNPq in the form of a scholarship.

"Life's short.
Anything could happen, and it usually does, so there is no point in sitting around thinking about all the ***ifs****,* ***ands*** *and* ***buts****."*
(Amy Winehouse)

Resumo

As músicas de sucesso são mais bem-sucedidas do que a média, onde fatores-chave tornam essas músicas qualitativamente superiores às outras. As técnicas atuais para prever músicas de sucesso exploram recursos que descrevem músicas individualmente. Propomos abordar esse problema de previsão através de uma forma multimodal, com a fusão de recursos musicais. Especificamente, descrevemos as músicas através de recursos de três modalidades: *música*, *artista* e *álbum*. Inicialmente, identificamos perfis de colaboração em uma rede musical composta por artistas de sucesso, revelando como os artistas se conectam profissionalmente pode impactar significativamente seu sucesso. Para aprofundar essas análises, usamos séries temporais e o teste de causalidade de Granger para avaliar se há uma relação causal entre perfis de colaboração e popularidade dos artistas. Finalmente, modelamos o problema de previsão de hits como duas tarefas distintas: *classification* e *placement*. A primeira é um problema clássico de classificação binária de aprendizado de máquina e é uma aplicação direta de nossas estratégias de fusão. A posterior é uma abordagem de modelagem que posiciona uma música em relação a um determinado ranking, prediz músicas de sucesso e fornece informações comparativas de popularidade de um conjunto de músicas. Além disso, enfatizamos os perfis dos artistas colaboradores como características importantes ao descrever suas músicas. Estudos empíricos extensos, usando diferentes *features* de cada modalidade, mostram a eficácia de nosso método que combina dados heterogêneos para ambas as tarefas.

Palavras-chave: Hit Song Science, Ciência de Dados, Aprendizado de Máquina, Redes Complexas, Mineração de Dados Musicais.

Abstract

Hit songs are more successful than average, where key factors make such songs qualitatively superior to others. Current techniques to predict hit songs exploit features that describe songs individually. We propose tackling this prediction problem through a multimodal form with songs' features fused, together. Specifically, we describe songs through features from three modalities: *music*, *artist* and *album*. Initially, we identify collaboration profiles in a musical network composed of successful artists, unveiling how artists professionally connect can significantly impact their success. Then, to deepen such analyses, we use time series and the Granger Causality test for assessing whether there is a causal relationship between collaboration profiles and artists' popularity. Finally, we model the *Hit Song Prediction* problem as two distinct tasks: *classification* and *placement*. The former is a classical machine learning binary classification problem and is a direct application of our fusion strategies. The latter is a modeling approach that ranks a song relative to a given chart, then predicts hit songs and provides comparative popularity information of a set of songs. Furthermore, we emphasize collaboration artists' profiles as important features when describing their songs. Extensive empirical studies using various features from the modalities confirm the effectiveness of our method, which fuses heterogeneous data for both tasks.

Keywords: Hit Song Science, Data Science, Machine Learning, Complex Networks, Music Data Mining.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The fast evolution in technology continues to drive changes in the way people discover and engage with music content. According to IFPI's Global Music Report, global annual revenues of physical music decreased from $23.3 billion units sold in 2001 to less than $5 billion units, as summarized in Figure 1.1. In contrast, revenue from digital music, mainly subscription and streaming services, have been regularly increasing in the past years. For example, in 2018, the music industry experienced steady and consistent growth with overall volume up 9.7% over 2017, driven by a 34% increase in paid streaming that offset track and album sales declines[1]. In the first half of 2019, this creative industry saw a six-month record of over 507 billion on-demand streams[2]. Moreover, total revenues from recorded music in the United States grew 18% to $5.4 billion at retail in the same period[3]. In such a huge industry, becoming successful is challenging, but can lead to millions in revenue.

[1]IFPI Global Music Report 2019, (January 01, 2020), bit.ly/GMR-2019
[2]2019 Nielsen Music Mid-Year Report, (January 01, 2020), bit.ly/mid-year-report-2019
[3]2019 RIAA Music Mid-Year Revenues Report, (January 01, 2020), bit.ly/riaa-mid-year-2019

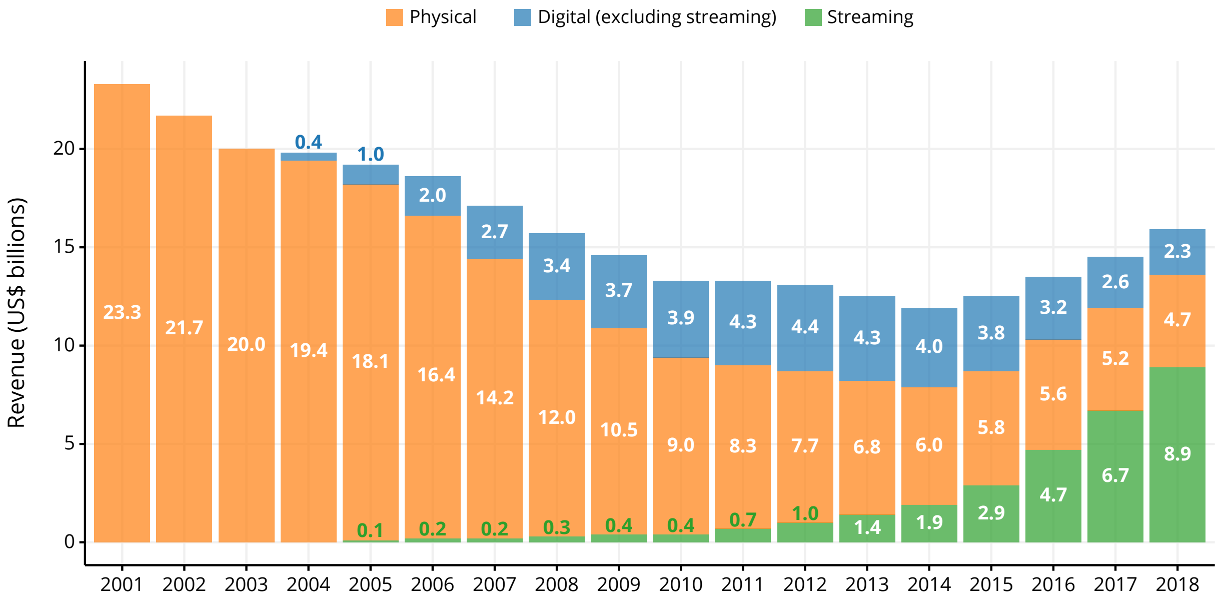


Figure 1.1: Global Recorded Music Industry Revenues 2001-2018 (US$ Billions). Source: IFPI Global Music Report 2019, (January 01, 2020), bit.ly/GMR-2019.

Perhaps the biggest challenge for music professionals in the industry is working with unpredictability. When releasing a single, professionals may face a complex task when trying to captivate different audiences. Besides offering unpredictable value to its listeners, one song can lead to a new *hit*, such as “Hello” for Adele or “Poker Face” for Lady Gaga. Moreover, some hit songs become acknowledged masterpieces, while other hit songs fall into oblivion as so-called *one-hit-wonders*. Such information implies an intriguing question: *what are the reasons for a song to achieve success and remain it for so long?* Discovering such reasons may lead to predicting whether a song will become popular, increase sales of physical and digital albums, improve the billing of on-demand audio streams services, or even help to predict the next music star.

Indeed, the ability to predict musical success offers huge benefits for many domains and audiences. For music industry CEOs, it may help maximize expected success by helping to decide whom to invest in to produce potential hits. Also, by properly investing in potential artist/music and distribution, the studio could increase sales of both physical and digital albums, improve revenue from on-demand audio streaming services or even launch the next *popstar* or *Summer hit*. Artists may also profit by identifying the most suitable songs to lead the album to early stardom. For music consumers, it may help to decide if an album is worth buying because it may potentially contain three to five hits, instead of being an *one-hit* only. This ability is the intrinsic motivation for a relatively new field of research known as *Hit song science* (HSS), which [51] define as “an emerging field of science that aims at predicting the success of songs before they are released on the market”.

HSS research is still in its infancy, and current attempts to solve the prediction problem are far from consensus [28, 81]. Not surprisingly, there has been an increasing interest in studying this burgeoning field. Its premise is that popular songs have similar attributes that make them appealing to people. Such attributes could then be explored to automatically predict whether a song will excel in popularity charts. However, there have been failed attempts at effectively learning hit song prediction by analyzing different factors that lead to the success of a song, with some researchers even concluding that such a task is impossible or not science [52, 2, 38, 81]. Such unsatisfactory outcomes do not mean, however, that popularity cannot be learned from analysing its intrinsic content or other resources.

In this dissertation, our premise is: music represents a multimodal item, and each of its facets may be mapped to a popularity aspect (e.g., melody, harmony, rhythm, artist’s reputation, musical collaboration profiles, and album information); next each aspect can be addressed, compared by statistical methods, and analyzed by an automated learner for a final decision on the popularity of a particular song. Overall, the goal is to *determine whether or not we can predict hit songs before they are released through a collaboration-aware characterization analysis and prediction.*

1.1 Motivation

Hit songs still make up the majority of a record company's profits, and as a result, labels invest billions into finding talent in hopes of achieving that gold. Creating a reliable scientific measure to predict whether a song offers the potential to become popular and commercially successful is a powerful and lucrative endeavor. One of the earliest studies on Hit Song Prediction [13] focuses on extracting acoustic and lyrical features from songs and using standard classifiers to separate hits from non-hits. The purpose was to determine if such a task was feasible or if the claims about the songs considered hits could be confirmed. A few years later, [52] stated that some subjective categories such as style and character mood of the song can be reasonably learned through techniques, but not popularity. However, after this study, many authors persisted in discussing the feasibility of predicting the popularity of a song by considering a suitable set of attributes.

Other approaches explore complementary perspectives on the same problem. For example, [45] highlight the changing musical tastes, then leading to an evolving popularity pattern. [11] also insert the video-clip of a song as a characteristic capable of affecting popularity. Moreover, [7] point out factors such as the preferential attachment by the artist, and [39] highlight the psychological parameters on the reasons why people prefer and are willing to listen to certain tracks. Likewise, [37] use Billboard rankings specifically for Rock music, analyze the complexity of a song based on audio signals to measure its impact on the popularity, and examine the popularity of a song track in the early days of release to predict future popularity. [29] confirm that the breadth of features leading to the popularity of a song exceeds the content of the track (audio and lyrics). Finally, [57] predict the popularity of a song by focusing on the social network Last.fm. They investigate three factors that could impact on popularity: lyrics content, artist's reputation and social context.

Even with such a diverse background, existing research in the area agrees that, besides the complex features to be measured, there are quantifiable qualities that contribute to the popularity of a song. Nonetheless, the range of song attributes that lead to popularity exceeds its intrinsic content, i.e., its audio-based features. Although popularity is affected by internal factors [13, 55, 28], some external factors have been ignored. For example, how artists connect professionally may be one of those untreated aspects. Indeed, artist collaboration is a strong force driving music today, as digital media enables musical collaboration among various artists. Specifically, according to data from the Billboard Hot 100, the number of songs executed by more than one artist has been experiencing a significant increase in recent years, as depicted in Figure 1.2**A**. By the end of 2018, these collaborations represented about 40% of the hit songs on average. In 2019, this number reached about 42% of songs on the chart. There is a recent trend in increasing even more

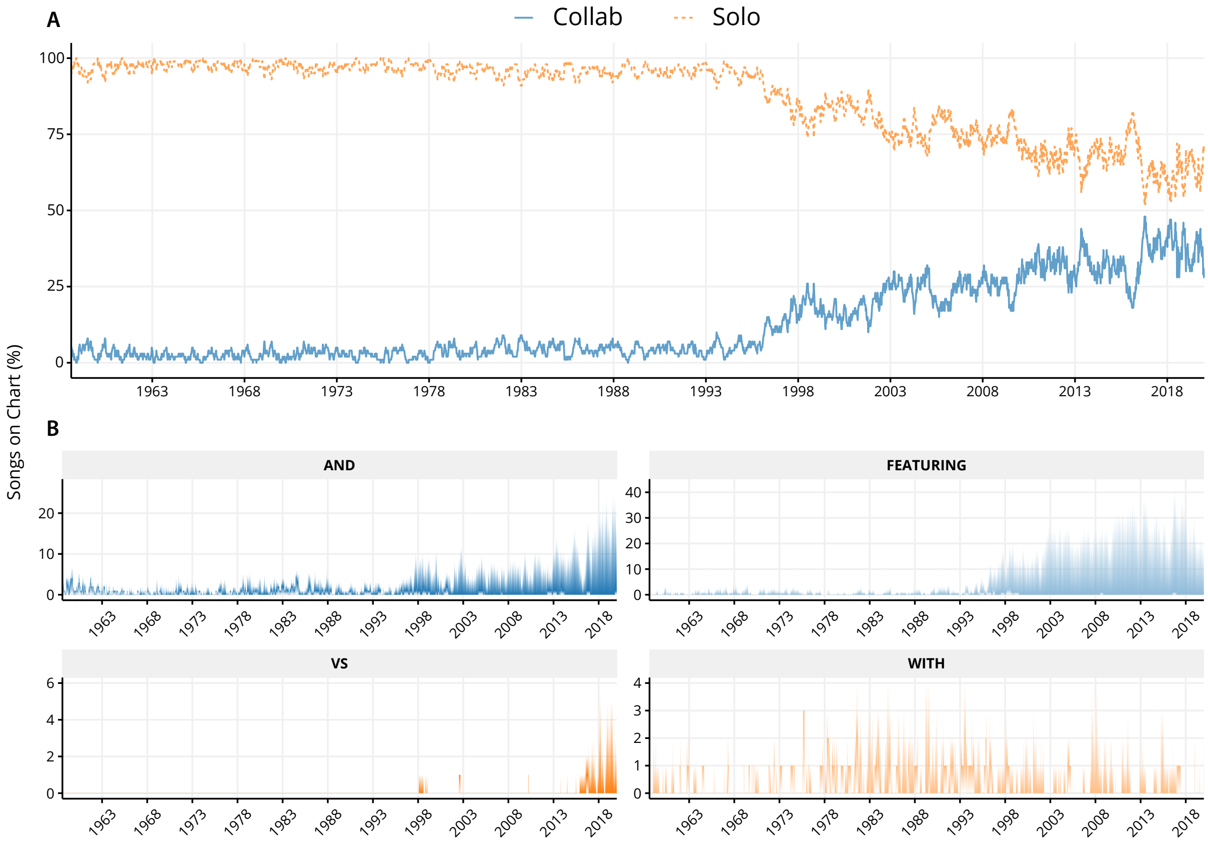


Figure 1.2: Billboard Hot 100 songs (1958 - 2020). (**A**) The dashed line represents hits without collaborations, while the solid one reflects musical collaborations. Clearly, collaborations are increasing in the US music industry with nearly half of all mapped music constituting collaborations. (**B**) Pairwise historical frequency of collaboration types, including *and* (set of artists with equal rights), *featuring* (when two artists collaborate on the same song), *slash*, *vs.* (DJ contest) and *with* (duet).

in the upcoming years.

In fact, in the 1960s, 70s, and 80s, collaborations were relatively rare (about 5% of charted songs) and generally took the form of duets (Figure 1.2**B**). The boom in collaborations started in the mid-1990s, when the number of collaborations increased significantly, with the duets dying and the artists on display taking over. Although such associations help artists bridge the gap between styles and genres then crossing over to new fan bases, they often run the risk of sounding artificial and non authentic as defined by professional critics. Nevertheless, there is no denying that collaborations are on their way to conquering the popular music domain. Despite such a rise in musical collaborations, factors driving the success of a collaborative process are not entirely understood [62, 27]. Therefore, investigating how music collaboration profiles can positively impact an artist's popularity is still worth pursuing. For instance, while talent and status attract social connections, the researchers usually ignore that social networks can independently promote success [44]. Still, previous research assesses the beneficial influence of social information and artistic quality in the Hit Song Prediction problem [59, 56, 28]. Such studies encourage the idea that non-musical factors can fulfill a key role as a popularity driver.

1.2 Research Goals

Music involves many features related to the composition (melody, harmony, rhythm, lyrics) and the social context (reachability and style of the artist, collaboration profiles, culture, etc.). With so many variables at hand, predicting song popularity remains a complex task that requires achieving a balance between different musical feature sources. Learning from heterogeneous information can offer the possibility to capture correspondences between different features and gain an in-depth understanding of musical success. Hence, this general objective can be divided into four specific Research Goals (RG), defined as follows:

Research Goal 1 (RG1). Identify the (potentially) topological measures and indicators that influence the popularity of both songs and artists (Chapter 3);

Research Goal 2 (RG2). Investigate the impact of these features on popularity over time, i.e., dynamically analyze whether features affect the popularity of an artist/song (Chapter 4);

Research Goal 3 (RG3). Verify the causal relationship between collaboration profiles and music success (Chapter 4); and

Research Goal 4 (RG4). Propose a machine learning approach to derive a song's popularity based on these groups of features and determine the best way for combining them to predict the success of a song (Chapter 5).

1.3 Main Contributions

The main contributions are summarized according to each RG, as follows.

RG1

1. We detect artists' clusters and their respective patterns of collaboration, focusing on analyzing the impact of such profiles on successful music artists;

2. We define four main categories of collaboration profiles: Interaction, Distance, Influence and Similarity; and

3. Our experimental analysis provides evidence that (*i*) there are, in fact, distinct success factors for musical collaboration profiles that are socially measurable, and (*ii*) there are common factors for successful collaboration in the music market.

RG2 and RG3

1. We explore the causal relationship between collaborative profiles and artist's success, by conducting two analyses in parallel based on daily time series of collaborative songs and musical success, and solo songs and musical success;

2. We validate that artistic popularity strongly affects establishing musical collaborations in a diverse profile;

3. We also confirmed some results obtained in the first research objective.

RG4

1. To the best of our knowledge, we are the first to define Hit Song Prediction problem as two independent tasks: *binary classification* and *hit song placement*;

2. We model Hit Song Prediction as a multimodal machine learning problem by considering three modalities: *music*, *artist* and *album*;

3. For both tasks, we evaluate the performance of our proposed representations and interpret the learned models to identify the most important features. Shortly, our empirical results show

   a) our proposed multimodal representation for songs outperforms state of the art algorithms in the Hit Song Prediction problem; and

   b) the features extracted from the *artist* modality, mainly collaborative information, are the most significant predictors of songs popularity.

1.4 Organization

The remainder of this work is organized as follows.

- **Chapter 2:** We review related work and summarize the main differences that distinguish this dissertation from previous research.

- **Chapter 3:** We identify collaboration profiles in a musical network composed of successful artists, arguing that the way in which artists professionally connect with each other can significantly impact their success.
- **Chapter 4:** We present deeper analyses using time series and Granger causality test for assessing whether there is a causal relationship between collaboration profiles and artist's popularity.
- **Chapter 5:** We present an unified framework for predicting musical success by using machine learning methods.
- **Chapter 6:** We conclude this dissertation by highlighting our findings and pointing out the future research directions.

Chapter 2

Related Work

Research over the music industry typically falls within the areas of recommendation systems [75, 74, 31, 77], human-centered computing [21, 26] and music retrieval [42, 47]. Nevertheless, ongoing efforts still try to solve a relevant and complex problem: *Hit Song Prediction*. Over the years, the number of artists and musical productions has considerably increased; and so has the number of attempts to discover the recipe for turning a song into a hit. Indeed, there are plenty of analyses on factors that potentially influence musical success from varied perspectives. These analyses are part of the emerging field of science known as *Hit Song Science* (HSS). In this chapter, the related work is divided into sections referring to the HSS (Section 2.1) and the final considerations, indicating the points where our work differs from the others (Section 2.2).

2.1 Hit Song Science (HSS)

Predicting song's popularity before its release is especially important for the music industry, as it allows improving revenues by focusing on potential hits. It may also help to identify key factors for a song to become popular and commercially successful. This task drives an emerging research field that aims to predict a song becoming a chart-topping hit, *Hit Song Science* (HSS). In HSS, popularity is regarded as a feature of a song, and the problem then is to map this feature to other resources that can be measured objectively [51]. Especially, related work follows two main directions. The first focuses on extracting general acoustic and lyric-based features (Section 2.1.1). However, such studies disregard the hypothesis that the song's popularity can be achieved indirectly through, for example, the reputation of the artist/album, social context, collaboration profiles, and so on. In this sense, the second direction considers more subjective information, that is, social information (Section 2.1.2). Such studies usually focus on a very specific social attribute, hence restricting the analysis of factors' on general musical success. Despite existing studies, there is still room for improvement in Hit Song Science and, subsequently, there is still much to gain.

2.1.1 Acoustic and Lyric-Based Features

In one of the earliest studies in the field, [13] extracted acoustic as well as lyric-based features, and then used standard classifiers to separate hit songs from non-hits. The authors show lyric-based attributes are slightly more useful than acoustic features for Hit Song Prediction problem. However, according to [52] and [51], the idea that song popularity can be predicted from such technical information contradicts the natural intuitions of any musically-trained composer. They describe a large-scale experiment to validate the state-of-the-art methods' ability to predict the popularity of musical titles based on global acoustic or human features. Both studies suggest acoustic features commonly used for music analysis are not informative enough to offer judgments on notions related to subjective aesthetics. Furthermore, they suspect that the previously cited study [13] is based on spurious data or biased experiments.

Following a similar path but contesting the results found by [52] and [51], [50] argue the viability of popularity prediction once it considers a set of relevant attributes. The authors investigated the UK top 40 singles chart from the past 50 years, distinguishing the top 5 from less popular (peak position 30 - 40) songs. The experiments show positive results using Machine Learning algorithms. Based on such a work, the site *Score a hit*[1] was created. Similarly, [23] used basic audio features, as well as more advanced features that capture a temporal aspect to tackle the dance hit song prediction problem. They explore several different classifiers to build and test prediction models. They obtained a good performance with logistic regression (AUC score: 0.81 and Accuracy: 80%). Along these lines, [37] use data collected from the Billboard Rock Songs Chart to investigate Hit Song Prediction problem. They analyze the song complexity based on audio signals, including *chroma*, *rhythm* and *timbre*, as well as the early stage popularity. They found that both groups of features (i.e., complexity and early stage popularity) are effective for different popularity patterns and combining the two types of features can be synergetic.

Recently, there were also attempts to apply deep learning to predict whether a song can be a hit. For instance, [79] used Convolutional Neural Networks (CNNs) with the Mel-spectrogram[2] of a song as the input for feature learning to predict a song's *play-count* in a music streaming platform. Their results unveiled that CNNs are indeed more

[1] Score a hit: `www.scoreahit.com`

[2] Mel-spectrogram is a Spectrogram with the Mel Scale (a perceptual scale of pitches judged by listeners to be equal in distance from one another).

effective than shallow models in predicting musical success. To provide a more comprehensive understanding and deeper insight into predicting music popularity based on acoustic information, [38] build classification models considering conventional acoustic features including MPEG-7 and Mel-frequency cepstral coefficient (MFCC) features. Their results show that, although there is still room for improvement, it is feasible to predict the popularity metrics of a song significantly better than random chance based on its audio signal. Recently, [81] assessed the potential success of a given song exploiting low- and high-level audio features and model Hit Song Prediction as a regression task. Particularly, using a wide and deep neural network model enabled the proposed approach to outperform baselines as well as approaches using low- or high-level features individually.

Unlike the previously described works, [3] tackle the problem of predicting hit songs as a classification task based on past information from the Spotify Top 50 Global chart, as well as acoustic features. Their main objective is to predict whether a song will be successful in the future, by making predictions in long term. The authors noticed that when considering acoustic information, the model's performance does not improve with a statistical significance, indicating that acoustic information may be completely overlooked. Finally, [43] introduced an innovative multimodal end-to-end Deep Learning architecture for predicting popularity in music recordings named as *HitMusicNet*. Their conducted experiments outperform previous studies by incorporating three musical modalities: audio, lyrics and metadata. Therefore, such findings validates the benefits of adopting multimodal strategies in prediction tasks.

2.1.2 Listener's Information

Previous knowledge of a song's success or about community's preferences can influence the musical taste of listeners. This was exactly the phenomenon studied by [59] through an impressive experiment. The authors created an artificial "music market", where $14,341$ participants downloaded unknown songs with or without knowledge of the choices of previous participants. Their conclusions confirm the hypothesis that social influence contributes both to inequality and unpredictability in cultural markets. Moreover, the real reason of experts fail to predict success is: when individual decisions are subject to social influence, markets do not simply aggregate pre-existing individual preferences. That is, social factors may play an important role in determining whether a song will be popular or not. Recognizing the importance of social influence in musical popularity, related work proposes predictive models for hit songs using only social information, ignoring the intrinsic features of songs. In particular, [11] investigate statistical patterns in people's musical tastes without considering the attributes of music track content. They also predict how long an album would remain in the popularity charts as well as position a new album on a chart in a certain week in the future by using sales data from the first few weeks. Their main findings include some interesting correlations, one of which emphasizes the role of marketing. Specially, good investment on marketing before starting sales of an album is crucial, since the data shows the higher the starting position of an album, the longer it is likely to stay in chart.

Likewise, [7] propose predicting success of songs through exploring social interactions. According to the authors, the success of a hit depends entirely on two factors: (1) its initial popularity observed after one week, and (2) contextual information of the album, the general popularity of the artist and the popularity of other tracks present in the album. Their method is based on data extracted from Last.fm and from the relationship among tracks, artists and albums. That is, their approach also does not use any information concerning the actual content of the songs. The method achieves good results in terms of AUC score, about 28% improvement compared with previous comparable work.

In a different scenario, [33] compare peer-to-peer file sharing information from songs with the popularity given by Billboard charts. The experiments indicate popularity trends for the songs on Billboard having a strong correlation with their respective popularity in the peer-to-peer network. Based on such a result, the authors propose a methodology that uses the aforementioned correlation to predict the success of a song. Focusing on *blogosphere* information (i.e., blogging behavior patterns), [1] investigate how blog posts can be used to predict the success of music and movies. Their experiments showed that traditional machine learning algorithms successfully learn to predict the trends of movie box office revenues and Amazon Sales Ranks, with a precision of 79.84% and 59.7%, respectively.

Still on social media information, [32] propose to collect the behavior of Twitter users-listeners based on hashtags related to songs for predicting popularity rankings. The reported results show high correlations between behavior in listening to music by Twitter users and the trend of song popularity in general. Following a similar approach, [57] predict the popularity of a song by focusing on the social network Last.fm. They also investigate three factors of a song that could impact its popularity: music content, artist's reputation and social context. Their main findings indicate that the content of the music is an important determinant of a music track's time duration in terms of weeks of popularity online. In addition, social attention of music listeners is another important determinant of future popularity online based on what happens during the early stage of a music track's diffusion. [64] also use data from social networks spanning music as well as books, photos, and URLs. Their results reveal that predictive models using temporal features achieve higher accuracy on various item types (network structure, early adopters' features and similarity) than all other feature types combined. Another study using Twitter users-listeners features was conducted by [2]. Based on comments posted on Twitter for 30 days before a given album released, the authors found it is possible to estimate the performance of an album. However, considering the Billboard ranking, they were unable to identify a statistically significant correlation.

In contrast with most aforementioned studies, [28] included songs both in and out of the top charts, endowing hence a larger predictive power. With such a diverse dataset, the authors correlate success with acoustic features and explore the predictability of musical success. Furthermore, they added a non-acoustic feature, the 'superstar' variable (i.e., whether the song's artist had appeared in the top charts recently), which greatly improve the prediction accuracy. These findings suggest social factors can play a significant role in the success of songs. Finally, with a different final task, [80] also use artist-based features but to explore popularity prediction for artists. They compared the performance of different methods including Support Vector Machines (SVMs) and long short-term memory (LSTM) neural networks.

Collaboration Networks. On a broader perspective, explaining or predicting the success of creative individuals through social network analysis has been a hot topic for decades. In a breakthrough work, [20] suggests the topology of an individual's social network impacts on personal success. Following approaches also focus on social collaborations [73, 8]. The connection between network topology and success was also found in the musical context. [73] analyze a network of creative collaboration among Broadway musical artists from 1945 to 1989. By applying statistical methods, they find that network measures significantly affect creativity in terms of financial and artistic success. In another setting, [8] measured creative collaboration in a music community where individuals compose songs together through overdubbing. The authors evaluate the relationship between metrics related to song- and individual-related measures and the likelihood of a song being *overdubbed* (i.e., add a track to a recording).

2.2 Overall Considerations

In general, approaches that consider only acoustic features and/or lyric-based information do not imply a proper musical representation for the prediction task (Section 2.1.1). Such unimodal representations disregard how external factors influence the popularity of a song. For example, the artists' reputation, social influence, the historical context or any other extrinsic feature that led a song to have a peak of success and become a hit. Further studies investigated music success with a joint representation by including non-musical factors, and the relationship between musical and social information (Section 2.1.2). Mostly, they found that the prediction accuracy can be improved when incorporating external factors. However, such studies generally focus on a specific social attributes which can represent limited musical representation.

Overall, our work differs fundamentally from those as we go further by introducing collaboration-aware modalities towards hit song prediction. Moreover, there is still no systematic and temporal study on collaboration information in the musical success context that could efficiently contribute to a song being a hit or not. In fact, most of the proposed works on HSS are limited to considering specific contexts (a given genre or a certain period); focusing on one or two attributes (acoustics and/or on lyric content), or relying on data of a particular social network. Hence, there is a strong potential for modeling music success through a multimodal and suitable combination of heterogeneous data.

In this sense, our study is innovative because it proposes to merge the music content and subjective social attributes and collaboration influence in order to predict success. This becomes explicit in Table 2.1, which shows a comparative analysis of existing HSS research studies. We are the first authors to consider collaboration profiles as predictors to assess the Hit Song Prediction problem. Furthermore, our work is the first one to employ the Learning to Place machine learning approach in the HSS context.

Table 2.1: A comparative analysis of existing Hit Song Science research studies

Task	Musical Features	Social Information	Year	Reference
Classification	acoustic and lyric-based features	*none*	2005	[13]
Experimental	artificial music market	listeners' information	2006	[59]
Regression	albums' average lifecycle	*none*	2006	[11]
Classification	global acoustic and manually-entered labels	*none*	2008	[52]
Classification	P2P queries	listeners' information	2009	[33]
Classification	listeners' music tastes	listeners' information	2009	[7]
Classification	blog posts, genre, artist and album information	listeners' information and artists' metadata	2010	[1]
Classification	acoustic features	*none*	2011	[50]
Classification	meta-information, basic acoustic features and temporal features	*none*	2014	[23]
Regression	music listening behaviors	listeners' information	2014	[32]
Classification	audio signal features	*none*	2015	[37]
Classification	music content, artists' reputation and social context	artists' reputation and social context	2016	[57]
Classification	data from social networks	listeners' information	2016	[64]
Regression	tweets	listeners' information	2017	[2]
Regression	acoustic features	*none*	2017	[79]
Classification	acoustic features and artists' reputation	artists' reputation	2018	[28]
Classification	acoustic features	*none*	2018	[38]
Classification & Regression	user songs operations and relations between songs and artists	listeners' information	2019	[80]
Classification	popularity information and acoustic features	*none*	2019	[3]
Regression	low- or high-level audio features	*none*	2019	[81]
Classification & Regression	text, audio and metadata	*none*	2020	[43]
Classification & Placement	**music-, artist- and album- related features**	**collaboration profiles and artists' metadata**	**2020**	**This dissertation**

Chapter 3

Collaboration Profiles and their Impact on Music Success

To further our understanding of how music collaboration affects artists' popularity, in this initial part of our research, we analyze and identify collaboration profiles in a musical success-based network; that is, a network composed only of successful artists. By detecting communities within this network, we identify collaboration profiles and analyze the impact of such profiles on musical success. Considering topological metrics, we define four main categories of collaboration profiles: *Interaction*, *Distance*, *Influence* and *Similarity*. Among them, we find that the first three affect musical success more intensely than *Similarity*. These findings provide evidence that (i) there are indeed distinct success factors for music collaboration profiles that are socially measurable, and (ii) there are common factors to successful collaboration in the music market. Such findings are important to motivate and set foundations for the contributions of this work, as described in the upcoming chapters.

Next, this chapter is organized as follows. In Section 3.1, we summarize fundamental concepts used in this work. In Section 3.2, we describe the dataset and introduce the proposed methodology. In Section 3.3, we detail the results and experimental evaluation. Finally, in Section 3.4, we discuss the overall considerations.

3.1 Fundamental Concepts

In this section, we present some fundamental definitions and concepts necessary for understanding this chapter. We start by describing all seven topological metrics used to characterize collaboration. Next, we summarize three types of correlation coefficients.

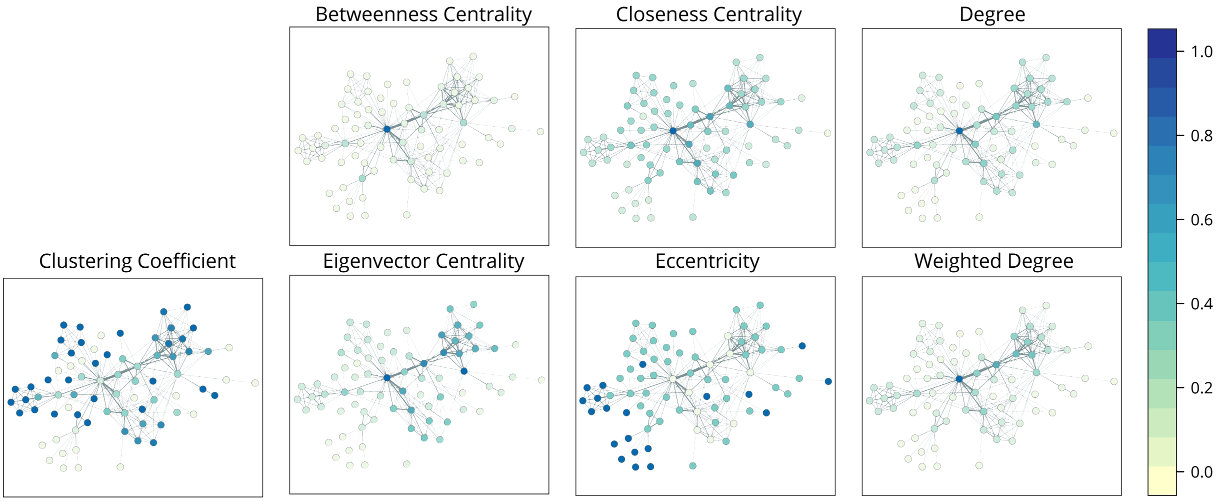


Figure 3.1: Topological metrics of a synthetic social network with 77 nodes and 254 edges. The nodes are colored by each topological measure.

3.1.1 Topological Metrics

In our study, we consider well known metrics related to individual graph nodes (Clustering Coefficient, Eigenvector, Degree and Weighted Degree) and to the whole graph (Closeness, Eccentricity, and Betweenness). We refer to related literature for complete definitions [49]. Nonetheless, we briefly define and put each in the musical context, whereas Figure 3.1 presents a synthetic social network colored by each measure's values.

Clustering Coefficient. Also known as the transitivity measure, it captures the degree to which the neighbors of a given node bind to each other. In other words, it measures the density of local links in the graph: the more interconnected the neighborhood of a node i, the greater its local clustering coefficient C_i. In the music context, the metric captures how connected an artist's collaborations are.

Eigenvector Centrality. It measures the influence of a node within a graph. It calculates the centrality degrees of each node in the graph, but not always with the same equivalence, as the central nodes that connect to other central nodes receive more weight than those that do not. Here, it gives more weight to artists who collaborate with central artists (that is, other highly collaborative artists).

Degree and Weighted Degree. Degree is the number of edges incident on a given node. Weighted degree is defined similarly by summing the weights of the incident edges. In our network, the degree measures the number of connections to other artists (nodes), while the weighted degree measures the number of interactions.

Closeness Centrality. It is a node centrality index based on the shortest path. Specifically, the value given to each node corresponds to the average of the shortest path between that node and all others in the graph. High (low) values of closeness should indicate that all other nodes are close (distant) to the one being measured. That is, the closeness of a node can be interpreted as a measure of the possibility of an artist to be relevant (in terms of collaboration) for several other artists, but with the possibility to be irrelevant for few others as well.

Eccentricity. It is also a node centrality index. For each node n, it calculates the shortest path between n and all other nodes in the graph, then the "longest" shortest path is chosen. Having such path identified, its reciprocal is calculated. High values indicate a positive meaning in terms of node proximity: if the eccentricity of n is high, all other nodes are close; if it is low, there is at least one node (and all its neighbors) that is far from n. The eccentricity of a node in a musical collaboration network can be interpreted as the easiness of an artist to be reached by all other artists in the network to collaborate in the future.

Betweenness Centrality. It is a different way to measure the node centrality using the shortest path by considering how many times a given node appears in the shortest path of the other nodes in the graph. High scores mean that the node is crucial to maintaining node connections for certain paths. Such connectors have the potential to be highly influential by inserting themselves into the dealings of other parties. In the music context, it indicates the capability of an artist to bring in communication with distant artists.

3.1.2 Correlation Coefficients

Correlation tests evaluate association between two variables, which can be performed using correlation coefficients. The Pearson correlation coefficient is the most commonly used, and is a parametric test recommended for normally distributed variables. Otherwise, non-parametric Kendall and Spearman correlation tests should be used. These methods are summarized as follows.

Pearson's correlation (r) [54] measures a linear dependence between two variables (x and y). The coefficient (r) varies between -1 (perfect negative correlation) and 1 (perfect positive correlation). The value 0 indicates there is no linear correlation, but does not guarantee the independence between the variables.

Spearman's correlation (ρ) [69] does not give any assumptions about the distribution of the data, but analyzes the appropriate correlation when the variables are measured on a scale that is at least ordinal. The method calculates the correlation between the classification of variables x and y. Its interpretation is similar to Pearson's: the closer ρ

is to ± 1, the stronger the monotonous relationship is.

Kendall's correlation (τ) [30] is a non-parametric test that measures the statistical dependence between two variables (x and y). Intuitively, the τ correlation between two variables is high when the observations return a similar (or identical, with $\tau = 1$) classification, and low when the observations have a different (or completely different, with $\tau = -1$) classification.

3.2 Methodology

This section presents our methodology for analyzing the impact of collaboration on musical artists' success. The dataset and results of such analyses are crucial to the next steps in our research.

First, we describe the process of building a dataset in Section 3.2.1. Then, we propose our definition of successful artists in Section 3.2.2. Next, we model a social network in Section 3.2.3. Finally, in Section 3.2.4, we perform semantic characterization to group similar topological metrics (summarized in Section 3.1.1) in categories, to later use them to detect collaboration profiles.

3.2.1 Data Collection

Billboard is a weekly American magazine specialized in music. Its website provides countless internationally recognized rankings that classify songs and popular albums. To model a success-based network of collaborations, we collected all the artists on Billboard's Artist 100[1] chart, a weekly ranking that lists the top 100 artists. We use the *billboard.py*[2] Python API for access Billboard's rankings and perform the data collection. In total, we collected 211 rankings between 2014 (July 26, 2014) and 2018 (July 28, 2018). As each chart consists of 100 artists' names, we collected 21,100 artists' names in total, which were deduplicated to achieve 1,135 distinct names.

For more information on each artist, we also collected data from Spotify, one of the most popular and used music streaming platforms. With Spotipy library[3], we obtained

[1] Billboard's Artist 100: `www.billboard.com/charts/artist-100`
[2] *billboard.py*: `github.com/guoguo12/billboard-charts`
[3] *Spotipy*: `spotipy.readthedocs.io`

full access to all of the music data provided by the Spotify platform. Then, for each artist collected on the Billboard chart, we also collected features and her/his ten most popular songs. In the same way, the artists (and their features) that collaborate in the execution of the top 10 songs were also collected. Therefore, the total number of artists grew to 2,152. For each artist, the following information was recovered[4]: the Spotify ID, name, popularity (a numeric value between 0 and 100, with 100 being the most popular), the total number of followers, and a list of the genres the artist is associated with. For example, Adele (the English singer-songwriter) has a popularity score of 87, about 20 million followers, and is associated with genres *british soul*, *pop* and *uk pop*. In contrast, Joss Stone (another English singer-songwriter) has a popularity score of 64, only 775,778 followers, and is associated with *pop rock*, *neo soul* and *british soul*.

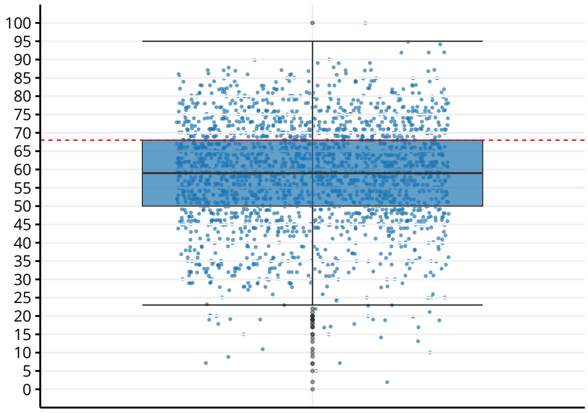
3.2.2 Definition of Success

To evaluate the association between collaboration and the artist's success, we must first define artists' success, which may correspond to popularity on social media and music platforms, sales profit, awards, etc. Here, we use two metrics collected from Spotify to define it as: *a successful artist presents both a high level of popularity and a large number of followers. To establish a threshold, we use the upper quartile of each success metrics' distribution (popularity and number of followers).*

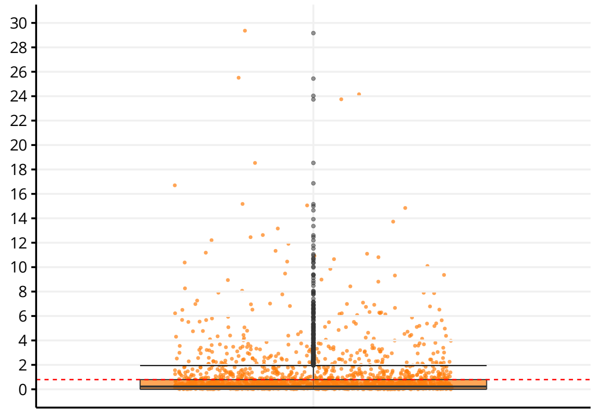
In Figures 3.2a and 3.2b, the upper quartiles of the popularity and number of followers distributions are around 70 and $1,000,000$, respectively. Specifically, 25% of artists have a popularity greater than 72, but not necessarily more than 1 million followers. For instance, the American singer Janelle Monáe has a popularity score of 70, but only $511,119$ followers (that according to data collected on January 2019; after her killing opening act at the Oscars 2020, her number of followers is above $800,000$ as of March 2020). Similarly, 25% of artists have more than 1 million followers, but they do not necessarily score 70 or more in popularity. A great example is the American singer Prince with $2,493,242$ followers, but 69 in popularity.

Furthermore, Figure 3.2c shows the intersection between both measures. That is, in our dataset, there are 354 artists who have a popularity greater than or equal to 70, as well as at least 1 million followers (e.g., the Canadian rapper Drake, with a popularity score of 100 and $23,732,186$ followers). Finally, we formally define the success s_i of an artist i as high if his/her popularity index $p_i \geq 70$ and number of followers $f_i \geq 1,000,000$.

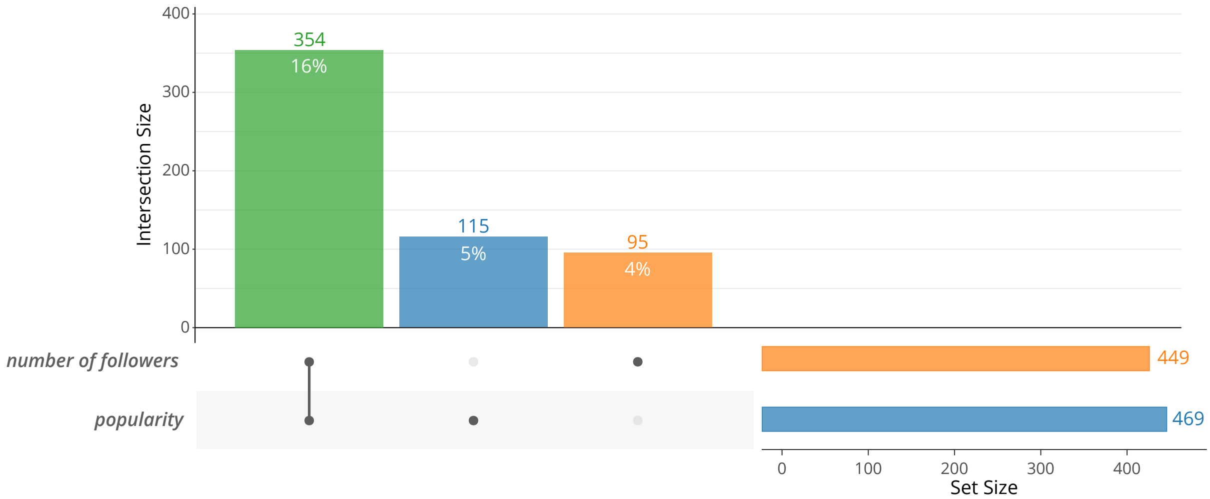
[4]Full dataset openly available at `bit.ly/apoena_datasets`



(a) Popularity



(b) Number of Followers (millions)



(c) Intersection between number of followers and popularity

Figure 3.2: (a,b) Boxplots of success measures. The boxes represent the lower and upper quartiles while whiskers show values below the 25th and above the 75th. (c) UpSet plots depicting quantitative intersection of the sets of each success measure across artists. The numbers above the bars show (left) the intersection of artists presenting both a high level of popularity and a large number of followers; (center) the set of artists with only a high level of popularity; and (right) the set of artists with only a large number of followers.

3.2.3 Social Network Modeling

To model the music collaboration network, we collected all the artists who participated in the execution of a single —either as participation (featuring) or collaboration (with). This is because, aside from Spotify not distinguishing the artists' role in a song execution, all the artists who collaborated in it represent an essential part in the success of a single. For example, "Despacito" (the most played song in history through streaming platforms as of March 2020) is stored as "Luis Fonsi *featuring* Daddy Yankee and Justin Bieber"; whereas "Under Pressure" (Queen's second number-one hit in their home country, after 1975's "Bohemian Rhapsody") is stored as a *collaboration* between Queen and

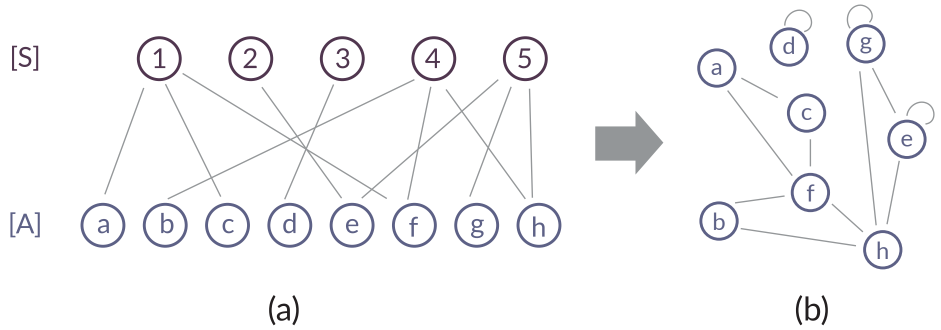


Figure 3.3: Bipartite graph projection: the two-mode social network for songs S_{1-5} and artists A_{a-h} (a) is projected onto a single one-mode social network (b).

Table 3.1: Collaboration social networks' statistics

	Original Network	**Filtered Network**
# Artists	2, 152	354
# Songs	10, 706	2, 144
# Collabs	5, 335	922
Modularity	0.793	0.478

David Bowie.

The dataset was modeled as a bipartite graph with nodes for songs and artists and edges connecting the individuals who collaborated in the execution of each song. As most network analysis techniques cannot be applied to bipartite graphs, we project the social network into an one-mode model, as explained next. In the bipartite graph, there are two groups of nodes: S the set of songs; and A the set of successful artists who collaborated in the execution of the songs present in S. Following the methodology proposed by [48], the bipartite model is then designed as a unimodal non-directed graph: every two nodes in A are connected by a link if they are connected to the corresponding node in S (in the bipartite representation). In other words, only artists are present as nodes, and edges exist between artists who worked on the same song in such a projection. In addition, songs that do not have more than one artist in their execution are modeled as self-loops on the artist's node. Figure 3.3 shows a simple example of such projection.

Before proceeding to the next step, we perform a network filtering for keeping only successful artists (definition in Section 3.1.2) on the network and getting rid of potential outliers. For instance, Lacy Mandigo, a contestant on Season 10 of the American reality talent show named *The Voice*, would be filtered out for scoring a popularity of only 10. Hence, the network includes only artists with a popularity rating greater or equal than 70, and at least 1,000,000 followers. A good example is the Canadian rapper Drake, the most popular artist in our dataset. Table 3.1 presents statistics of the originally collected network and the new, filtered one: total number of artists, total number of songs, total number of musical collaborations, and network modularity.

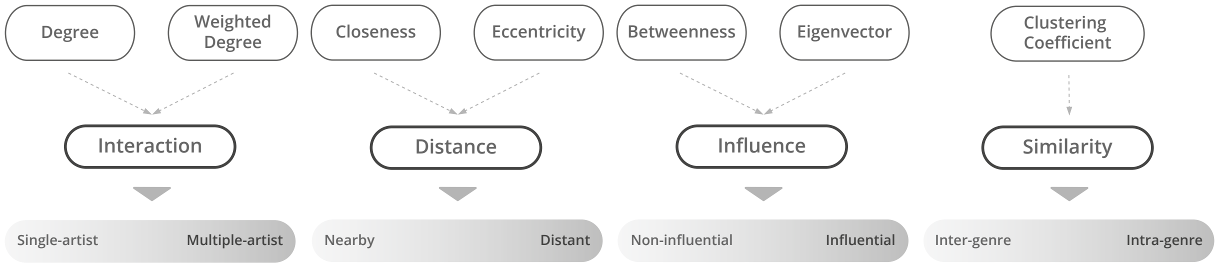


Figure 3.4: Semantic characterization of four collaboration categories (*Interaction*, *Distance*, *Similarity* and *Influence*), their composing metrics and magnitude range.

3.2.4 Semantic Characterization

From the topological metrics (Section 3.1.1) of the modeled collaborative success-based network, we define four categories that semantically characterize collaboration. Each category has two levels of subcategories indicating its magnitude. They are defined according to the specific social metrics' features as follows.

1. Interaction (degree and weighted degree). This first category is based on the artist (node) connectivity and is directly related to Degree and Weighted Degree metrics (quantify the number of connections/interactions of a node). High values of these metrics suggest a collaboration composed by more than one artist; likewise, low values indicate non-collaborative artists. Thus, the category ranges from single-artist (or band) to multi-artist collaborations.

2. Distance (closeness and eccentricity). The *Distance* category is based on proximity by Closeness and Eccentricity metrics. Note that Closeness should always be compared to Eccentricity [60]: a node with high eccentricity and high closeness is very likely to be central in the graph. In fact, the values of such metrics can be considered as an "average tendency to node proximity or isolation" [60]. Therefore, high values indicate pivotal artists in the graph, which are close to the other nodes. This suggests central artists in the network collaborate with individuals from different places; in contrast, isolated nodes (i.e., that are locally connected) represent artists who collaborate with fellow countrymen. In other words, the concept of proximity here refers to the reachability of different cultures, indicating a greater geographical distance between connections. Thus, the artist can have a nearby (low degree of proximity, therefore less central in the network) or a distant (high degree of proximity, more central) collaboration.

3. Similarity (clustering coefficient). *Similarity* is the most complex category and related to the Clustering Coefficient (captures the density of edges in the neighborhood of a node/artist). The more local links within the neighborhood of a node, the higher its local clustering will be. In our case, a link represents a musical collaboration. Therefore, an edge can be defined by a collaboration between artists who are present in the same professional/social circle. It is safe to say that these associations are usually made up of artists from the same musical genre, where they often share the same producer, events, audience, and so on. Consequently, a high clustering coefficient indicates similar connections (probably, of the same musical genre); furthermore, low values imply in more diverse collaborations and, in general, between different genres. For example, country singers usually collaborate among themselves, such as Bryan White and Shania Twain; the same is true for other genres, such as Pop and many collaborations of Lady Gaga.

4. Influence (betweenness and eigenvector). As the name already says, this category incorporates the concept of network influence. Betweenness and Eigenvector metrics accurately explore this concept, quantifying the importance of a node in a social network. A node with high values of these metrics has the potential to be highly influential, having access to different network regions not reachable by other vertices. For that reason, the category is composed of influential and non-influential artists.

Overall, Figure 3.4 illustrates the semantic characterization, by showing how individual metrics are combined into each category (exception of Similarity) and the range of their magnitude.

3.3 Results and Evaluation

Our goal is to identify collaboration profiles and assess their impact on the success of musical artists. To do so, we need to detect groups of artists and their respective collaborative patterns. Hence, in Section 3.3.1, we conduct a cluster analysis based on the four categories described in Section 3.2.4. Finally, we perform the statistical correlation analysis to evaluate the correlation between collaboration profiles and the artist's success measure in Section 3.3.2.

3.3.1 Cluster Analysis

Cluster analysis or clustering is an unsupervised Machine Learning technique that involves grouping data points into specific clusters based on similar properties and/or features. These clusters may reveal patterns related to the phenomenon under study. In our context, we are studying musical collaboration from four categories: *Interaction*,

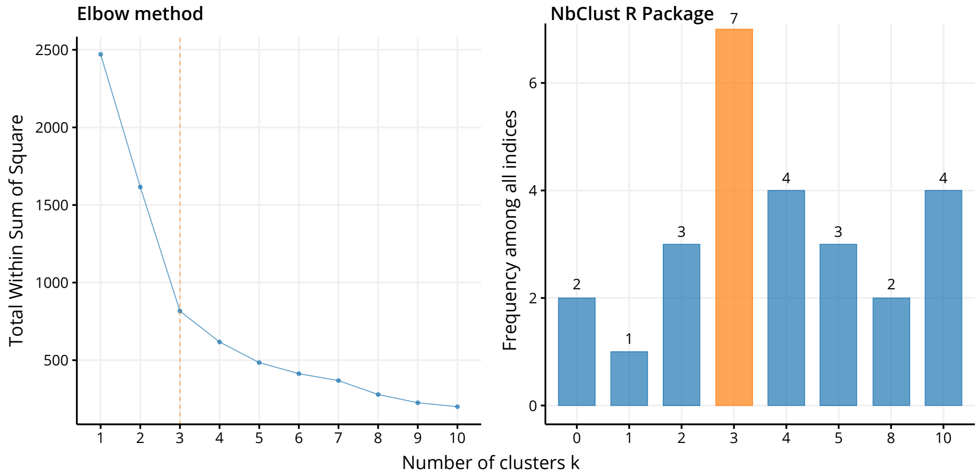


Figure 3.5: The optimal number of clusters (k = 3), according to Elbow Method (left) and the *NbClust* R Package (right).

Distance, *Similarity* and *Influence* (described in Section 3.2.4). To do so, we can use a clustering algorithm to classify each data point (i.e., artist) into a specific collaboration profile, according to such categories. Specifically, we calculate the topological metrics of each artist present in the musical collaboration network and then apply the clustering algorithm to group those with similar topological characteristics.

There are different clustering algorithms that can be applied to finding subgroups of observations within a dataset. Here, we use K-means, which is the simplest and most commonly used cluster method for dividing a dataset into a set of k groups. The first steps of this algorithm is to define the number of clusters to work with. We do so by using a common solution to identify the optimum number of clusters: the Elbow method. It runs K-means on the dataset for a range of values of k (e.g., k from 1 to 10); for each value of k, it calculates the sum of squared errors (SSE); it then considers a line chart of the SSE for each value of k; finally, if the line looks like an arm, then the "elbow" on the arm is the best k value. The results are shown in Figure 3.5. In the left panel, there is a clear bend (or "elbow") at $k = 3$. This bend indicates that additional clusters beyond the third one would negatively affect the results by increasing k.

The Elbow method is often ambiguous and not very reliable, especially if the data is not very clustered. Hence, we also consider a complementary approach to verify the previous results. Here, we use the *NbClust* [10] for computing about 30 methods at once, to find the optimal number of clusters. It provides 30 indices (e.g., *Gamma* [5], *Silhouette* [58], *Gap* [71], and so on) that determine the relevant number of clusters in a dataset and offers the best clustering scheme from different results. Moreover, it provides a function to perform K-means and hierarchical clustering with different distance metrics and aggregation methods. It can simultaneously compute all the indices and determine the number of clusters in a single function call.

Figure 3.5 (right panel) shows the *NbClust* method produced a histogram of up to

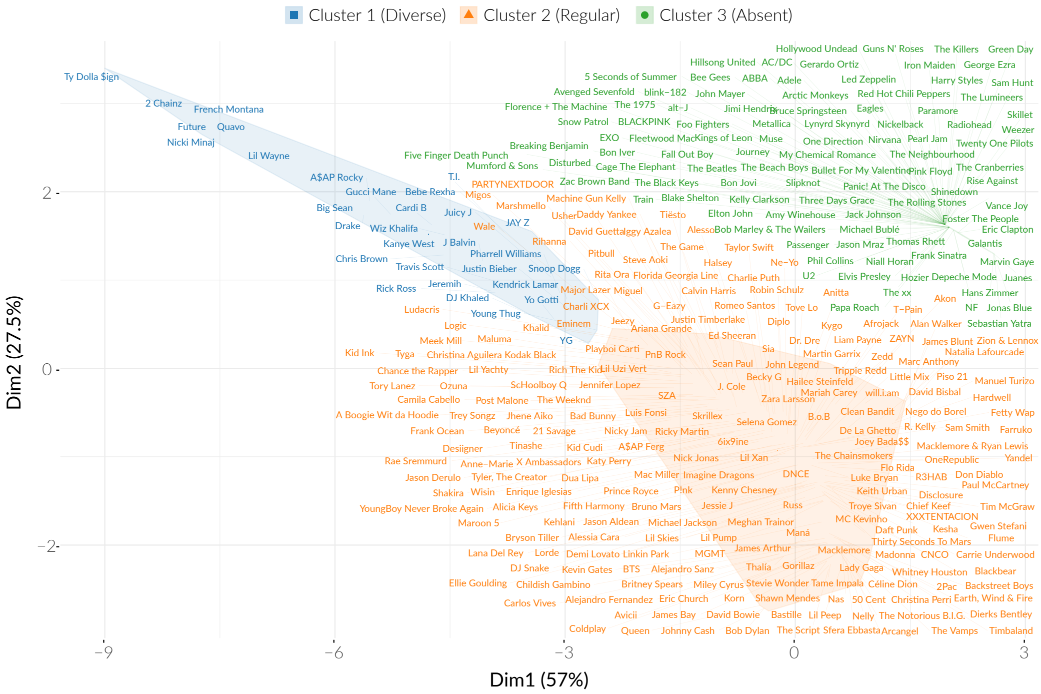


Figure 3.6: Clustering of collaboration profiles. The results are generated with K-means algorithm with number of clusters $k = 3$. The algorithm is based on the topological metrics of each artist.

10 possible number of clusters (cluster configurations never selected are omitted from the X-axis as their frequency would be zero). Atop each vertical bar is the total number of the 30 indices used to estimate the optimal number of clusters. Note that seven of the 30 indices proposed 3 as the best number of clusters. Therefore, we can safely affirm that $k = 3$ is the optimal number of clusters for our dataset.

We now apply K-means again on the dataset with $k = 3$ and illustrate the clustering results by using different colors according to each cluster assignment in Figure 3.6. As the dataset is multi-dimensional, we perform Principal Component Analysis (PCA) to plot data points according to the first two principal components coordinates. According to Figure 3.6, there is a natural division of three clusters: *Cluster 1* in the upper left corner; *Cluster 2* in the opposite lower corner to the right; and *Cluster 3* in the upper right corner. Note K-means clusters data into k groups based on their similarity. Hence, each group is represented by its center that corresponds to the mean of points assigned to the group.

To identify the collaboration profiles of each cluster, we examine the behavior of each pattern according to the four main categories defined in Section 3.2.4. We can infer 16 different standard collaboration profiles from the four categories and their two levels of magnitude, and Table 3.2 presents their characteristics (represented by a collaboration category). In summary, to define the standard profiles, we use the characteristics to represent threshold levels: 1 for a high metric value, that is, greater than or equal to 0.5;

Table 3.2: Standard Collaboration profiles inferred from the four categories of collaboration and their two levels of magnitude. The last column shows the corresponding profile for each cluster identified in the dataset (colored according to the radar plots in Figure 3.7).

Profiles	**Interaction**		**Distance**		**Genre**	**Influence**		**Discovered Cluster**
	Degree	Weighted Degree	Eccentricity	Closeness	Clustering	Betweenness	Eigencentrality	
1A 2A 3A 4A	0	0	0	0	0	0	0	Absent
1A 2A 3A 4B	0	0	0	0	0	1	1	-
1A 2A 3B 4A	0	0	0	0	1	0	0	-
1A 2A 3B 4B	0	0	0	0	1	1	1	-
1A 2B 3A 4A	0	0	1	1	0	0	0	-
1A 2B 3A 4B	0	0	1	1	0	1	1	-
1A 2B 3B 4A	0	0	1	1	1	0	0	Regular
1A 2B 3B 4B	0	0	1	1	1	1	1	-
1B 2A 3A 4A	1	1	0	0	0	0	0	-
1B 2A 3A 4B	1	1	0	0	0	1	1	-
1B 2A 3B 4A	1	1	0	0	1	0	0	-
1B 2A 3B 4B	1	1	0	0	1	1	1	-
1B 2B 3A 4A	1	1	1	1	0	0	0	-
1B 2B 3A 4B	1	1	1	1	0	1	1	Diverse
1B 2B 3B 4A	1	1	1	1	1	0	0	-
1B 2B 3B 4B	1	1	1	1	1	1	1	-

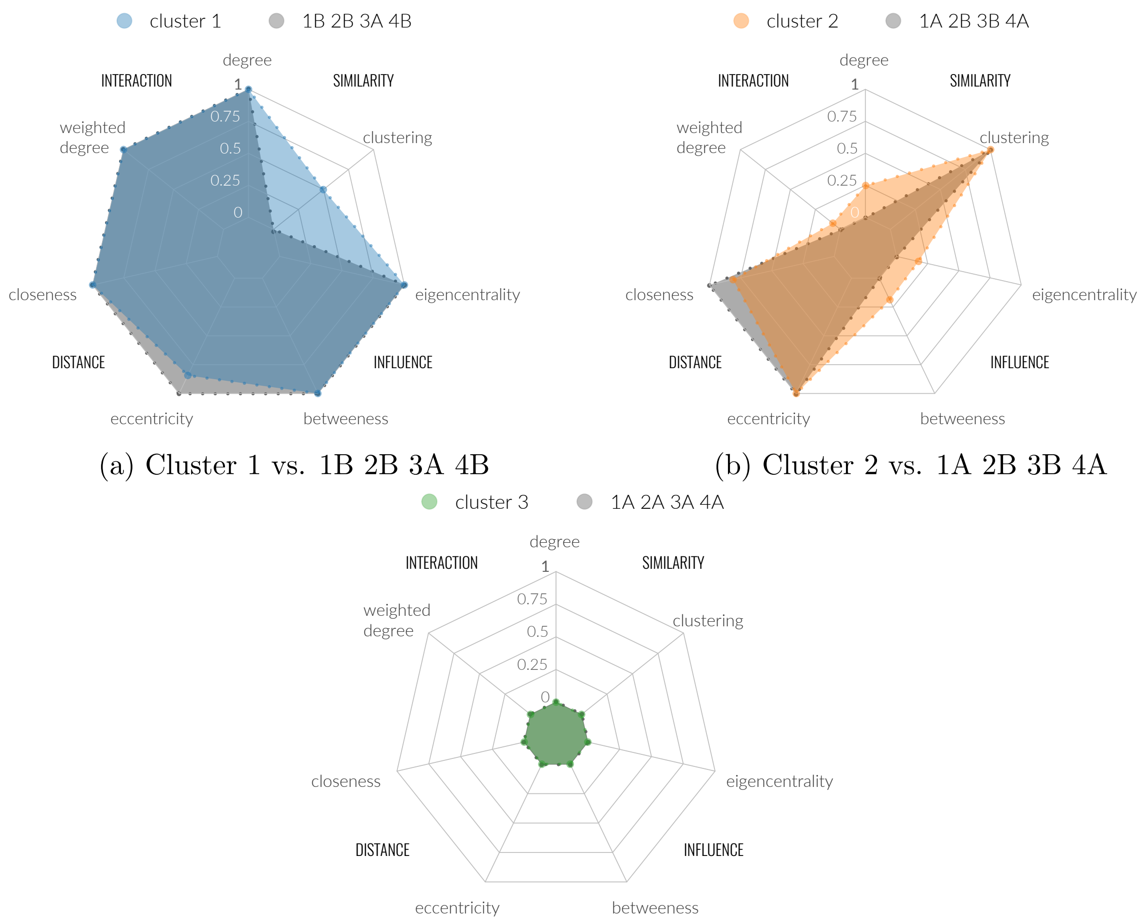
or 0 for low metric values, that is, less than 0.5. Then, we plot radar charts for each profile, as well as for each cluster detected[5]. In the latter, we first calculate the average of each topological metrics and normalize the values into a $[0, 1]$ range. Radar charts are extremely suitable for showing outliers and similarities; then, we can clearly identify which profile each cluster belongs to based on the generated plots (Appendix A.1).

Figure 3.7 shows the result of comparing collaboration profiles with each cluster. Note that *Cluster 1* presents high metric values related to the *Interaction*, *Distance* and *Influence* categories. However, it has an intermediate value for *Similarity*. On the other hand, *Cluster 2* only presents high values for *Distance* and *Similarity*. Finally, *Cluster 3* presents only minimum values for all the categories. To summarize, we name each identified cluster as follows: *Cluster 1* as Diverse Collaboration (Diverse); *Cluster 2* as Regular Collaboration (Regular); and *Cluster 3* as No Collaboration (Absent). According to the radar plots, the collaborative characteristics of *Diverse* are more similar to profile **1B 2B 3A 4B**; the *Regular* cluster is more similar to the **1A 2B 3B 4A** pattern; and the *Absent* cluster is identical to profile **1A 2A 3A 4A**, as identified by their color in Table 3.2.

3.3.2 Statistical Analysis

After identifying the collaboration profiles present in the network, we verify if there is a correlation between an artist's success and his/her collaboration profile by considering

[5]Visualizations available on the project page at bit.ly/comusic_visualizations



(a) Cluster 1 vs. 1B 2B 3A 4B

(b) Cluster 2 vs. 1A 2B 3B 4A

(c) Cluster 3 vs. 1A 2A 3A 4A

Figure 3.7: Comparing the collaboration profiles of each cluster. Each topological metric is represented as an axis starting at the center. Each metric value is plotted along its individual axis, where all variables are connected to form a polygon. Here, the polygon represents exactly the characterization of a collaboration profile.

a set of analyses from different perspectives. First, we perform the normality test of Shapiro-Wilk [63] over the the whole dataset to verify if they follow a normal distribution. The Shapiro-Wilk method provides a statistic test to assess whether a sample has a normal distribution. It is based on the correlation between the data and the corresponding normal scores. Then, we apply three different correlation metrics (Section 3.1.2) to inform the intensity, direction, and significance of the association. Finally, we compute the t-test to determine the statistical significance level. T-test (or Student's T-test) is any hypothesis test that uses statistical concepts to reject or not a null hypothesis when the statistic test follows a Student's t-distribution.

Initially, we present the results of the normality test, because such a method helps to confirm the conclusions on the subsequent analyses. According to the results in Table 3.3, all p-values are much smaller than the significance level of $\alpha = 0.05$, informing that the data distribution is significantly different from a normal distribution. In other words, we cannot assume normality for any of the data samples.

Table 3.3: Shapiro-Wilk test results. Null hypothesis H_0: the distributions are Normal

Sample	**Statistic**	**p-value**	**Normal**
Popularity	0.93788	$5.269e^{-11}$	no
Followers	0.6737	$< 2.2e^{-16}$	no
Degree	0.78467	$< 2.2e^{-16}$	no
Weighted Degree	0.76878	$< 2.2e^{-16}$	no
Eccentricity	0.74696	$< 2.2e^{-16}$	no
Closeness	0.78364	$< 2.2e^{-16}$	no
Clustering	0.87115	$< 2.2e^{-16}$	no
Betweenness	0.58553	$< 2.2e^{-16}$	no
Eigencentrality	0.66979	$< 2.2e^{-16}$	no

Table 3.4: Statistical correlation between: Popularity on Spotify vs. Topological Metrics, and Number of Spotify Followers vs. Topological Metrics, where $\star\star\star$ = Strong, $\star\star$ = Moderate, $\star$ = Weak, no star = Very Weak correlations

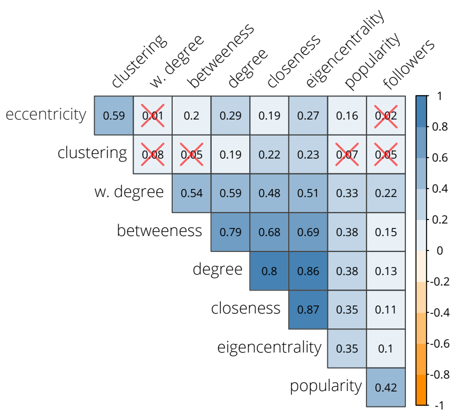
Popularity						
Metric	Pearson		Spearman		Kendall	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Degree	0.5146218 $\star\star\star$	$< 2.2e^{-16}$	0.5132529 $\star\star\star$	$< 2.2e^{-16}$	0.3813272 $\star$	$< 2.2e^{-16}$
Weighted Degree	0.4774128 $\star\star$	$< 2.2e^{-16}$	0.4562222 $\star\star$	$< 2.2e^{-16}$	0.3331673 $\star$	$< 2.2e^{-16}$
Eccentricity	0.3134519 $\star$	$1.647e^{-09}$	0.2151929	$4.457e^{-05}$	0.1623648	$7.095e^{-05}$
Closeness	0.3315823 $\star$	$1.569e^{-10}$	0.4756752 $\star\star$	$< 2.2e^{-16}$	0.3450073 $\star$	$< 2.2e^{-16}$
Clustering	0.04256444	0.4247	0.1071778	0.04388	0.06908103	0.07314
Betweenness	0.4412301 $\star\star$	$< 2.2e^{-16}$	0.5091025 $\star\star\star$	$< 2.2e^{-16}$	0.3800098 $\star$	$< 2.2e^{-16}$
Eigencentrality	0.4287696 $\star\star$	$< 2.2e^{-16}$	0.4875358 $\star\star$	$< 2.2e^{-16}$	0.3526067 $\star$	$< 2.2e^{-16}$
Followers						
Metric	Pearson		Spearman		Kendall	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
Degree	0.1997008	0.0001554	0.1849627	0.0004688	0.1307009	0.0004795
Weighted Degree	0.2496251	$1.979e^{-06}$	0.3029554 $\star$	$5.985e^{-09}$	0.2175485	$4.906e^{-09}$
Eccentricity	0.1004326	0.05906	0.02492095	0.6403	0.01792927	0.6521
Closeness	0.1414757	0.007679	0.160929	0.00239	0.1100344	0.002724
Clustering	0.04100568	0.4418	0.07265466	0.1726	0.05100541	0.174
Betweenness	0.1932948	0.0002537	0.2147099	$4.64e^{-05}$	0.1531739	$5.268e^{-05}$
Eigencentrality	0.1589018	0.002716	0.1501669	0.004634	0.1038141	0.004559

Table 3.5: Rule of Thumb for interpreting the size of a Correlation Coefficient

Correlation Coefficient	**Strength Description**
$\pm 0.81 - \pm 1.00$	Strongest
$\pm 0.61 - \pm 0.80$	Strong
$\pm 0.41 - \pm 0.60$	Moderate
$\pm 0.21 - \pm 0.40$	Weak
$\pm 0.00 - \pm 0.20$	Weak to No Relationship

Note: Hair, J. F. Jr., Babin, B., Money, A., and Samouel, P. (2003). Essentials of Business Research Methods. New York: John Wiley & Sons.

The statistical correlation evaluation of all three studied coefficients is summarized in Table 3.4. Such metrics consider the null hypothesis that there is a correlation between two variables. Considering only popularity, the results show only the p-value of the Clustering metric is above $\alpha = 0.05$. Therefore, all three correlation coefficients are statistically significant for all social metrics, except for the clustering coefficient. On the other hand, for the number of followers, both the clustering coefficient and the eccentricity reach a p-value bigger than the significance level ($\alpha = 0.05$), failing to reject the null hypothesis. That is, we can conclude that there is not a significant linear correlation between number of followers and those metrics.



(a) Kendall's Correlation Matrix

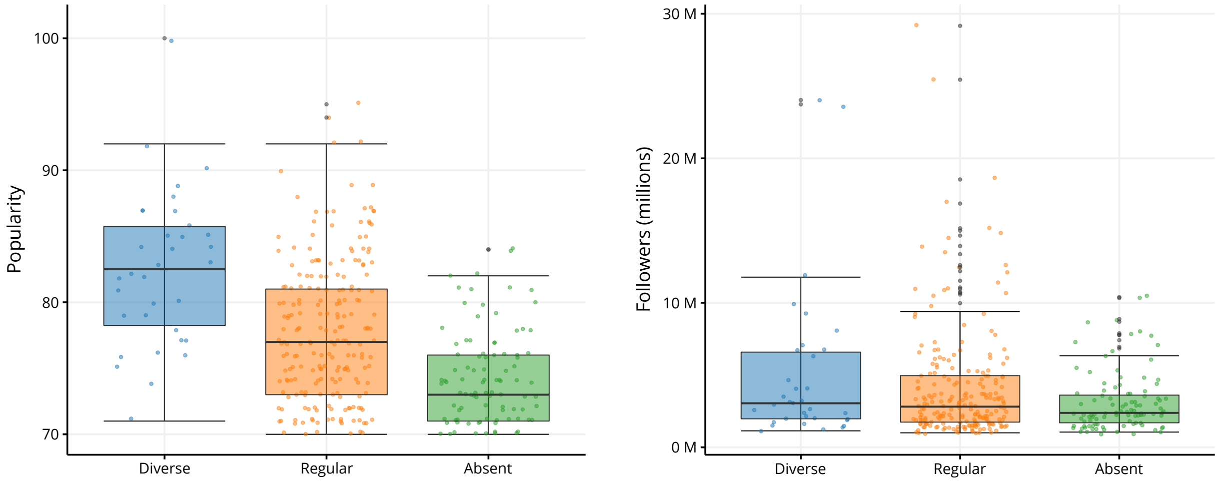
(b) Spearman's Correlation Matrix

Figure 3.8: Visual representation of Kendall and Spearman correlation matrices. Colored boxes represent a significant correlation coefficient between the variables ($p < 0.05$). Light and dark blue represent respectively positive and negative correlation. Correlations with $p > 0.01$ are considered as insignificant, then represented by a red cross. Full scatter plots of the variables are given in the Appendix A.2.

As our data is not normally distributed, non-parametric Kendall and Spearman correlation tests should be applied. Figure 3.8 shows the correlation matrices for both tests. Analyzing only the correlation between popularity and the network metrics, we notice there is a direct association between all the data points. In addition, the tests achieve a moderate and weak (between ± 0.60 and ± 0.21) relationship between popularity and most metrics with statistical significance (Table 3.5). In fact, only the eccentricity measure exhibited a weaker (between ± 0.21 and ± 0.40) association in all three methods. Regarding the number of followers, all three correlation coefficients are very low. However, there is a strong correlation between popularity and the number of followers, as expected.

Figure 3.9 shows boxplots for each cluster related to the success metrics (popularity and number of followers). In each plot, the central rectangle spans the first quartile to the third quartile; a segment within the rectangle shows the median; and upper and lower stems (whiskers) show the minimum and maximum locations. In summary, the chart shows the localization of 50% of the most likely values, the median and the extreme values. Figures 3.9a and 3.9b show the boxplots of the popularity's measure and number of followers of each cluster, respectively.

Figure 3.9a shows that all three clusters have different average levels of popularity, decreasing from *Diverse* to *Absent*. In the first cluster, the mean and median are equivalent to approximately 85. The second one shows the mean and median values less than 80. The last cluster has the lowest rates, close to the minimum values of each cluster (around 70). In fact, 50% of the artists in this cluster are less popular than half of the artists in *Diverse*. We also notice that all three clusters present outliers in their boxplots. Nonetheless, the atypical values continue to follow a descending pattern from the first to the third cluster. In addition, in *Absent*, the cluster has less dispersed data, with a



(a) Popularity per cluster

(b) Followers per cluster

Figure 3.9: Boxplots comparisons of each cluster in relation to the success metrics.

uniform popularity distribution.

On the other hand, the clusters in Figure 3.9b have equivalent median values regarding the number of followers. Nevertheless, with respect to the mean, the rates are also decreasing from *Diverse* to *Absent*. *Diverse* has median values close to 5,000,000 of followers and a maximum value greater than 10,000,000. *Regular* has an average of less than 5,000,000 and a maximum value around to 10,000,000 of followers. *Absent* is the one that presents the smallest mean and maximum value. In fact, 75% of the artists in this cluster have a number of followers less than 2,500,000. As the popularity boxplot, the three clusters present outliers. In the opposite way, the outlying values escape the descending pattern previously observed. Particularly, the most atypical data value belongs to the second cluster (29,165,243). In addition, the third cluster also presents the lowest variability and standard deviation, meaning a high predictability.

3.3.3 Discussion on Empirical Results

The results presented in this chapter show the collaboration profiles detected in each cluster of the musical network. Since the average degree of the network is about 10 and the maximum degree value on the Regular cluster is equal to 30, we define that highly collaborative profiles are those with a Degree value greater than or equal to 30; profiles with moderate collaboration have between 10 and 30 musical collaborations; and non-collaborative profiles with less than 10 collaborations. According to the data, the *Diverse* group is characterized by the profile **1B 2B 3A 4B**: the artists of this cluster have a profile with high interaction (collaborative), a high degree of distant collaborations, a medium level of connections between different genres (diversity), and musical collaborations between influential artists (influence). The *Regular* cluster presents the profile **1A 2B 3B 4A**: such artists are not very collaborative (low interaction), with a considerable degree of distant collaboration and between the same musical genres (medium diversity), and rare musical collaborations between influential artists (low influence). Finally, the artists in *Absent* are characterized by the collaboration profile **1A 2A 3A 4A**; which means this cluster contains non-collaborative artists. In fact, the majority of this cluster consists of bands, which usually do not collaborate by definition.

With the collaboration profiles of each cluster defined, we can discuss their impacts on the artists' success. According to the results in Section 3.3.2, there is a moderate to strong relationship between the metrics of interaction, closeness and influence, and Spotify's measure of popularity. These results indicate such metrics may positively impact an artist's popularity. Therefore, successful artists are very likely to present highly collaborative profiles with distant and influential collaborations. We did not find correlation between popularity and the clustering coefficient, thus we cannot infer that the diversity of musical genres impacts on musical success. According to Figure 3.2, there is a weak but statistically significant relationship between the interaction and influence metrics and the number of followers. However, for the proximity and clustering metrics, no significant relationship was detected. In this case, there is insufficient statistical evidence to support the claim that the number of an artist's followers impacts (positively or negatively) their success. Then, Figure 3.2 also shows the number of followers is strongly correlated with the popularity of the artist (which is expected to some degree).

Finally, with Figure 3.8, we validate some outcomes. For instance, *Diverse* presents the profile that more likely infers musical success, and this is the exact cluster composed of the most successful artists. Likewise, *Absent* has the less collaborative profile, which does not necessarily expand musical success. As reported in Figure 3.8, this is the cluster that presents the lowest values of success metrics in comparison to the other clusters. Therefore, we may conclude that successful artists are more likely to have a high degree of collaboration between influential and diversified artists. Similarly, for most cases, those who prefer to pursue a non-collaborative musical career may be missing an opportunity to enhance and expand their potential. Also, these success characteristics affect more significantly than the musical genre factor. While the first observations may seem intuitively obvious, the latter conclusion is less evident. This is because, typically, highly connected entities tend to achieve some metrics of popularity (regardless of how popularity is defined and what it implies). However, it is counter-intuitive to believe that musical genres are not as influential in an artist's success.

3.4 Overall Considerations

In this chapter, we identified collaboration profiles present in a musical network. Moreover, we analyzed the relationship between such collaborative patterns and the artists' success. Specifically, using data from Billboard and Spotify, we modeled a social network and defined a measure of success. Then, to identify the collaborative profiles, we applied six topological metrics and defined four main categories: *Interaction, Distance, Similarity* and *Influence*. Next, by using the K-Means algorithm, we identified three clusters with distinct collaboration profiles. To measure the relationship between these profiles and musical success, we conducted a statistical analysis.

Through an extensive data analysis, we identified moderate and statistically significant correlations between popularity and metrics for degree and centrality. However, we found no substantial relationship in terms of the number of followers and social metrics. Our results provide strong evidence that clusters with a high degree of interaction, influence, and diversity, are more likely to present successful artists. Moreover, these success features affect more significantly than the musical genre factor.

Limitations. Despite the relevant results, we cannot forget that correlation does not imply causation. If there is a direct mathematical correlation between variables, it does not indicate a relationship. Other factors can represent the cause and not all success factors were covered here. Examples of factors include the number of awards received by artists, the number of views of music clips, the success of the songs, among others. Such external factors can improve the correlation analysis between musical success and collaborative patterns. Therefore, in the next chapter, we further investigate the causality relationship using the Granger causality test.

Chapter 4

Causality Analysis on Collaboration Profiles and Music Success

The broad range of characteristics that lead to the success of a song exceeds its intrinsic content, namely the audio features and the lyrics. Factors such as the artist's preferred attachment [7], society and culture [13] or psychological parameters on the reasons for preferring a track and listening exposure to tracks [51], can also fulfill a key role. Following an alternate direction, we argued that the way artists connect professionally can significantly affect musical success. Using data from Billboard and Spotify, we identified (Chapter 3) three collaboration profiles in a musical network composed of 354 successful artists: Diverse, Regular and Absent. Through a statistical analysis, we observed that the most successful artists are more likely to have highly collaborative profiles among influential and diverse artists (belonging to the Diverse profile). On the other hand, those who prefer to pursue a non-collaborative music career (belonging to the Absent profile) may be missing an opportunity to improve and expand their potential.

Though the correlation between collaboration profiles and musical success is well established, there is no research into the *causality* in such relationship. Understanding both statistical terms (together) is very important, not only for drawing insights but more importantly, for a correct conclusion at the end. However, determining causation is never perfect in the real world, and there are different methods to find evidence on causal relationships, including the Granger causality[1] test. Overall, we explore the potential of Granger causality [19] to assess the existence of a causal relation between collaboration profiles and musical success.

The remainder of this chapter is organized as follows. In Section 4.1, we describe the dataset and introduce the proposed methodology. Next, we detail the results and experimental evaluation in Section 4.2. Finally, we conclude with overall insights in Section 4.3.

[1] *Granger's causality* is a statistical concept of cause-and-effect relationship that is based on prediction. Note: it is different than the term *causality* in other contexts such as Pearl Causal Model (PCM) [53]

Table 4.1: Statistics of the top 30 artists selected (Col = Collaborations)

Diverse			**Regular**			**Absent**		
Artist	Solo	Col	Artist	Solo	Col	Artist	Solo	Col
Carlos Vives	188	71	XXXTENTACION	61	32	5 Seconds of Summer	186	4
Drake	142	97	Ariana Grande	88	77	Adele	52	2
Future	202	134	Bad Bunny	22	57	Arctic Monkeys	131	0
J Balvin	91	141	Ed Sheeran	170	57	One Direction	125	17
Kanye West	124	89	Eminem	333	150	Panic! At The Disco	172	8
Kendrick Lamar	111	70	Khalid	42	38	Queen	893	72
Nicki Minaj	188	170	Marshmello	29	75	Red Hot Chili Peppers	438	0
Travis Scott	45	32	Nicky Jam	102	112	Sebastian Yatra	38	69
Ty Dolla $ign	44	167	Ozuna	33	71	The Beatles	825	14
Wiz Khalifa	204	130	Post Malone	47	31	Twenty One Pilots	67	5

4.1 Methodology

Our methodology aims at investigating if there is a causal relationship between collaboration profiles and musical success. In summary, our methodology is: build a dataset containing 30 popular artists from three different collaboration profiles (Section 4.1.1); from this data, use ego networks to model such artists' collaborations (Section 4.1.2), and define a temporal success measure (Section 4.1.3); create three time series of musical success, musical collaborations and solo songs of the selected artists (Section 4.1.4); and apply the Granger causality test to check if patterns of collaboration can influence the musical success or vice versa (Section 4.1.5).

4.1.1 Data Collection

In the previous chapter, we detected three collaboration profiles: Diverse, Regular and Absent. From such profiles, we now select the 10 most popular artists. That is, in total, 30 successful artists are chosen. Using the Spotipy library, we collected the entire discography of such artists. The *Spotipy* provides the Spotify catalog information about a certain artist's work, where four types of albums are available: *album*, *single*, *appears_on* and *compilation*. Here, we did not filter for a specific type, then considering all four. Table 4.1 lists the names of all the artists present in our dataset[2], as well as the total number of solo and collaborative songs of each one. Note that in the Diverse profile, the number of solo songs and collaborations are high and similar, in general; in the Regular profile, the number of solo songs and collaborations is slightly lower, but also well balanced. However, for the Absent profile, the number of solo songs is extremely

[2]Full dataset openly available at `https://homepages.dcc.ufmg.br/~mirella/projs/bade`

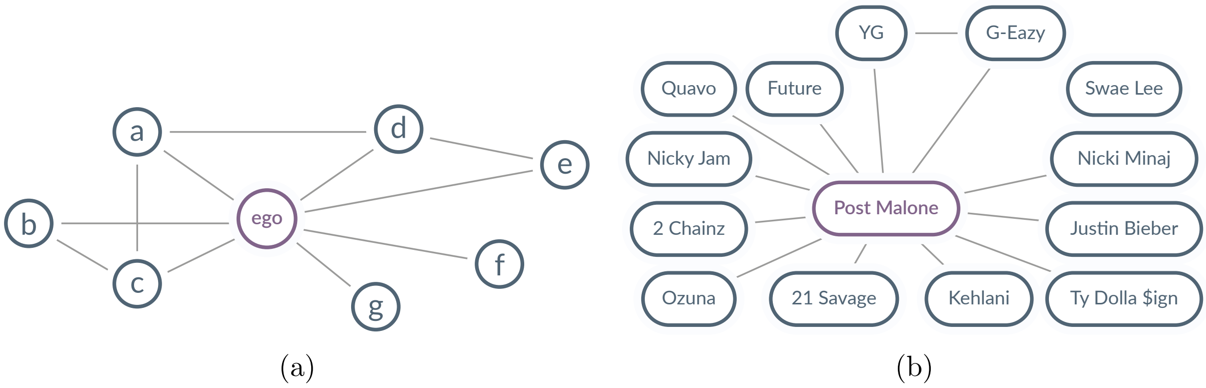


Figure 4.1: (a) A generic example of the ego network model: a top 30 artist, (*ego*), and his/her collaborating artists, ($alter_{a-g}$). (b) A real ego network example for the rapper *Post Malone*.

higher than the number of collaborations, which is very limited in relation to the other profiles.

4.1.2 Ego Network Modeling

From the dataset, we create ego networks for each of the 30 selected artists. Ego networks have proved to be a valuable tool for understanding the relationships that individuals establish with their peers. In an ego network, the individual (termed *ego*) is at the center of the graph, and the edges connect her/his to the peers (termed *alters*) with whom she/he interacts [70]. As depicted in Figure 4.1, we model 30 ego networks, where the top 30 artists represent the ego nodes and their musical collaborations, the alter nodes.

4.1.3 Temporal Success Measure

Musical success can be defined through popularity on social media and music platforms, sales profit, awards, etc. Another common approach is to rely on pop charts, such as the Billboard charts[3]. In this case, the ranking of a song/album/artist in a chart is a time domain signal, describing its popularity over time. To define the temporal success measure, we first collected and grouped rankings of 18 top Billboard charts by the name

[3] *Billboard charts*: https://www.billboard.com/charts

of the top 30 artists. Next, we calculate the popularity over time of these artists from the *rank_score* of the artists/albums/songs top charts. The *rank_score* is the inverted rating on a chart. That is, the $rank\_score_{jt}(i)$ of a chart j at time t of a song/album/artist i is obtained by Equation 1.

$$rank\_score_{jt}(i) = max\_rank_{jt} - rank_{jt}(i) + 1, \tag{4.1}$$

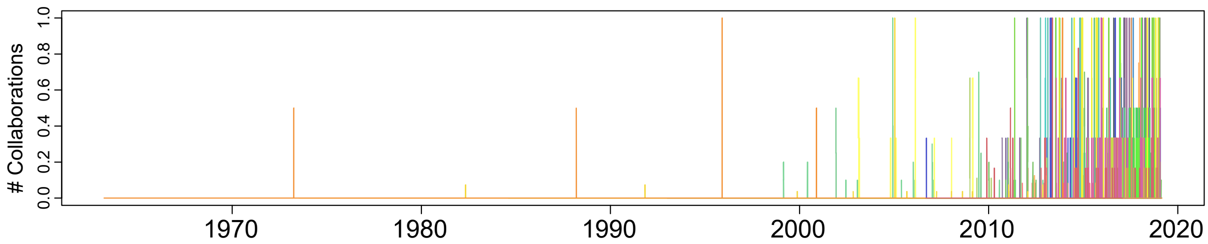
where max_rank_{jt} is the lowest rank of the j chart at time t, and $rank_{jt}(i)$ is the rank of the song/album/artist i. For instance, in a weekly Top Album Sales chart, the album ranked highest has a *rank_score* of 100 and the album ranked lowest has a *rank_score* of 1. After calculating the *rank_scores* for each of the 18 charts for each of the artists over time, we aggregate these scores to obtain our temporal success measure, S_t. Then, we define the temporal success measure of an artist i with Equation 4.2.

$$S_t(i) = \sum_{j=1}^{18} rank\_score_{jt}(i). \tag{4.2}$$

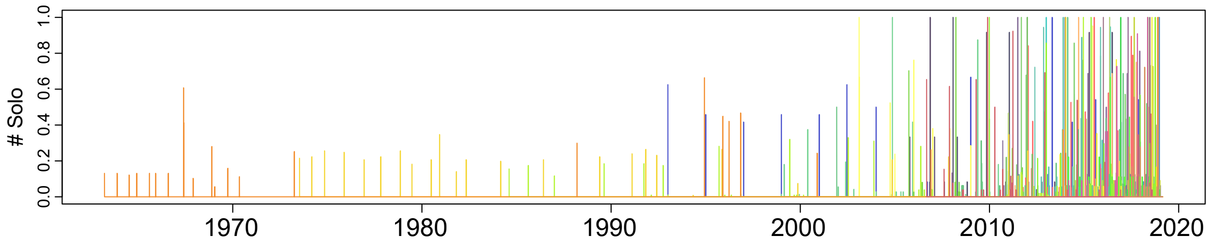
4.1.4 Time Series

Time series analysis is a powerful technique that helps to understand the distinct temporal data patterns and to predict how levels of a variable shall change in the future, by considering what has happened in the past. To study the dynamic properties of the ego networks modeled in Section 4.1.2, we conduct two parallel analyses based on daily time series of (i) collaborative songs and musical success, and (ii) solo songs and musical success. Hence, in the end, each of the 30 ego networks produces three time series defined as follows.

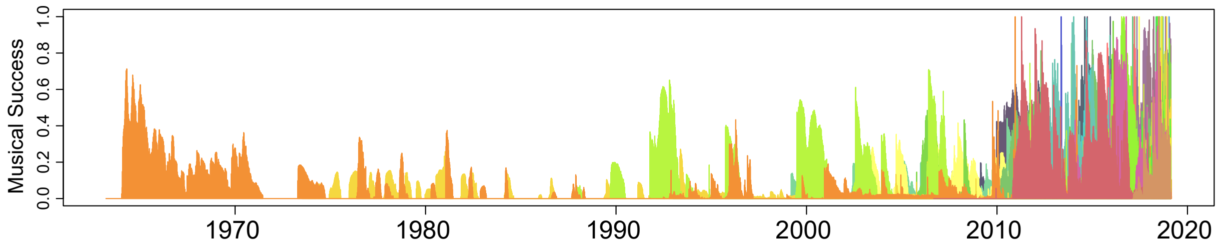
- **Collaborative songs.** Daily time series with the total number of musical collaborations normalized over time. The elapsed time of each of these series was collected from the release dates of all the artist's songs. See Figure 4.2a.
- **Solo songs.** Daily time series with the total number of solo songs normalized over time. The elapsed time of each of these series was collected from the release dates of all the artist's songs. See Figure 4.2b.
- **Musical success.** Daily time series with the artist's success measure defined in Section 4.1.3 normalized over time. See Figure 4.2c.



(a) Collaborative songs time series



(b) Solo songs time series



(c) Musical success time series

Figure 4.2: The three time series created for each of the 30 ego networks. Each color represents a top 30 artist's time series. The time series vary from 1963 to 2019.

4.1.5 Granger Causality Analysis

Finally, we analyze the effect of musical collaborations on musical success over time, applying the Granger causality test. Technically, the Granger causality test is a method to determine if one time series is useful in forecasting another. Since predictability remains a fundamental feature of causal attribution, Granger causality (GC) is also frequently interpreted in cause-effect context. However, GC should not be confused with the deep meaning of the word. GC is *limited* exclusively to identifying a statistical cause-and-effect relationship between two variables when there is a temporal precedence relationship between them.

The basic principle of the Granger causality test is to check if there is a lagged cause relationship between two or more variables. For example, we can investigate the causality between two time series of world prices of Arabica and Robusta coffees [72]. Let X_t and Y_t be the Arabica and Robusta coffees variables, respectively. X_t Granger-causes Y_t, if the forecast obtained in the current price of Robusta coffee can be improved by considering the lagged price information for Arabica coffee. In order to test the Granger causality, the series Y_t and X_t must be stationary, i.e., the series must be integrated with order zero $I(0)$. According to [14], the following VAR (Vector Autoregression) model can be estimated to test the causality between two variables:

$$Y_t = \beta_1 + \beta_2 Y_{t-1} + \beta_3 X_{t-1} + \varepsilon_{1t} \tag{4.3}$$

$$X_t = \beta_4 + \beta_5 X_{t-1} + \beta_6 Y_{t-1} + \varepsilon_{2t}, \tag{4.4}$$

where ε_{1t} and ε_{2t} are white noises, (β_2, β_5) are the current and lagged coefficients of endogenous variables (determined within the model) and (β_3, β_6) are the current and lagged coefficients of exogenous variables (determined outside the model). These coefficients measure the influence of past values of each variable. In this system of equations, each variable is a function of its value with a lag and the value of the other variable with a lag. After verifying if all the variables are stationary, we can use the VAR model with n lags in Equations 4.5 and 4.6.

$$Y_t = \sum_{i=1}^{n} \alpha_i Y_{t-i} + \sum_{j=1}^{n} \beta_j X_{t-j} + \varepsilon_{1t} \tag{4.5}$$

$$X_t = \sum_{k=1}^{n} \gamma_k Y_{t-k} + \sum_{l=1}^{n} \delta_l X_{t-l} + \varepsilon_{2t} \tag{4.6}$$

Equation 4.5 postulates that current values of Y_t are related to past values of Y_t as well as to lagged values of X_t. On the other hand, Equation 4.6 postulates a similar behavior for the variable X_t. In general terms, since the future cannot predict the past, if the variable X_t *Granger-causes* the variable Y_t, then changes in X_t must temporarily precede changes in Y_t. In order to test whether X_t *Granger-causes* Y_t, we attempt to reject or accept the following hypotheses by means of F-tests as follows:

$$H_0 : \sum_{j=1}^{n} \beta_j = 0$$

$$H_1 : \beta_j \neq 0,\ \textit{for at least one } j,$$

where j is the number of lags. Hence, the Null Hypothesis (H_0) in which X_t does not *Granger-cause* Y_t is tested against the Alternative Hypothesis (H_1) in which at least one lag of the variable X_t *Granger-causes* Y_t.

Based on the estimated coefficients for Equations 4.5 and 4.6, four different hypotheses about the relationship between Y_t and X_t can be formulated, as follows.

Unidirectional *Granger-causality* $X_t \rightarrow Y_t$: when the estimated coefficients of the lags of X_t in Equation 4.5 are jointly different from zero ($\sum_{j=1}^{n} \beta_j \neq 0$), and the estimated coefficients of the lags of Y_t in Equation 4.6 are jointly zero ($\sum_{k=1}^{n} \gamma_k = 0$).

Unidirectional *Granger-causality* $Y_t \rightarrow X_t$: when the estimated coefficients of the lags of X_t in Equation 4.5 are jointly zero ($\sum_{j=1}^{n} \beta_j = 0$), and the estimated coefficients of the lags of Y_t in Equation 4.6 are jointly different from zero ($\sum_{k=1}^{n} \gamma_k \neq 0$).

Bidirectional (or feedback) *Granger-causality*: when the sets of lagged coefficients of X_t and Y_t are statistically different from zero in both regressions ($\sum_{j=1}^{n} \beta_j \neq 0$ and $\sum_{k=1}^{n} \gamma_k \neq 0$).

Independence between X_t and Y_t: when, in both regressions, the sets of lagged coefficients of X_t and Y_t are not statistically different from zero ($\sum_{j=1}^{n} \beta_j = 0$ and $\sum_{k=1}^{n} \gamma_k = 0$).

4.2 Results and Evaluation

This section goes over our analyses through the times series evaluation and a discussion over the results. Following [72], the first step in the Granger causality test is to verify if the time series are stationary. Hence, we apply the Augmented Dickey-Fuller (ADF) [17] and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) [35] tests. For both time series, the null hypothesis H_0 of non-stationary can be rejected at the 5% significance level. Tables 4.2 and 4.3 present, respectively, the results for ADF and KPSS to verify the order of integration of the level series for the two analyses. According to the results, all POP series are integrated of order one $I(1)$ at a significance level of 5%. However, the great majority of $COLLAB$ and $SOLO$ series are stationary with $I(0)$ at a significance level of 5%. For all the time series group, we consider the maximum integration order $m = 1$.

In cases where both series have the same integration order, the co-integration test must be applied between the variables. Therefore, we perform the Johansen co-integration test for the 14 pairs of time series that presented the same integration order (See Tables 4.2 and 4.3). As highlighted by [72], the lag selection procedure is a crucial step for the Granger causality test. In this methodology, in order to determine the number of lags, p, we first estimate an unrestricted VAR(p) model and select the model that minimized the information criteria of Akaike (AIC), [22] (HQ) and [61] (SC).

Based on Tables 4.2 and 4.3 and following [72], we add to the VAR(p) models, additional m lags in each of the variables of each of the equations – where p is the appropriate maximum lag for the variables, and m represents the maximum integration order for the time series. Next, we test the Granger causality as follows. For expository purposes, suppose VAR(p) has two equations, for X_t and for Y_t. We test the hypothesis that (only) the coefficients of the first lagged values of X_t are zero in equation Y_t, using the F-test. Subsequently, we perform the same analysis for the coefficients of the lagged values of Y_t in equation X_t.

4.2.1 Hyphoteses

As mentioned earlier, we conduct two parallel analyses based on daily time series of collaborative songs and musical success, and solo songs and musical success, from each of the top 30 artists selected from the three collaboration profiles (*Diverse*, *Regular* and *Absent*). In both analyses, we test the null hypothesis H_0 that collaborative profiles do not *Granger-cause* artistic success. From the results of both analyses, we can distinguish four hypotheses as follows.

Hypothesis 1. The way in which artists professionally connect with each other is useful in forecasting their musical success. This hypothesis imagines a situation in which an artist's collaborative patterns have a direct impact on his/her success.

Hypothesis 2. The musical success of an artist is useful in forecasting his/her collaborative profile. This hypothesis imagines a situation in which the musical success of artists drives the way in which they collaborate.

Hypothesis 3. The musical success of an artist assists in forecasting their collaborative profile, as well as their profile helps in forecasting their success. This hypothesis imagines a situation in which artists, before collaborating with other people, consider the success level of such individuals. In the same way, the growth of artistic success is affected by the way artists collaborate professionally.

Hypothesis 4. There is no causality relationship (in the Granger sense) between collaboration profiles and musical success.

4.2.2 Discussion

For clarification purposes, we start our discussion by summarizing the three base profiles: *Diverse* has highly collaborative and important artists; *Regular* has collaborative, regular artists; and *Absent* has non-collaborative artists. Now, we further discuss the

Table 4.2: Granger causality test for the first analysis. Column 1 groups the results by profile. Column 2 shows the selected top 30 artists. Columns 3 and 4 show the data characteristics of the series $COLLAB$ (number of musical collaborations) and POP (musical success) respectively. If the series itself is stationary, we represent it by $I(0)$. If the series is integrated with order n, we represent it by $I(n)$. Column 5 shows the lag orders, where the lag order chosen is underlined. Column 6 uses ✓ to indicate that the two series are co-integrated. Columns 7 and 8 represent the P-value of the Granger causality test: ***, **, * and . statistically significant at 0%, 1%, 5% and 10%, respectively

	Artist	$COLLAB$	POP	Lag	Co-integration	$COLLAB \rightarrow POP$	$POP \rightarrow COLLAB$
Diverse	Drake	$I(0)$	$I(1)$	15 — $\underline{7}$		5.756e−07 ***	0.9831
	J Balvin	$I(1)$	$I(1)$	$\underline{20}$ — 14	✓	4.988e−06 ***	4.988e−06 ***
	Carlos Vives	$I(1)$	$I(1)$	17 — $\underline{15}$	✓	0.2477	< 2.2e−16 ***
	Nicki Minaj	$I(0)$	$I(1)$	$\underline{14}$		0.032 *	0.9881
	Kanye West	$I(0)$	$I(1)$	$\underline{56}$ — 14 — 7		< 2.2e−16 ***	1
	Kendrick Lamar	$I(0)$	$I(1)$	14 — $\underline{10}$ — 9		5.827e−1 ***	2.571e−06 ***
	Future	$I(0)$	$I(1)$	$\underline{61}$ — 7		< 2.2e−16 ***	< 2.2e−16 ***
	Ty Dolla $ign	$I(0)$	$I(1)$	14 — $\underline{7}$		0.491 *	0.5503 **
	Travis Scott	$I(0)$	$I(1)$	21 — 14 — $\underline{7}$		0.3772	0.7933
	Wiz Khalifa	$I(0)$	$I(1)$	20 — $\underline{14}$		6.872e−10 ***	6.872e−10 ***
Regular	XXXTENTACION	$I(0)$	$I(1)$	$\underline{7}$		0.6671	0.7559
	Post Malone	$I(0)$	$I(1)$	14 — $\underline{7}$		0.2291	5.379e−11 ***
	Ozuna	$I(1)$	$I(1)$	20 — $\underline{14}$	✓	0.000236 ***	0.003717 **
	Bad Bunny	$I(1)$	$I(1)$	$\underline{7}$	✓	0.2568	0.994
	Khalid	$I(1)$	$I(1)$	17 — $\underline{14}$	✓	4.359e−07 ***	3.546e−14 ***
	Ed Sheeran	$I(1)$	$I(1)$	19 — 17 — $\underline{14}$	✓	1.371e−11 ***	< 2.2e−16 ***
	Ariana Grande	$I(0)$	$I(1)$	$\underline{7}$		0.6814	0.0009666 ***
	Nicky Jam	$I(1)$	$I(1)$	99 — $\underline{59}$ — 15 — 14	✓	< 2.2e−16 ***	< 2.2e−16 ***
	Eminem	$I(0)$	$I(1)$	$\underline{54}$ — 14 — 7		3.344e−15 ***	1
	Marshmello	$I(1)$	$I(1)$	97 — 21 — 14 — $\underline{7}$	✓	0.112	8.122e−07 ***
Absent	Panic! At The Disco	$I(1)$	$I(1)$	14 — $\underline{7}$	✓	< 2.2e−16 ***	0.1697
	5 Seconds of Summer	$I(1)$	$I(1)$	$\underline{7}$	✓	0.3544	5.002e−15 ***
	Twenty One Pilots	$I(0)$	$I(1)$	$\underline{7}$		0.0001239 ***	0.9952
	The Beatles	$I(0)$	$I(1)$	20 — 17 — $\underline{14}$		1	1.711e−14 ***
	Queen	$I(1)$	$I(1)$	98 — 21 — $\underline{14}$ — 7	✓	0.9869	0.9922
	Red Hot Chili Peppers	$I(0)$	$I(1)$	$\underline{1}$		–	–
	Arctic Monkeys	$I(0)$	$I(1)$	20 — 15 — $\underline{14}$		0.7243	2.728e−12 ***
	Sebastian Yatra	$I(1)$	$I(1)$	16 — 8 — $\underline{7}$	✓	6.852e−11 ***	0.722
	Adele	$I(0)$	$I(1)$	20 — 15 — $\underline{14}$		0.7243	2.728e−12***
	One Direction	$I(0)$	$I(1)$	36 — $\underline{7}$		0.001496 **	2.043e−05 ***

Table 4.3: Granger causality test for the second analysis. Column 1 groups the results by profile. Column 2 shows the selected top 30 artists. Columns 3 and 4 show the data characteristics of the series *SOLO* and *POP* series respectively. If the series itself is stationary, we represent it by $I(0)$. If the series is integrated with order n, we represent it by $I(n)$. Column 5 shows the lag orders, where the lag order chosen is underlined. Column 6 uses ✓ to indicate that the two series are co-integrated. Columns 7 and 7 represent the P-value of the Granger causality test: ***, **, * and . statistically significant at 0%, 1%, 5% and 10%, respectively

	Artist	*SOLO*	*POP*	Lag	Co-integration	$SOLO \to POP$	$POP \to SOLO$
Diverse	Drake	$I(0)$	$I(1)$	$\underline{15}$		0.121	$< 2.2e{-16}$ ***
	J Balvin	$I(0)$	$I(1)$	$\underline{15}$		0.3623	$< 2.2e{-16}$ ***
	Carlos Vives	$I(0)$	$I(1)$	$\underline{21}$		0.9976	$< 2.2e{-16}$ ***
	Nicki Minaj	$I(0)$	$I(1)$	$\underline{19}$ — 7		0.4811	$< 2.2e{-16}$ ***
	Kanye West	$I(0)$	$I(1)$	$\underline{15}$ — 7		0.1103	$1.671e{-1}$ ***
	Kendrick Lamar	$I(0)$	$I(1)$	$\underline{22}$ — 7		0.477	$< 2.2e{-16}$ ***
	Future	$I(0)$	$I(1)$	$\underline{7}$		0.4268	0.9981
	Ty Dolla $ign	$I(0)$	$I(1)$	$\underline{22}$ — 15 — 7		$2.403e{-09}$ ***	$< 2.2e{-16}$ ***
	Travis Scott	$I(0)$	$I(1)$	$\underline{15}$		1	$< 2.2e{-16}$ ***
	Wiz Khalifa	$I(0)$	$I(1)$	$\underline{18}$ — 7		0.8481	$< 2.2e{-16}$ ***
Regular	XXXTENTACION	$I(0)$	$I(1)$	$\underline{15}$ — 7		0.766	$< 2.2e{-16}$ ***
	Post Malone	$I(0)$	$I(1)$	$\underline{15}$		0.9966	$< 2.2e{-16}$ ***
	Ozuna	$I(0)$	$I(1)$	$\underline{15}$		0.7705	$< 2.2e{-16}$ ***
	Bad Bunny	$I(0)$	$I(1)$	98 — 71 — $\underline{65}$ — 20 — 13 — 7		$< 2.2e{-16}$ ***	$< 2.2e{-16}$ ***
	Khalid	$I(1)$	$I(1)$	17 — $\underline{16}$	✓	$2.735e{-05}$ ***	$< 2.2e{-16}$ ***
	Ed Sheeran	$I(0)$	$I(1)$	$\underline{22}$ — 14		0.872	$2.147e{-14}$ ***
	Ariana Grande	$I(0)$	$I(1)$	22 — $\underline{15}$ — 7		0.2358	$< 2.2e{-16}$ ***
	Nicky Jam	$I(1)$	$I(1)$	29 — $\underline{22}$	✓	1	$< 2.2e{-16}$ ***
	Eminem	$I(0)$	$I(1)$	20 — $\underline{19}$ — 7		0.9331	$< 2.2e{-16}$ ***
	Marshmello	$I(0)$	$I(1)$	14 — $\underline{7}$		0.05908 .	0.2674
Absent	Panic! At The Disco	$I(0)$	$I(1)$	$\underline{22}$ — 7		0.9999	$< 2.2e{-16}$ ***
	5 Seconds of Summer	$I(0)$	$I(1)$	$\underline{7}$		0.9876	1
	Twenty One Pilots	$I(0)$	$I(1)$	$\underline{15}$		0.9999	$< 2.2e{-16}$ ***
	The Beatles	$I(0)$	$I(1)$	38 — $\underline{22}$ — 8 — 7		0.6306	$< 2.2e{-16}$ ***
	Queen	$I(0)$	$I(1)$	98 — $\underline{22}$ — 7		0.999	$1.697e{-07}$ ***
	Red Hot Chili Peppers	$I(0)$	$I(1)$	22 — $\underline{21}$ — 14 — 7		1	$< 2.2e{-16}$ ***
	Arctic Monkeys	$I(0)$	$I(1)$	$\underline{52}$ — 35 — 7		0.4089	$4.898e{-08}$ ***
	Sebastian Yatra	$I(0)$	$I(1)$	$\underline{15}$		$3.355e{-05}$ ***	$< 2.2e{-16}$ ***
	Adele	$I(0)$	$I(1)$	$\underline{52}$ — 35 — 7		0.4089	$4.898e{-08}$ ***
	One Direction	$I(0)$	$I(1)$	$\underline{19}$ — 7		$1.57e{-05}$ ***	$< 2.2e{-16}$ ***

results from three perspectives: the collaborative songs, the solo ones, and some relevant general points crossing all information with the three profiles.

Collaborative Songs. Table 4.2 suggests that musical collaboration helps predicting artistic success in 40% of *Diverse* profiles, just as success helps in predicting collaborations. In 20% of cases (*Drake* and *Kayne West*), there is unilateral causality where the success of the artist strongly influences the incidence of their musical collaborations. Only one artist (*Carlos Vives*) presented an opposite result, where his musical associations directly affect his success. The other 30% did not present statistically significant results. Likewise, for 40% of the *Regular* profile, musical collaboration helps predicting artistic success, just as success helps predicting collaborations. Now, in 30% of the cases (*Ariana Grande*, *Marshmello* and *Post Malone*), there is a unilateral causality, where, this time, the artist's musical associations directly affect success. Only one artist (*Eminem*) has produced a divergent result, and his success affects the incidence of his musical collaborations. The other 20% did not present statistically significant results. For the *Absent* profile, the tests indicate no bilateral causality. In fact, in 30% of cases (*Panic! At The Disco*, *Twenty One Pilots* and *Sebastian Yatra*), there is unilateral causality, where success helps predicting collaborations. In 50% of cases, there is a unilateral causality between the artist's musical associations and their success. The other 20% did not present statistically significant results.

Solo Songs. The results summarized in Table 4.3 suggest that in 80% of *Diverse* cases, solo songs assist in predicting artistic success. This time, only one artist (*Ty Dolla $ign*) presented bilateral causality, and the remaining 10% did not present statistically significant results. Then, in 70% of *Regular* cases, solo songs assist in predicting artistic success. Two artists (*Bad Bunny* and *Khalid*) presented a bilateral causality, and one artist reports no statistically significant results. Finally, for the *Absent* profile, the tests also indicate that in 70% of the cases solo songs assist in predicting artistic success. Two artists (*Sebastian Yatra* and *One Direction*) presented a bilateral causality, and one artist indicated no statistically significant results.

General Points. If two or more time series are cointegrated, then there must be *Granger causality* between them —either unidirectional or in both directions; however, the opposite is false. According to Table 4.2, the time series of *Bad Bunny* and *Queen* are cointegrated, but no evidence of causality is found. In both cases, such a conflict may have occurred because the size of the sample is too small to satisfy the asymptotics on which the cointegration and causality tests depend. Similarly, the sample size of the *Red Hot Chili Peppers* series was also too small to perform the causality test.

Table 4.4 summarizes the final results of the Granger causality test for the analyses considered: percentage of artists, the test result, and the confirmed hypothesis. Analyzing the two collaborative profiles (*Diverse* and *Regular*), when collaborating, most artists consider the level of success of their collaborators. Likewise, the growth of an artist's

Table 4.4: Summary of main results (Tables 4.2 and 4.3)

	COLLAB & POP				*SOLO & POP*		
	Artists (%)	Causality Relation	Hypothesis		Artists (%)	Causality Result	Hypothesis
Diverse	40%	*COLLAB* ↔ *POP*	Hypothesis 3	Diverse	80%	*SOLO* → *POP*	Hypothesis 1
	30%	–	Hypothesis 4		10%	*POP* ↔ *SOLO*	Hypothesis 3
	20%	*POP* → *COLLAB*	Hypothesis 2		10%	–	Hypothesis 4
	10%	*COLLAB* → *POP*	Hypothesis 1				
Regular	40%	*COLLAB* ↔ *POP*	Hypothesis 3	Regular	70%	*SOLO* → *POP*	Hypothesis 1
	30%	*COLLAB* → *POP*	Hypothesis 1		20%	*POP* ↔ *SOLO*	Hypothesis 3
	20%	–	Hypothesis 4		10%	–	Hypothesis 4
	10%	*POP* → *COLLAB*	Hypothesis 2				
Absent	50%	*COLLAB* → *POP*	Hypothesis 1	Absent	70%	*SOLO* → *POP*	Hypothesis 1
	30%	*POP* → *COLLAB*	Hypothesis 2		20%	*POP* ↔ *SOLO*	Hypothesis 3
	20%	–	Hypothesis 4		10%	–	Hypothesis 4

Hypotheses Summary

1. The way artists professionally connect to each other is useful in forecasting their musical success.
2. The musical success of an artist is useful in forecasting his/her collaborative profile.
3. The musical success of artists assists in forecasting their collaborative profile, as well as their profile helps in forecasting their success.
4. There is no (Granger) causality relationship between collaboration profiles and musical success.

success is affected by the same musical collaborations. That is, the results indicate the presence of contemporary feedback between the two variables, forming a cycle between collaborations and musical success.

Regarding only the *Diverse* profile, there is a significant unilateral causal relationship between the solo songs and the artists' success, as well as a unilateral causal relationship between artistic success and musical collaborations. The first observation was predictable: as we selected only the most popular artists from each profile, it was expected that their solo songs would affect their musical success. However, the second observation shows that, contrary to expectations, the musical success of an artist affects the forecast of musical collaborations more than the opposite. Although it goes against our assumptions, such results still make sense. In Chapter 3, we found that *Diverse* was the most collaborative profile and composed of the most successful artists. Now, we can conclude that a large number of collaborations performed by the artists of this profile is because they are successful artists. That is, other musicians, seeking to increase their professional success, will always attempt to collaborate with the artists in such a group.

Regarding the *Regular* profile, the results also suggest the existence of a significant unilateral causal relationship between the solo songs and the artists' success. However, in this case, there is also a unilateral causal relationship between musical collaborations and artistic success. As in the *Diverse* profile, the first observation was expected. Nevertheless, unlike the previous analysis, the second observation is consistent with our thesis. Being a profile composed of artists with success and an ordinary amount of collaborations, the eventual musical collaborations would cause (in Granger sense) the increasing of these artists' success.

Finally, regarding the *Absent* profile, the results suggest the existence of a significant unilateral causal relationship between the solo songs and the artists' success, once again. Moreover, for most artists, a unilateral causal relationship between musical collaborations and artistic success has been identified. As the *Absent* profile is composed of non-collaborative artists, this observation was less evident. However, through the current results, we can conclude that the few collaborations performed by these artists exposed a strong impact on their success. That is, as shown earlier (Chapter 3), the reason these artists are the least successful (compared to the other profiles) may be related to this causal relationship. In short, because they collaborate very little, they may be missing the opportunity to improve and expand their musical success.

4.3 Overall Considerations

In this chapter, we collected data from the 30 most successful artists from three different collaboration profiles. We then defined a temporal measure of musical success by grouping the *rank_score* of 18 Billboard charts. From our dataset, we created three time-series for each of the top 30 artists: collaborative songs, solo songs and musical success. With data from these time series and the success measure, we applied the Granger causality test to explore the causal relationship between collaborative profiles and artist's success. By conducting two analyses in parallel, the results for the causality test showed the causal relationship between the two variables is not obvious in general.

Combining the current results with observations from the last chapter, we were able to better understand the relationship between artists' collaborative patterns and their musical success. Specifically, we have evidence to support that artistic success strongly affects establishing musical collaborations in a *Diverse* profile. In addition, we confirmed the presence of a cycle between musical collaborations and musical success in the most collaborative profiles (*Diverse* and *Regular*). Finally, we conclude *Absent* may be the profile with less successful artists with exceptions such as Adele, as it is a non-collaborative profile.

Limitations. Despite the relevant results, the dataset employed provides some constraints. By considering a limited number of successful artists, the data analysis risks becoming generic and biased. Therefore, including a more complete dataset, containing both popular and non-popular artists is certainly one of the following study steps. Another notable limitation is the questionable causal notion incorporated in Granger causality concept. According to [53], econometric concepts such as Granger causality are classified as statistical rather than causal. Hence, to further understand the causal relationships between collaborative profiles and artist's success, we plan to deeply investigate techniques of the Pearl Causal Model (PCM).

Chapter 5

Collaboration-Aware Multimodal Hit Song Prediction

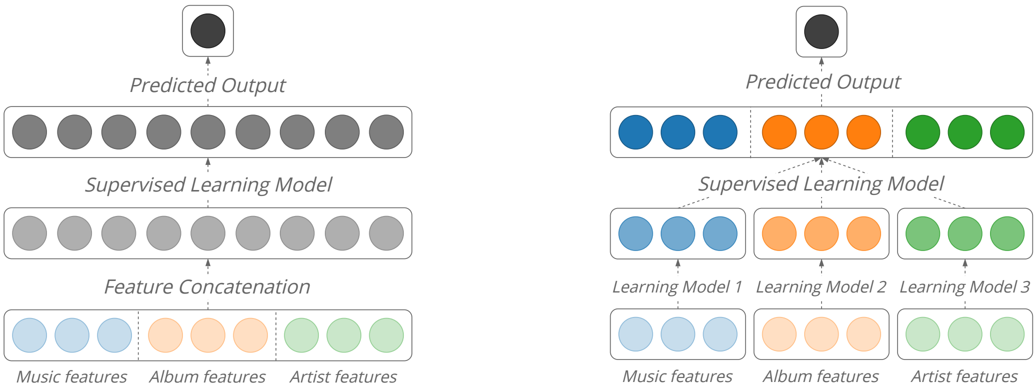
Features extracted from songs' audio, albums and artists' collaborations can provide complementary information to identify potential factors related to songs' popularity. Consequently, a proper description of a song may help machine learning algorithms to predict whether it will become a hit or not. Prior studies on hit song prediction focus on describing them by exploiting features from unique modalities, such as only using audio-based data. In this chapter, we propose including different modalities simultaneously to properly describe songs. Thus, we extract acoustic features, characteristics of the album, artistic collaboration, among others, by considering multiple modalities. We claim that each modality is potentially relevant to describe a song and, consequently, predict its popularity. Since prior chapters strongly favor artistic collaboration as an important metric of song success, we emphasize its use.

The remainder of this chapter is organized as follows. Section 5.1 gives a quick overview on *Multimodal Learning* followed by our proposed methodology to assess the Hit Song Prediction problem, in Section 5.2. We detail the results and experimental evaluation of two distinct tasks in Sections 5.3 and 5.4. Finally, Section 5.5 addresses overall considerations.

5.1 Fundamental Concepts

In this section, we briefly review *multimodal machine learning*, which builds models that can process and relate information from multiple modalities. Here, we focus on multimodal data fusion, one of the five challenges surrounding such a multi-disciplinary field[1]. Technically, multimodal fusion reflects data integration from two or more modalities

[1] The other four challenges on multimodal learning are: representation, translation, alignment, and co-learning (more information in [6]).



(a) Early Fusion (b) Late Fusion

Figure 5.1: Multimodal data fusion approaches for *Hit Song Prediction*. (a) Early Fusion: concatenates features from different modalities, then, a classifier is trained using this common feature vector in order to form the final prediction model. (b) Late Fusion: a classification model is trained separately for each modality, and the individual results are merged into a final prediction model.

to perform a prediction task [41, 76, 25]. It usually allows more robust predictions by providing access to multiple modalities that represent the same phenomenon. Such access may also capture complementary information not visible in the individual modalities and operate with missing modalities.

[6] classify multimodal fusion into: *model-based* – multimodal extensions of existing models; and *model-agnostic* – not directly dependent on a specific learning model. Here, we focus on the latter by using different learning methods for the prediction task. Such approaches are generally *early* (feature-based) or *late* (decision-based) fusion [4] as simplified in Figure 5.1.

5.1.1 Early Fusion

Early fusion, or feature-based, requires training only a single learning model by aggregating data (often by concatenating their representations) from all the modalities. Its learning phase is simple, as only one model is involved, and allows to explore the interactions between features from distinct modalities. Yet, combining features into a common representation is usually challenging. The overall flowchart for early fusion is illustrated in Figure 5.1a.

5.1.2 Late Fusion

Late fusion, or decision-based, learns semantic concepts directly from unimodal features and focuses on the individual strength of modalities. It allows using distinct models on each modality, as different predictors can model each individual modality better, then providing flexibility. Also, it is easier to handle over missing modalities as the predictions are trained individually. However, because late fusion operates on inferences and not on the raw input, it ignores low level interactions between the modalities. The overall flowchart for late fusion is illustrated in Figure 5.1b.

5.2 Methodology

We now detail the new methodology to tackle Hit Song Prediction problem as defined in Section 5.2.1. Then, Section 5.2.3 describes how we model the solution by including artists collaborations, whereas Section 5.2.2 covers the actual multimodal features used for the multimodal strategy. Last, we present the dataset in Section 5.2.4.

5.2.1 Problem Definition

We define Hit Song Prediction as two distinct tasks summarized as follows:

- *Binary Classification:* Given a song, the problem is to predict whether it will be a hit or not; and
- *Placement:* Given a set of existing hit songs sorted by a popularity measure, the problem is to predict in which position a new song will be placed at its release.

Indeed, current work on Hit Song Prediction (e.g., [3, 43]) treats the problem as a regression (or ranking) or binary classification. For a thorough analysis, we regard both tasks as independent. Overall, the main novelty is to solve them by adopting multimodal strategies that consider not only the acoustic features of a song but also its artists' collaborative interactions. Next, we define such tasks.

Binary Classification. Let $\mathcal{X}$ denote a set of songs ordered by their release date, and $\mathcal{Y} = \{1, 0\}$ be the label space (i.e., 1 = *hit* or 0 = *non-hit*). The task of binary classification is to learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ from the training set $\{(x_i, y_i) \mid 1 \leq i \leq m\}$, where $x_i \in \mathcal{X}$ is an instance characterizing the features of a song[2], and $y_i \in \mathcal{Y}$ is the corresponding target value. In this task, the train-test split must be done at a given time t, as the chronological order of the songs is important for the prediction.

Placement. Let $\mathcal{H} = \{s_1, s_2, \cdots, s_n\}$ be a set of hit songs ranked by a popularity measure and s' a new hit song. The placement task aims to find the right position for s' between two instances $(s_i, s_j) \in \mathcal{H} \mid rank(s_i) > rank(s') > rank(s_j)$. This task was first addressed in the context of book sales by [78].

5.2.2 Multimodal Features

Different information may be associated with a particular song, such as lyrics [13], Musical Instrument Digital Interface (MIDI) [55], listener-based information [57], and so on. Learning from multimodal sources improves machine learning models, as a single modality with complete knowledge of the raw data is unusual. That is, representing music only through acoustic features is a restrictive choice. Hence, we explore three modalities as follows.

Music Features. We divide music features into two categories: *internal*, which depends exclusively on resources extracted from the audio, or acoustic fingerprints; and *external*, which considers aspects of the musical ecosystem, e.g., textual and numeric metadata. As internal, we use acoustic fingerprints that are objective (key*, loudness*, mode*, time signature*, and tempo*) or subjective (acousticness*, danceability*, energy*, instrumentalness*, liveness*, speechiness*, and valence*).[3] We also use song duration in milliseconds and the explicitness of lyrics. As external, we consider metadata such as track number, number of artists who have performed the song, number of countries in which the song can be played, and number of years the song has been in the Hot 100 since its release.

Artist Features. Chapters 3 and 4 show artists' professional connections may affect musical success. Therefore, to capture this social information, we consider four social network node-dependent metrics (Clustering Coefficient, Eigenvector, Degree and Weighted Degree) and three graph-dependent metrics (Closeness, Eccentricity, and Betweenness) [34].

[2] In early fusion, x_i vector refers to concatenation of features, whereas in late fusion, x_i vector refers to one modality at a time.

[3] Terms marked with * are further explained in A.3.

We also consider all three collaboration profiles identified through the topological metrics (Chapter 3): *Diverse* for highly collaborative and influential artists; *Regular* for normally collaborative artists; and *Absent* for non-collaborative bands and artists. Finally, we consider the number of genres an artist is associated with and the number of artist's albums.

Album Features. We consider two album metadata: *album_type* that classifies albums in *album*, *single* or *compilation*; and *album_total_tracks*, i.e., number of tracks in an album.

Uniqueness. A key property of a multimodal strategy is *uniqueness*, which is necessary to achieve interpretability, i.e., to assign physical meaning to a phenomenon [36]. Here, artist's and album's features are unable to represent a unique song. For instance, Elvis Presley, who was one of the most productive artists of all time recording over 387 albums[4], could not represent a singular song. Likewise, *21*, Adele's second album ranked as the *Greatest Billboard 200 Album of All Time*,[5] is composed not by one, but 11 different songs. Therefore, to establish uniqueness, we attach acoustic fingerprints in both modalities (*artist* and *album*), as established in the *music* modality.

5.2.3 Collaboration-aware Multimodal Solution

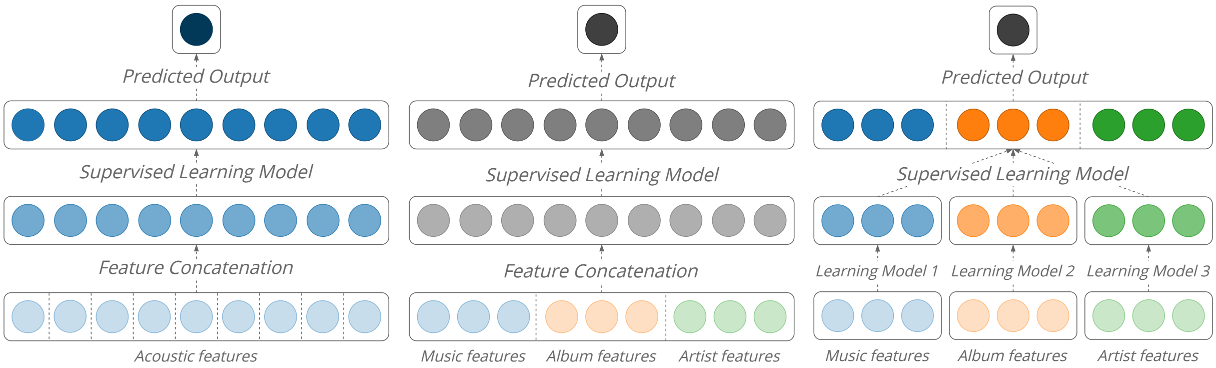
Generally, a *modality* refers to how something exists, is experienced or expressed. Music is characterized as multimodal when expressed by multiple modalities (e.g., melody, rhythm, artist's reputation, collaboration profiles, album information), and each of its modality may be mapped to a popularity metric. Here, we focus on three musical modalities: musical and acoustic features; album features; and artist features, including social and collaboration metrics.

As music may be described by using different modalities, defining how to fuse data obtained from each modality is important. Here, we combine the multimodal features into one multimodal representation using the aforementioned model-agnostic strategies (Section 5.1). Then, we define three data fusion strategies: two using early fusion (data integration before being used by a supervised learning model) and one late fusion (data merging after evaluating each modality), as follows.

EF-music. It is an early fusion strategy for integrating only music related features. After the features being aggregated (through simple concatenation), they become input for a single multimodal supervised learner to predict the output variables. The overall flowchart for **EF-music** is illustrated in Figure 5.2a.

[4]Discogs (10 August, 2020). www.discogs.com/artist/27518-Elvis-Presley

[5]Caulfield, Keith (12 November, 2015). bit.ly/albums-artists-all-time



(a) **EF-music** (b) **EF-all** (c) **LF**

Figure 5.2: System framework for the defined three data fusion strategies. (a) EF-music: early fusion strategy for integrating only music related features. (b) EF-all: early fusion strategy for integrating three musical modalities (music, artist and album). (c) LF: late fusion strategy for trained one classifier separately for each musical modality.

EF-all. It integrates all three musical modalities (Section 5.2.2) by vector concatenation. The overall flowchart for **EF-music** is illustrated in Figure 5.2b.

LF. As a late fusion model, each modality has a specific unimodal supervised learner to predict a correct outcome from labeled data. Thus, the fusion process stems from a weighted voting process, in which predicted probability vectors of each model are summed and averaged. The overall flowchart for **LF** is illustrated in Figure 5.2c.

5.2.4 Experimental Setup

While evaluating our models, we are more interested in investigating if considering social features (specifically, collaboration features) positively impact the performance of predicting hit songs rather than comparing the models' performance with a baseline. Therefore, our experiments address two research questions relevant to each problem task, as follows.

RQ1: *Are acoustic features enough to efficiently predict a hit song?*
RQ2: *Do musical collaboration features affect hit song prediction?*

We investigate them for both binary classification and placement tasks independently. To do so, we propose the following three-step methodology:

1. Choose an appropriate machine learning model for Hit Song Prediction (on the considered task);

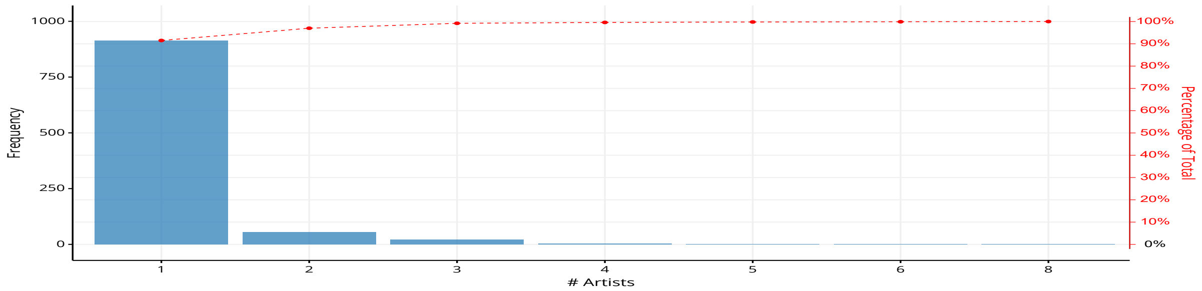


Figure 5.3: Pareto plot of the frequency (blue columns) and the cumulative percentage (red line) of the total number of artists on a song.

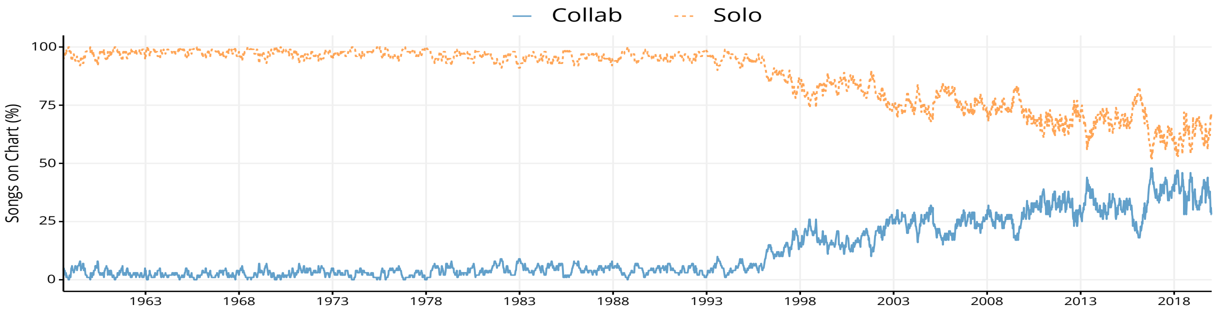


Figure 5.4: Billboard Hot 100 songs (1958 - 2020). The dashed line represents hits without collaborations, while the solid reflects musical collaborations. Clearly, collaborations are increasing in the US music industry with nearly half of all mapped music constituting collaborations.

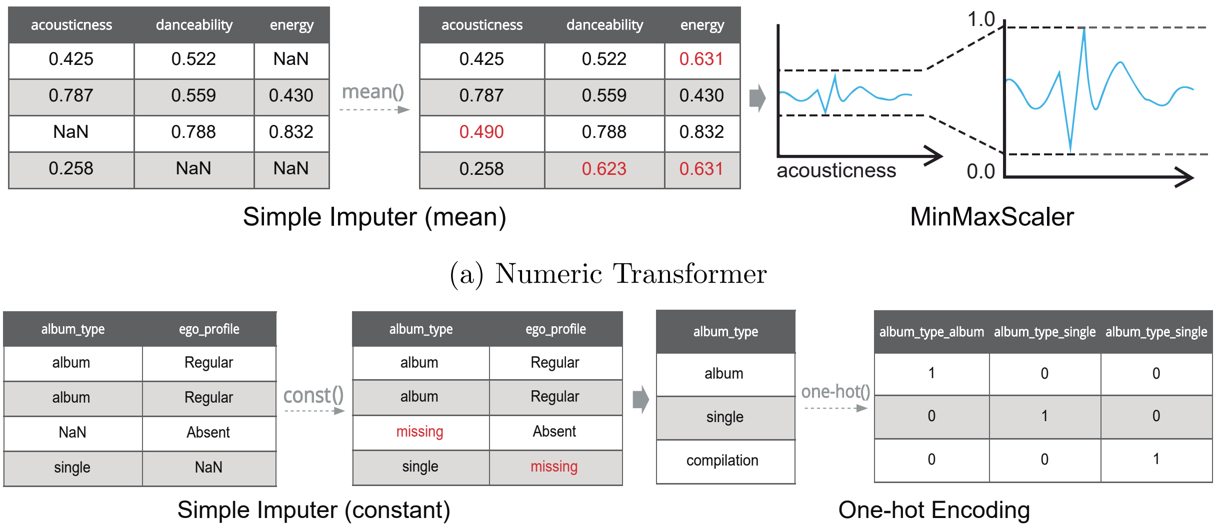
2. Build models for different musical modalities by using multimodal fusion strategies; and

3. Evaluate the built models considering a proper dataset.

As the third step is common to both classification and placement tasks, it is further explained next. Then, steps (1) and (2), which are specific for each task, are explained in Sections 5.3 and 5.4, respectively.

Data

We base our experiments on the freely available *MusicOSet* [68], an open dataset of musical elements (artists, songs and albums) suitable for music data mining (which is also a contribution of this research). The dataset contains 56 years of the Billboard Hot 100 charts, from January 01, 1962, to December 31, 2018. To simplify the modeling process, we filter the data by considering only songs with one or two artists, which represent 97% of the dataset according to Figure 5.3. Hence, there is no loss of generality, as the modeling can be extended for songs with more featured artists. In practice, we call the main artist on the song as *ego* and the featured artist as *alter*. Also, recall Figure 1.2 that is repeated here as Figure 5.4 for practical reasons. It shows considerable growth of



(a) Numeric Transformer

(b) Categorical Transformer

Figure 5.5: Data Preprocessing flowcharts. (a) Two-Step Numeric Transformation: Simple Imputer, replacing missing values using the mean along each column; and MinMax Scaler, scaling each feature into a $[0, 1]$ range. (b) Two-Step Categorical Transformation: Simple Imputer, replacing missing values with a constant = 'missing'; and MinMax Scaler, creating a binary column for each category.

musical collaborations from the mid-90s on. Then, to avoid noisy data and for predicting contemporary hits, a second filtering considers the charts from 1995 on. In the end, we also reduce possible bias resulting from changes in the phonographic sector due to technological innovations (e.g., easier distribution, commercialization and dissemination).

Data Preprocessing

Correctly processing data through the learning models requires to handle different ranges and missing data for both numeric and categorical features. Hence, we perform a two-step numerical and categorical transformations as follows. In the numerical transformation (Figure 5.5a), as only 0.28% of the songs have missing values in our dataset, for each numeric attribute, we fill the missing values with its mean value. Although the mean imputation can distort the distribution for the missing variable, it works well with small numerical datasets and is the easiest and fastest way to impute missing values. After this step, all attributes are normalized into a $[0, 1]$ range with the MinMax Scaler [18]. In categorical transformation (Figure 5.5b), for each categorical attributes, we fill missing values with a constant value, avoiding null problems. Finally, to adjust the data to the input format of most machine learning models, we binarize these features through the One-hot Encoding technique [18].

5.3 Hit Song Binary Classification

The main goal is to predict whether or not a given song will be a hit by appearing at the Billboard Hot 100 charts. Such goal is then defined as two tasks, and this section presents the first one: hit song binary classification. We solve it through three data fusion strategies (Section 5.2.3) that consider music, artist and album features from songs. The first prediction task considers classifying an instance into one of two categories: '1' stands for a song that will be a hit, and '0' otherwise. Next, Section 5.3.1 introduces the learning methods for this task. Then, Section 5.3.2 describes how to use such classifiers for prediction, whereas Section 5.3.3 goes over experimental evaluation and discussion.

5.3.1 Learning Methods

For hit song binary prediction, it first selects the best classifier for each data fusion strategy. Hence, we initially use 19 well-established classification models, briefly defined as follows. We refer to related literature for complete definitions [46]. Next, we describe the strategies in which these classifiers are used and how we select the best one for Hit Song Prediction.

- ***Linear models***
  - *Logistic Regression* is a supervised learning algorithm which is mostly used for binary classification problems. Logistic regression is also known as *logit* regression and is used to explain the relationship between a discrete set of classes and one or more independent variables. Logistic Regression uses as cost function, the 'sigmoid function' also known as the 'logistic function'. The sigmoid function maps any real value into another value between 0 (non-hit) and 1 (hit), that is, maps predictions to probabilities.
  - *Perceptron* is a binary classifier based on a linear model. It works by learning a series of weights that correspond to the input features. Each pair of weight and input features is multiplied, and then the results are summed. If the summation is above a certain threshold, the algorithm predicts one class; otherwise, the prediction belongs to a different class.
  - *Ridge Classifier* is a variant of the Ridge Regressor. It first converts binary targets to $[-1, 1]$ and then treats the problem as a regression task, minimizing a penalized

residual sum of squares. Then, the predicted class corresponds to the sign of the regressor's prediction.

– *Stochastic Gradient Descent (SGD)* is a simple yet very efficient approach to discriminative learning of linear classifiers under convex loss functions such as (linear) Support Vector Machines and Logistic Regression.

- ***Tree-based***

  – *Decision Tree* is a supervised learning algorithm that can be used for solving regression and classification problems. In classification, a Decision Tree creates a training model to predict the class of the target variable by learning simple decision rules inferred from prior data (training data).

- ***Support Vector Machines (SVM)***

  – *Support Vector Classification (SVC)* is the classifier variant of SVM. It is based on the idea of finding a hyperplane that best separates a multidimensional space into different classes based on the provided kernel function. Overall, the main objective is to segregate the given dataset in the best possible way, by selecting a hyperplane with the maximum possible margin between support vectors in the given dataset.

  – *NuSVC* is similar to SVC, but it accepts slightly different sets of parameters and has different mathematical formulations. It uses a parameter (nu) to control the number of support vectors.

  – *Linear SVC*, on the other hand, is another implementation of SVC for the case of a linear kernel.

- ***Naive Bayes*** methods are a set of supervised learning algorithms based on the Bayes Theorem. Bayes' theorem describes the probability of an event, with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. There are different Naive Bayes classifiers:

  – *Multinomial Naive Bayes (MNB)* is the Naive Bayes algorithm for multinomially distributed discrete features; e.g., word count in text classification.

  – *Bernoulli Naive Bayes (BNB)* implements the Naive Bayes algorithm for data that is distributed according to multivariate Bernoulli distributions; e.g., binary-valued (Bernoulli, Boolean) variables.

  – *Complement Naive Bayes (CNB)* is an adaptation of the standard MNB algorithm that is particularly suited for imbalanced data sets. It uses statistics from the complement of each class to compute the model's weights.

- ***Neural Networks***

- *Multilayer Perceptron (MLP)* learns a non-linear function approximator for either classification or regression. It differs from logistic regression as there can be one or more non-linear layers (called *hidden layers*) between the input and the output layer.

- ***Nearest Neighbors***
  - *Nearest Neighbors Classification* is a non-parametric algorithm that classifies the test set into some class according to its K nearest neighbors. It stores all available instances and classifies new instances based on a similarity measure (e.g., distance functions).
  - *Nearest Centroid Classification*, similarly to K-means clustering algorithm, represents each class by the centroid of its members.

- ***Ensemble***
  - *Random Forest* is an ensemble approach that can be used to perform both classification and regression tasks. The algorithm combines several decision trees in randomly selected data samples to determine the final classification. Each decision tree is executed in parallel and, in the end, the algorithm selects the best solution through voting.
  - *AdaBoost* is a popular boosting algorithm that fits a sequence of weak learners (i.e., models that are only slightly better than random guessing) on repeatedly modified versions of the data. The predictions from all of them are then combined through a weighted majority vote to produce the final prediction.
  - *Bagging* is an ensemble meta-estimator that fits base classifiers (each) on random subsets of the original dataset and then aggregates their individual predictions (by either voting or averaging) to form a final prediction.
  - *Extra Trees*, or Extremely Randomized Trees, implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
  - *Gradient Boosting* is a generalization of boosting to arbitrary differentiable loss functions. It is an accurate and effective off-the-shelf procedure that can be used for both regression and classification problems in a variety of areas including Web search ranking and ecology.

Table 5.1: Best classifiers for early fusion strategies (EF-music and EF-all), sorted by weighted F1-Score.

	EF-music				**EF-all**		
Classifier	F1-Score	Accuracy	AUC	Classifier	F1-Score	Accuracy	AUC
RandomForest	0.786	0.787	0.832	MLP	0.834	0.833	0.894
ExtraTrees	0.780	0.788	0.844	SVC	0.829	0.823	0.892
GradientBoosting	0.770	0.761	0.831	ExtraTrees	0.825	0.835	0.900
Perceptron	0.769	0.766	0.817	NuSVC	0.824	0.818	0.888
AdaBoost	0.766	0.755	0.830	GradientBoosting	0.822	0.819	0.884

5.3.2 Setup and Metrics

As music may be describe using different modalities, defining how to fuse data obtained from each modality is important. Section 5.2.3 introduced three multimodal data fusion strategies for different classification models to predict whether a song will be a hit or not. Each classification model is trained with 75% of the data (chronologically split from 1995 on), leaving 25% for testing. In this task, all the songs in the dataset are considered, hits and non-hits. Nevertheless, the dataset contains a more significant number of non-hits (899,068 – 98.7%) than hits (11,959 – 1.3%). To overcome this disproportionate ratio, we randomly duplicate observations from the minority class (with replacement) in the training set to reinforce its signal. Note the resampling is done only on the training set or the performance measures could get skewed. The test set continues with a high imbalance level, to mimic real world data, where only few songs can be considered hits. To select the best model, we use the weighted F1-score, a common metric for classification task. We also consider other two evaluation metrics: accuracy and the area under the curve (AUC). We select the best hyperparameters using Grid Search (Table A.2). The experimental results for each step are presented next.

5.3.3 Experimental Evaluation

We now present the experimental evaluation of the aforementioned binary classification models by comparing their performance (Section 5.3.3) and investigating feature importance (Section 5.3.3).

Table 5.2: Best classifiers for late fusion strategy (LF) per modality, sorted by weighted F1-Score.

	Music				**Artist**				**Album**		
Classifier	F1-Score	Accuracy	AUC	Classifier	F1-Score	Accuracy	AUC	Classifier	F1-Score	Accuracy	AUC
RandomForest	0.796	0.800	0.847	RandomForest	0.809	0.818	0.866	RandomForest	0.797	0.800	0.848
ExtraTrees	0.793	0.802	0.864	GradientBoosting	0.809	0.805	0.872	GradientBoosting	0.791	0.785	0.851
AdaBoost	0.789	0.782	0.848	AdaBoost	0.797	0.792	0.855	ExtraTrees	0.779	0.791	0.855
MLP	0.787	0.779	0.854	Bagging	0.790	0.798	0.826	AdaBoost	0.778	0.770	0.844
GradientBoosting	0.786	0.779	0.852	ExtraTrees	0.760	0.792	0.860	NuSVC	0.773	0.762	0.853

Performance Comparison

For fair evaluation, we train and test each model individually. As the train-test split follows a chronological order, we test the models against unseen data (e.g., whether a song released in 2020 will be a hit based on data up to 2019). Table 5.1 presents the five classifiers with the highest F1-scores, their accuracy and AUC score for both strategies: EF-music considers only music modality, and EF considers all modalities. The best performing classifier for **EF-music** is Random Forest (RF), with a weighted F1-Score of 0.786. Its accuracy and AUC score also present high values, making it a very good choice for this model. For **EF-all**, the best classifier is MLP (Multilayer Perceptron), with higher values than those of EF-music. Consequently, we choose MLP as the best for this model. In addition, Support Vector Machine-based classifiers (SVC and NuSVC), which are among the most commonly used classifiers for Hit Song Prediction, also performed well in this model.

Finally, evaluating the **LF** strategy is more complex as there is one classifier for each modality. Hence, we evaluate them individually in Table 5.2. Note the modalities have four common classifiers in the top: RandomForest, ExtraTrees, AdaBoost and GradientBoosting. This may indicate a consistency in the classifiers for the modalities, thus being equally good choices. We choose RandomForest for all three modalities, as it outperforms the others in F1-scores. The late fusion happens only after each modality is evaluated separately, i.e., each of the three classifiers first predicts a class for a given song (according with the respective modality). Then, integrating such classes achieves a final result. Here, the integration comes from a Soft Voting Classifier, which considers different weights for each modality (class-probabilities) to predict the final class. After tuning its parameters with a grid search, we obtain F1-score of 0.82 and AUC of 0.865.

Having selected the best classifier for each model and tuned their hyperparameters, we may now compare them to find the best one for the hit song classification task. Besides computing their F1-scores, to better visualize the performance of the models, we use ROC curves and AUC score (the higher the AUC score, the higher the capacity of predicting a hit). Figure 5.6 presents such results which show the **EF-all** outperforms the others in this task, with AUC of 0.90. Also, the **LF** performs better than **EF-music**, i.e., enhancing the artist modality with collaboration features improves the results. We are now able to answer **RQ1**, as the models considering social collaborative features **do indeed** generate

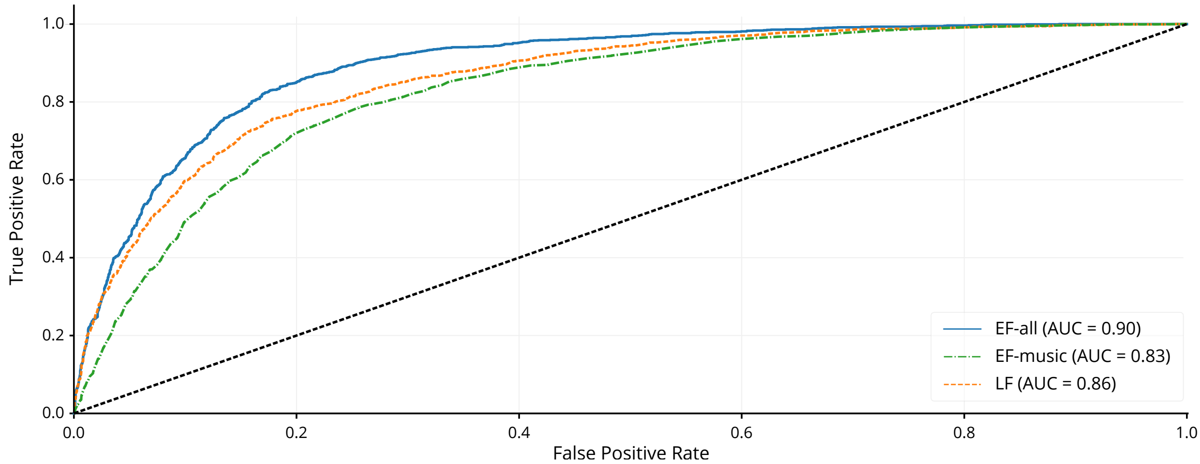


Figure 5.6: ROC curve performance measurement and area under the curve (AUC) score for the proposed strategies for the binary classification task. Each model is indicated with a different color and shape.

more accurate predictions than those without them.

When comparing the two models that consider the collaboration features, **EF-all** outperforms **LF** with AUC score of 0.90 against 0.86. As discussed in Section 5.1, a possible reason is early fusion allows low-level interactions among features, which are very important in our context because music is better defined considering all of its aspects together. Therefore, we choose **EF-all** as the most efficient model for the binary classification task. Next, we analyze the key features that influence its performance.

Feature Importance

Machine learning algorithms can produce good predictions, but their *black-box* nature does not help in understanding highly trained models. Yet, understanding how features influence prediction is still relevant. Hence, we use the SHAP (SHapley Additive exPlanations) [40] method to interpret predictions by computing the contribution of each feature to the results.

The global importance of features included in **EF-all** is illustrated in Figure 5.7 by summary plots. In such plots, all features are vertically sorted by their average impact in the predictions. The feature importance plot (left) is useful, but there is no information beyond the relative importance. The summary plot (right) can further show the positive and negative relationships of the predictors with the target variable, combining feature importance with feature effects. Each point indicates a Shapley value for a feature and an instance. The position on the y-axis is determined by the feature and on the x-axis by the Shapley value, i.e., the impact that feature has on the model's prediction for that song.

Also, Figure 5.7 (right) reveals the direction of feature effects, such as explicit songs (red) having a high and positive impact on the quality rating (the *high* is in red

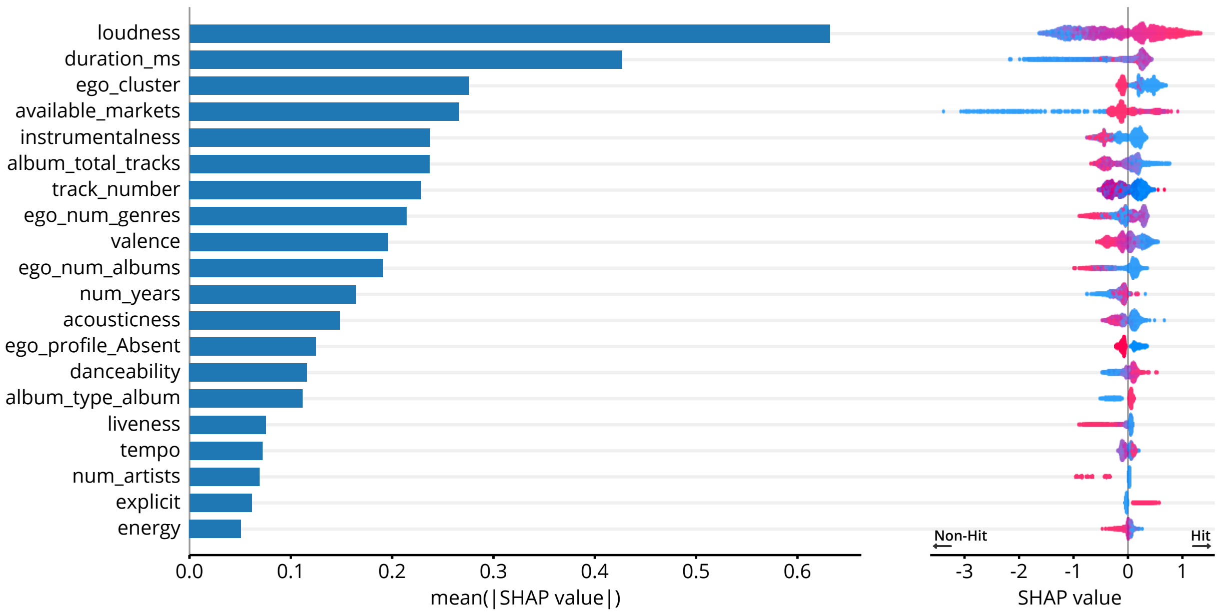


Figure 5.7: Top 20 most significant variables. Left: relative importance as the mean absolute Shapley values. Right: SHAP summary plot of the hit prediction model (each x point is the impact of a feature y on the Hit Song Prediction model for an instance, with values from low/blue to high/red).

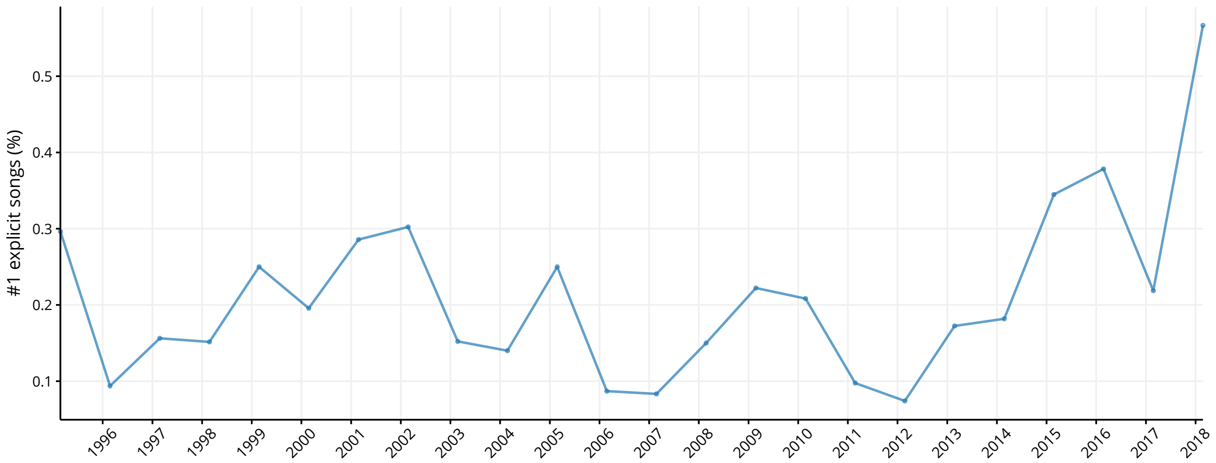


Figure 5.8: Presence of average explicit lyrics in Billboard Hot 100 #1 hit songs in 1995–2018. From 2014 on, at least 20% of top-chart songs have parental advice labels.

color, and the *positive* impact is shown on the x-axis). Such behavior is consistent with the expected and illustrated by Figure 5.8. More than 50% of Billboard Hot 100 number one songs in 2018 feature explicit lyrics. The taste for expletive-filled lyrics has grown since 2012, except for 2017. From 2014 on, at least 20% of number 1 songs have the label of parental advice. Additionally to the direction, the summary plot provides the distribution of effect sizes, such as the long tails of some features. The general trend of long tails reaching to the left, but not to the right, means that extreme values of such measurements can significantly raise non-hit prediction. It also means that features with low global importance (e.g., *ego_num_albums* and *liveness*) can still be important for specific instances.

Overall, the summary plot emphasizes the relationship between a feature value and its impact on the prediction. As expected, most musical features are present in the top 20 of global importance, with *loudness* and *duration_ms* (i.e. track duration in milliseconds) having the maximum impact on the quality rating. Also, both features presented similar effects, with high values associated with positive impact on hit song prediction. In acoustics, *loudness* represents the subjective perception of sound pressure and is directly proportional to the square of the amplitude of vibration. This is compatible with previous intuitions and scientific knowledge. According to [50], there is an evident trend for music to become relatively longer and louder. Hence, the increasing importance of such metrics has become more useful for telling apart a hit from a non-hit.

For *album* modality, there are two features among the most significant predictors: low values of *album_total_tracks* and *album_type_album* feature equal to 1 (i.e., when the album has seven tracks or more), the hit song predictions increases. However, the global importance of both features is quite different, with the number of tracks on a song's album being much more significant than that song being released within an album. In other words, hit songs tend to appear on albums composed of few songs. In the music industry, albums released with one to three tracks are called *singles*. A *single* is frequently a song considered commercially viable enough by the artist and the recording company to be released separately from an album. Therefore, the result is consistent with reality.

Finally, the *ego_cluster* feature is the third most important predictor, changing the predicted absolute hit probability on average by 30% percentage points (0.3 on x-axis). This artist-based feature indicates which cluster an artist is part of, i.e., which is his/her collaboration profile: Diverse, Regular and Absent. Note the Absent profile is among the top 20 most influential features. As suggested by Figure 5.7 (right), an artist who has such a profile (i.e., *ego_profile_Absent* = 1) negatively drives the predictive model to the *non-hit* class; likewise, when equal to 0, the corresponding Shapley values are positive. This means that the collaborative information of the artist significantly affects the accuracy of the model for predicting successful songs, especially when the artist has a collaborative profile (Diverse or Regular). Such results answer our second research question, that is, musical collaboration features indeed *affect* (improve) the prediction of hit songs.

5.4 Hit Song Placement

In this section, we face Hit Song Prediction as a placement task. That is, we employ a machine learning approach (described in Section 5.4.1) that learns to place a new song into a chart ranked by a popularity measure. Popularity of a song can be defined through several aspects such as sales profit, reputation on social media and music platforms, awards received, and so on. Another common approach is to rely on pop charts, such as the Billboard Hot 100, a weekly ranking that lists the top 100 songs based on sales, radio airplay, and streaming activity. Therefore, we can use an inverse-point system of the Billboard ranking score, *rank_score*, as a song popularity measure. The *rank_score* of song i is obtained as $rank\_score(i) = max\_rank - rank(i) + 1$, where max_rank is the lowest rank of the chart and *rank(i)* is the rank of the song.

At first glance, it may seem that simply learning a good regression model is enough for this task, as the target is a continuous output measure. However, in general, traditional approaches of prediction and regression of heavy-tailed outcomes show limited performance on predicting high-value instances [24, 78]. Here, heavy-tailed means a variable with distribution made up of mostly less popular items, with few popular ones, i.e., well-known hits. Most creative industries are driven by sales of a small handful of the most popular releases, such as blockbuster movies [12], art auctions [16], book sales [78] and, mostly relevant to this work, hit songs [9].

To tackle heavy-tailed outcome prediction problem in the book sales domain, [78] introduced the *Learning to Place* (L2P) algorithm, which learns to place a new instance into an ordinal ranking of known instances. Although the *rank_score* (ranges in [1 to 100]) does not follow a heavy-tailed distribution, we adapt the L2P algorithm to predict the position of a new song into a pop chart (given a sequence of previously ranked songs), as the music industry has many common factors with the book industry. Next, we briefly describe the L2P method (Section 5.4.1) and outline how we model it for hit song placement (Section 5.4.2). Finally, we experimentally evaluate the model's performance (Section 5.4.3).

5.4.1 Learning to Place (L2P)

As a classical supervised learning method, in the L2P task one learn from a set of well-labeled input data and uses learned models to predict a quantitative outcome for a given test instance. It has two phases: *training*, which trains a classifier to predict pairwise preferences between each pair of training songs; *testing*, which places a new song q in the given sequence of songs from the training set ranked by *rank_score*. Algorithm 1 summarizes both stages in the context of hit songs.

During training (lines 3-6), for each pair of songs $\{i, j\}$, it concatenates their feature vectors $\{f_i, f_j\}$ (line 4). Next, the problem "becomes" a binary classification based on

Algorithm 1: Learning to Place for Hit Song Prediction

```
   Input: Training songs S, target variable vector t, test feature vector f_q and
          classifier C
1  y = [ ]                                                          # label vector
2  I = [ ]                                                        # voting counter
   # Training Phase
3  foreach pair of train songs (i, j) ∈ S × S, i ≠ j do
4  |   X_ij = ConcatenateFeatVector(f_i, f_j)
5  |   y_ij = CreatePairwisePreferences (t_i, t_j)
6  C.train(X_ij, y_ij)                                           # train the model
   # Testing Phase
7  intv = sort(unique(t))                                     # unique intervals
8  foreach test song q do
9  |   foreach train song i ∈ S do
10 |   |   X_iq = ConcatenateFeatVector(f_q, f_i)
11 |   |   ŷ_iq = C.predict(X_iq)
12 |   |   I = Voting(t̂_iq)                                      # voting process
13 h = GetHighestInterval(I)                        # get the most voted interval
14 t̂_q = mean(intv[h − 1], intv[h])                        # get predicted place
15 return t̂_q
```

the target variable with results: 1 or -1 (line 5). Then, L2P uses the training data as input to a classifier C to predict whether the *rank_score* for i is greater (or small) than j's. During testing (lines 7-14), each test song q is compared with each training song $i \in S$ using the model learned in the training phase to predict the pairwise relations (lines 8-11). Next, L2P treats each training song as a "voter", and sorts them by *rank_score*, dividing the target variable axis into intervals (line 12). After voting, L2P gets the most voted interval, h, and obtains the predicted place $\hat{t}_q$ as the midpoint of h (lines 13-14).

5.4.2 Setup and Metrics

The Billboard Hot 100 chart is weekly released on Tuesdays. It lists the 100 most popular current songs across all genres, ranked by sales (physical and digital), radio play, and online streaming in the United States. To assess L2P performance over time, we analyze each final week's Hot 100 chart of every 2018 month (the last year in the dataset). Unlike the previous task, we use only early fusion to model the problem of placing hit songs. Given that L2P concatenates the two feature vectors $\{f_i, f_j\}$ for each pair of songs (in both training and testing), an automatic column selection of each modality becomes unfit. Hence, the data fusion happens at only one level: feature level (i.e., early fusion, Section 5.2.3).

In total, we train 24 models (two per month), using the *Leave-One-Out* approach to split the data. Then, the L2P is applied once for each data point, using all other songs as a training set and using the selected instance as a test set (singleton). In this task, all the hit songs in the dataset are considered (i.e., all songs featured in the weekly Billboard Hot 100 at least once). We chose the Random Forest as the binary classifier of L2P. According to [78], this meta estimator has good performance (i.e., does not overfit) and provides interpretability of features and results. Finally, to evaluate those models, we consider the following analyses.

Quantile-Quantile (Q-Q) plots. It compares the deviations between true and predicted distributions. This graphical method sorts each data set by value, and then plots them against each other. The closer the values form a straight line, the higher chance to come from a similar distribution.

Kolmogorov-Smirnov (KS) test. It is a common statistical metric for distance between two underlying one-dimensional probability distributions. It is between 0 and 1, and represents how two data sets are similar. A better model presents smaller KS distance.

Earth Mover's Distance (EMD) – first Wasserstein distance. It is a statistical unbounded distance. Small earth mover distance indicates higher similarity between distributions (i.e., better reproducing underlying distribution).

Mean Absolute Error (MAE). As the predicted outcome returned by L2P is a continuous value, we use the MAE loss to compute the average of the absolute difference between the actual and predicted values, without considering their direction. We chose this regression loss metric because it returns errors that are more easily interpretable and is not sensitive towards outliers.

5.4.3 Experimental Evaluation

We now experimentally evaluate L2P for assessing the research questions from Section 5.2.4. We first compare the performance of the early fusion strategies described in Section 5.2.3 (Section 5.4.3), and then conduct a feature importance analysis to find which modalities most affect the hit song placement (Section 5.4.3).

Performance Comparison

For a fair comparison, we train and test each model separately, and discuss our findings per metric.

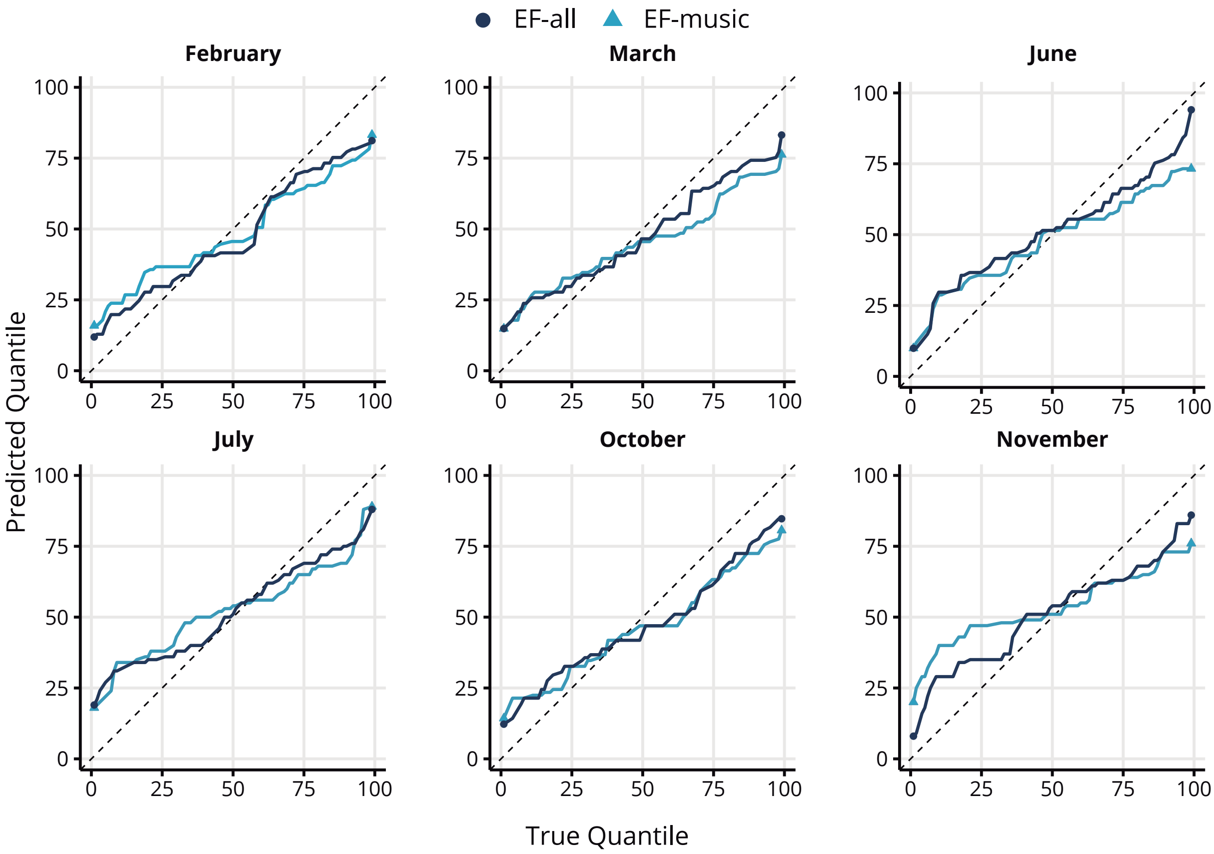


Figure 5.9: Quantile-Quantile (Q-Q) plots for the most significant months, comparing the predicted outcomes with the true distribution for EF-all and EF-music. Each model is indicated with a different color and shape. The dashed line indicates identity. Full Q-Q plots of the other months are given in the Appendix A.4.

Q-Q plot. We assess distribution of predicted outcomes for each month, and show the most relevant in Figure 5.9. Best performing models are near the $x = y$ line (at 45°), and we study those deviating from it. If distributions are linearly related, the Q-Q points form one line, but not necessarily on $x = y$. Overall, the **EF-all** strategy gives the output closest to the ground truth on 45°. Most plots present a deviation from such a line at the high-end. **EF-all** produces the smallest deviation at high quantiles, whereas **EF-music** produces larger deviations on both low- and high- ends.

KS and EMD. While visual inspection is a useful analysis, having a quantitative measure of the similarity between two distributions is crucial. Thus, we compute the well-known KS test and the EMD distance. Table 5.3 summarizes the resulting distances. It confirms that **EF-all** provides the smallest KS and EMD distances, for all months. Although the **EF-music** strategy presents lower KS distances compared to **EF-all**, the observed outcomes are not as discrepant as those resulting from the EMD distance.

MAE. Despite the relevant results, simply comparing the deviations between both distributions is insufficient to evaluate the models by themselves, as the error between predicted and actual *rank_score* is not directly gauged by such metrics. Hence, to quantify

Table 5.3: Performance evaluation for all months. Kolmogorov-Smirnov (**KS**) test and Earth Mover's Distance (**EMD**) compare the similarity between the prediction and the actual distributions. Mean Absolute Error (**MAE**) quantifies performance through the prediction errors

Month	*KS*		*EMD*		*MAE*	
	EF-music	**EF-all**	**EF-music**	**EF-all**	**EF-music**	**EF-all**
January	0.21	0.16	9.61	3.71	0.86 ± 0.18	0.41 ± 0.16
February	0.22	0.16	10.50	4.45	0.89 ± 0.21	0.53 ± 0.18
March	0.21	0.19	8.39	5.67	0.67 ± 0.14	0.47 ± 0.13
April	0.26	0.18	10.89	4.99	0.74 ± 0.17	0.47 ± 0.15
May	0.21	0.17	10.25	5.34	0.74 ± 0.15	0.56 ± 0.13
June	0.28	0.19	8.65	5.83	0.68 ± 0.17	0.52 ± 0.15
July	0.25	0.20	8.29	5.43	0.63 ± 0.17	0.50 ± 0.17
August	0.24	0.17	10.71	7.15	0.72 ± 0.15	0.54 ± 0.14
September	0.22	0.17	9.26	5.96	0.66 ± 0.18	0.51 ± 0.15
October	0.20	0.17	7.56	4.13	0.72 ± 0.16	0.49 ± 0.14
November	0.27	0.18	7.89	4.40	0.68 ± 0.20	0.48 ± 0.16
December	0.26	0.17	6.60	3.41	0.68 ± 0.13	0.38 ± 0.12

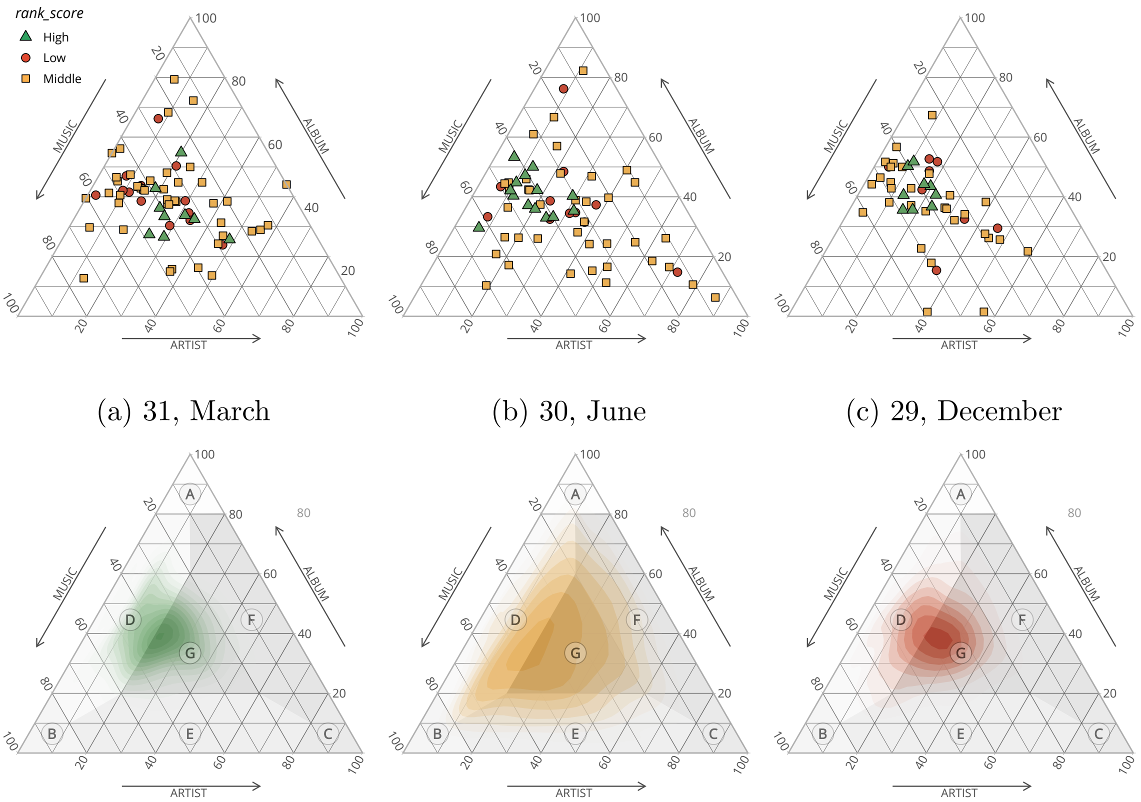
the models' performance using prediction errors, we calculate the MAE regression loss metric. Table 5.3 also summarizes the computed mean absolute errors. Following our other results, **EF-all** achieves the best performance (i.e., lower values). Then, **EF-music** offers higher MAE, and inferior performance under Q-Q plots and KS distance. Such a result points out its inefficiency at an accurate prediction.

General Points. Notably, the **EF-all** strategy yields the best performance for all studied scenarios and in all comparative analyses. We can now safely answer the first research question (**RQ1**): considering the music modality alone does not achieve an efficient prediction. Next, we assess which modalities significantly contribute to the hit song placement by feature importance analysis.

Feature Importance

Following [78], given our three musical modalities (music, artist, album), we asses the relative importance of each specific multimodal feature by training three individual models. We predict the *rank_score* of each song using all three models separately in L2P. We then compare them to the actual *rank_score* of songs and normalize the absolute errors E_{music}, E_{artist} and E_{album}, so that they sum up to one. Finally, we use a ternary plot to inspect the source of errors for different songs.

Figure 5.10 shows ternary plots for songs in three significant time windows (March, June and December). We also color each point and set different shapes based on actual *rank_scores*. Overall, the central left side (Region **D**) has the highest density of absolute errors, being more evident in Figures 5.10b, 5.10c, 5.10d and 5.10e. Most songs in this area have a high-middle actual *rank_score*, which indicates that relying only on *music* and *album* modalities returns the largest prediction error; i.e., having only *music* and/or



(a) 31, March (b) 30, June (c) 29, December

(d) High (e) Middle (f) Low

General explanation. The values of the three modalities *music*, *artist* and *album* sum 100%. The concentration of each modality is 100% in each corner of the triangle and 0% at its opposite line. Besides its three corners, the ternary plot can be divided into seven regions: (**A**) contains at least 80% of E_{album}; (**B**) contains at least 80% of E_{music}; (**C**) contains at least 80% of E_{artist}; (**D**) contains no more than 20% of E_{artist}; (**E**) contains no more than 20% of E_{album}; (**F**) contains no more than 20% of E_{music}; and (**G**) contains at least 20% of each modality.

Figure 5.10: Ternary diagram plots for feature importance. (a, b, c) normalized absolute error for feature group importance per month: a point shows the three L2P normalized values (E_{music}, E_{artist}, E_{album}); and *low*, *middle*, *high* correspond to popularity metric within $[0, 20]$, $(20, 80]$, and $(80, 100]$ respectively. (d, e, f) normalized absolute error accumulated in a year per *rank_score*. Full ternary diagram plots of the other months are given in the Appendix A.5.

album information is not enough for a good prediction. Hence, we can affirm the *artist* modality strongly contributes to the hit song placement task.

The middle of the triangle presents the second-highest density in most months, clearer in Figure 5.10a. In these cases, most songs in the central area have a high *rank_score*, indicating great hit song placement requires excelling in all three dimensions: *music*, *artist*, and *album*. Finally, there is a small concentration of outlier songs with average popularity in the right corner (Regions **E**, **F**, **C** and Figure 5.10b). This behavior means that having only *artist* information can negatively affect model performance for such isolated cases.

5.5 Overall Considerations

In this chapter, we addressed the problem of Hit Song Prediction [51]. Here, we define this problem from two independent tasks: *binary classification* and *placement*. In the former, given a collection of songs, a classifier is trained to predict whether an unreleased song will be a hit or not. While the latter trains a classifier to place a new song into a popularity chart. To tackle both tasks, we model them as multimodal problems, by considering three different modalities: *music*, *artist* and *album*, focusing primarily on artistic collaboration features.

Our proposed methodology sheds light on two research questions defined in Section 5.2.4: (**RQ1**) By comparing the performance of models with different fusion strategies (considering different musical modalities), we found that relying exclusively on internal musical characteristics is not enough to obtain efficiently hit song predictions; (**RQ2**) The feature importance analysis allowed to identify the most significant features that drive hits prediction. We found the *artist* modality contains the most significant predictors, mainly social interaction information. Such results demonstrate not only the relevance of handling Hit Song Prediction as a multimodal problem but also the importance of relying on information from the artists' collaboration profiles. To sum up, our results reveal it is indeed possible to successfully predicting whether or not a given song will be a hit or a non-hit.

Limitations. One limitation of our work is its dataset comprising music charts from the U.S. only. Hence, a natural extension is to consider data beyond U.S., such as European, Latin American and Asian charts. Still, collecting such data and getting it ready for all tasks performed here present some serious challenges, starting at the lack of open online information.

Chapter 6

Conclusion and Future Work

In this chapter, we summarize the main results achieved so far (Section 6.1) and present open problems and future work derived from this dissertation (Section 6.2).

6.1 Conclusions

In this work, we explore the relationship between collaboration profiles and musical success towards predicting whether a given song will rise to a high position in top charts. We categorize such analyses into research goals, discussed next. In addition, we cite the publications that are directly and/or indirectly related to each research goals (all of them were published during the Master's program).

RG1: Identify the (potentially) intrinsic features and indicators that influence the popularity of both songs and artists. [67] We proposed an initial study to analyze and identify music collaboration profiles in a musical success-based network. Specifically, we focused on investigating the impact of these profiles on successful music artists. Through six topological metrics, we defined four key categories of collaboration profiles: *Interaction*, *Distance*, *Influence* and *Similarity*. Among them, we found that the first three affect musical success more intensely than *Similarity*. That is, successful artists are more likely to have a high degree of collaboration between influential and diversified artists. These findings suggest there is a powerful sense that collaborations enhance artists' prospects of having a successful song, not only boosting opportunities for unknown artists but also reinvigorating the careers of more established ones. Nevertheless, those who prefer to pursue a non-collaborative musical career may be missing an opportunity to enhance and expand their potential. Our results provide strong evidence that (*i*) there are, in fact, distinct success factors for musical collaboration profiles that are socially measurable, and (*ii*) there are common factors for successful collaboration in the music market. Overall, our analysis motivates the further study of the causal relationship between such profiles and musical success.

RG2 and RG3: Investigate the impact of these features on popularity over time, and verify the causal relationship between collaboration profiles and music success. [65] Resuming the initial analysis from RG1, we proposed to establish relations of causality between collaboration profiles and artists' popularity. Previous findings reveal that the way artists connect professionally may affect their musical success. Therefore, we further such analyses by using time series and the Granger causality test for assessing whether there is a causal relationship between collaboration profiles and artistic success. Our experimental evaluation reveals varied relationships linking collaboration profiles and musical success, indicating a direct connection. In particular, we identified the presence of a cycle between collaborations and musical success. Namely, before collaborating with others, most artists consider the level of success of their collaborators. It means collaboration can be seen as a musical career propeller, working as a fast-track route to becoming a household name, especially when reaching out to more established artists. Furthermore, based on findings regarding to RG1, we found that the few collaborations performed by non-collaborative artists exposed a strong impact on their success. However, such artists are part of the group with the lowest average success rate. This indicates that, because they collaborate very little, they may be missing the opportunity to grow their reach and inspire themselves to new creative heights. Whereas our results showed an unclear causal relationship between collaboration profiles and musical success, in general, these findings offer a novel perspective on success in the music industry, unraveling how collaboration profiles can contribute to an artist's popularity.

RG4: Propose a machine learning approach to derive a song's popularity based on these groups of features and determine the best way for combining them to predict the success of a song. [68, 66] As a result of the well-established correlation between collaboration profiles and musical success, a natural next step is to propose a methodology for predicting hit songs. Particularly, we tackle Hit Song Prediction problem from two distinct tasks: *binary classification* and *placement*. To assess both tasks, we model a song as a multimodal representation using information from three modalities: *music*, *artist* and *album*. In order to connect and verify our results found so far, we focused primarily on artistic collaboration features. This research shows not only the relevance of handling Hit Song Prediction as a multimodal problem but also the importance of relying on social information, specially from artists' collaboration profiles. Finally, these positive results point out that popularity **can indeed** be learned from different heterogeneous information. Furthermore, they suggest that features commonly used for represent music content are not informative enough to outperform collaboration-aware strategies.

Relevance Insights. Our work differs from the current state of the art in two crucial ways. First, although considering social aspects in hit song prediction is beneficial from the analytical perspective, this is the first time that the collaboration between artists and their profiles are modeled as features for a machine learning approach. Second, the multimodal perspective brings the necessary complexity to analyzing music in all its facets. Therefore, combining the multimodal representation with a collaboration-aware model means a big step towards advancing both fields of *Hit Song Science* and *Multimodal Music Analysis* [15], providing potential impact to the music industry.

6.2 Future Work

One limitation of our work is the dataset that includes music charts from the U.S. only, generating a cultural monopolization. Furthermore, despite the significant results, we consider relatively few factors in each of the three modalities, i.e., music, artist and album. Another limitation is the statistical notion incorporated in the Granger Causality concept rather than actual causal inference. Such limitations point to the following future direction:

- Since our data source considers only mainstream and popular music from the U.S. industry, cultural and gender diversity is precarious. Hence, as future work, a natural extension is to consider data beyond U.S., such as European, Latin American and Asian charts. Still, collecting such data and getting it ready for all tasks performed here present some serious challenges, starting at the lack of open online information.
- We also plan to include other interactions in social media in our multimodal approach, and other characteristics on this context, such as artist reputation. Our proposed prediction models may be improved with a more complete dataset.
- Finally, to further understanding the causal relationships between collaborative profiles and artist's success, we plan to deeply investigate techniques of the Pearl Causal Model (PCM).

Appendix A

Further Information

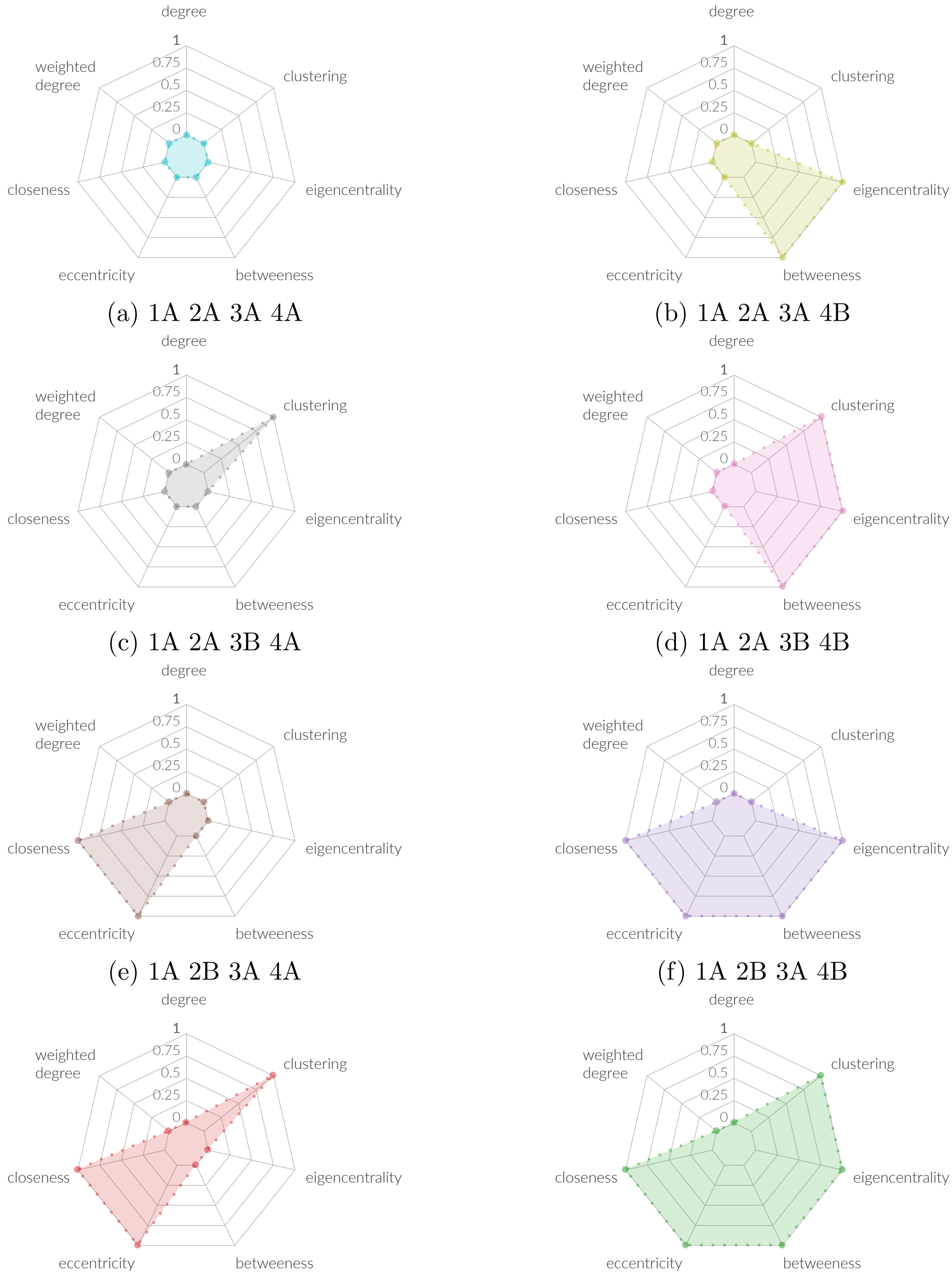
A.1 Collaboration Profiles

We infer 16 different collaboration profiles from the four categories defined in Section 3.2.4. Table A.1 and Figures A.1 and A.2 presents their characteristics, each one represented by a topological metric. In summary, we use the characteristics to represent the threshold levels: 1 for a high metric value, that is, greater than or equal to 0.5 ; or 0 for low metric values, that is, less than 0.5.

Table A.1: Standard Collaboration Profiles

	Interaction		Distance		Genre	Influence	
Profiles	Degree	Weighted Degree	Eccentricity	Closeness	Clustering	Betweenness	Eigencentrality
1A 2A 3A 4A	0	0	0	0	0	0	0
1A 2A 3A 4B	0	0	0	0	0	1	1
1A 2A 3B 4A	0	0	0	0	1	0	0
1A 2A 3B 4B	0	0	0	0	1	1	1
1A 2B 3A 4A	0	0	1	1	0	0	0
1A 2B 3A 4B	0	0	1	1	0	1	1
1A 2B 3B 4A	0	0	1	1	1	0	0
1A 2B 3B 4B	0	0	1	1	1	1	1
1B 2A 3A 4A	1	1	0	0	0	0	0
1B 2A 3A 4B	1	1	0	0	0	1	1
1B 2A 3B 4A	1	1	0	0	1	0	0
1B 2A 3B 4B	1	1	0	0	1	1	1
1B 2B 3A 4A	1	1	1	1	0	0	0
1B 2B 3A 4B	1	1	1	1	0	1	1
1B 2B 3B 4A	1	1	1	1	1	0	0
1B 2B 3B 4B	1	1	1	1	1	1	1

We use radar charts to represent the collaborative characterization of each 16 profiles presented in Table A.1. Radar Charts are usually used for comparing multiple quantitative variables. This makes them useful for seeing patterns in the data. In short, each metric is provided with an axis that starts at the center; each metric value is plotted along its individual axis and all variables in a dataset and connected together to form a polygon. Here, the polygon represents exactly the characterization of a collaboration profile.

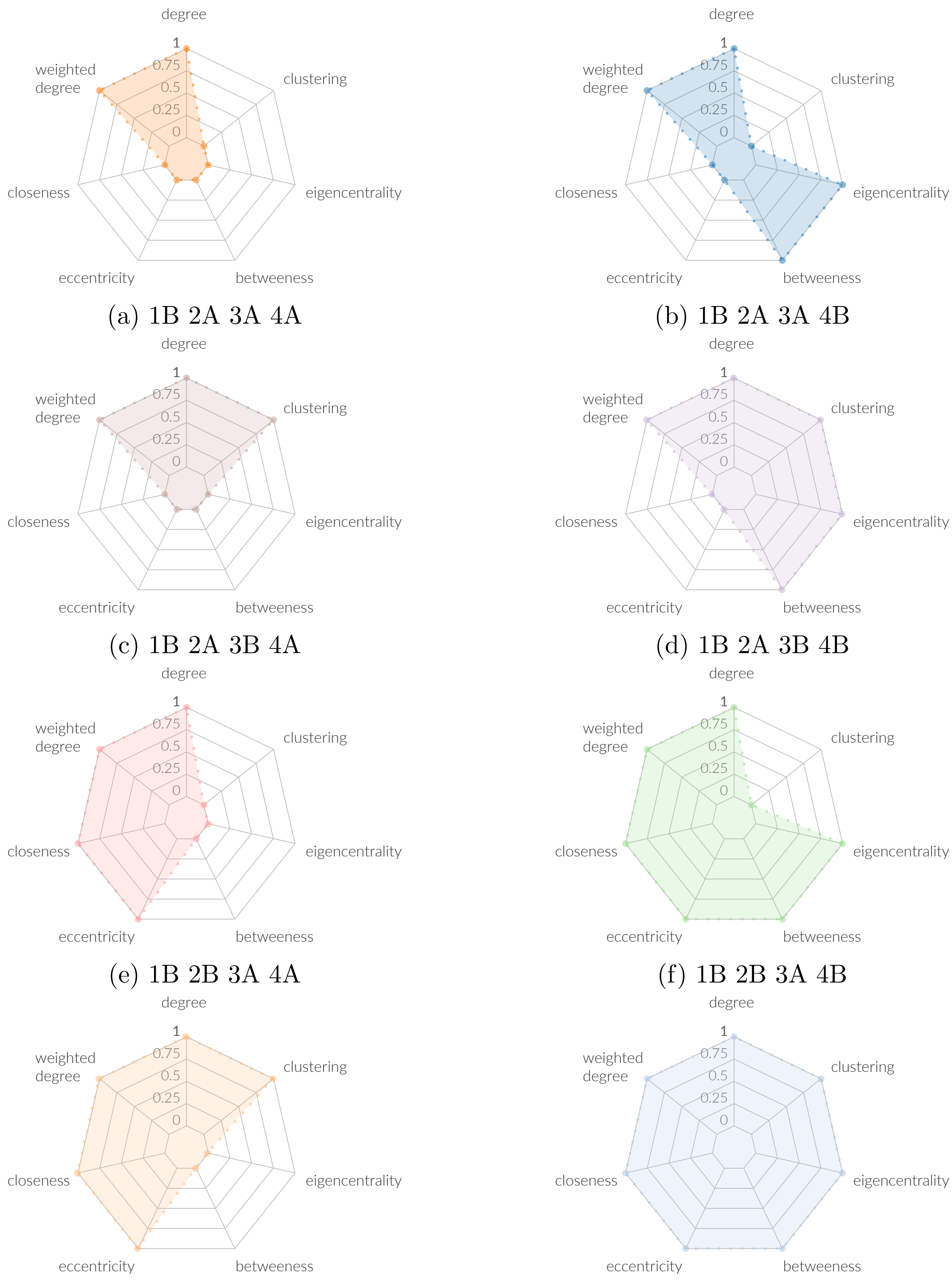


(a) 1A 2A 3A 4A
(b) 1A 2A 3A 4B
(c) 1A 2A 3B 4A
(d) 1A 2A 3B 4B
(e) 1A 2B 3A 4A
(f) 1A 2B 3A 4B
(g) 1A 2B 3B 4A
(h) 1A 2B 3B 4B

Figure A.1: Radar Plots of each collaboration profile (Part 1).

A.2 Correlation Tests

Scatterplots of each pair of numeric variable collected from Spotify are drawn on the top-right part of Figure A.3. Pearson's (r), Spearman's (ρ) and Kendall's (τ) correlations



(a) 1B 2A 3A 4A (b) 1B 2A 3A 4B

(c) 1B 2A 3B 4A (d) 1B 2A 3B 4B

(e) 1B 2B 3A 4A (f) 1B 2B 3A 4B

(g) 1B 2B 3B 4A (h) 1B 2B 3B 4B

Figure A.2: Radar Plots of each collaboration profile (Part 2).

is displayed on the bottom-left. Variable distribution is available on the diagonal.

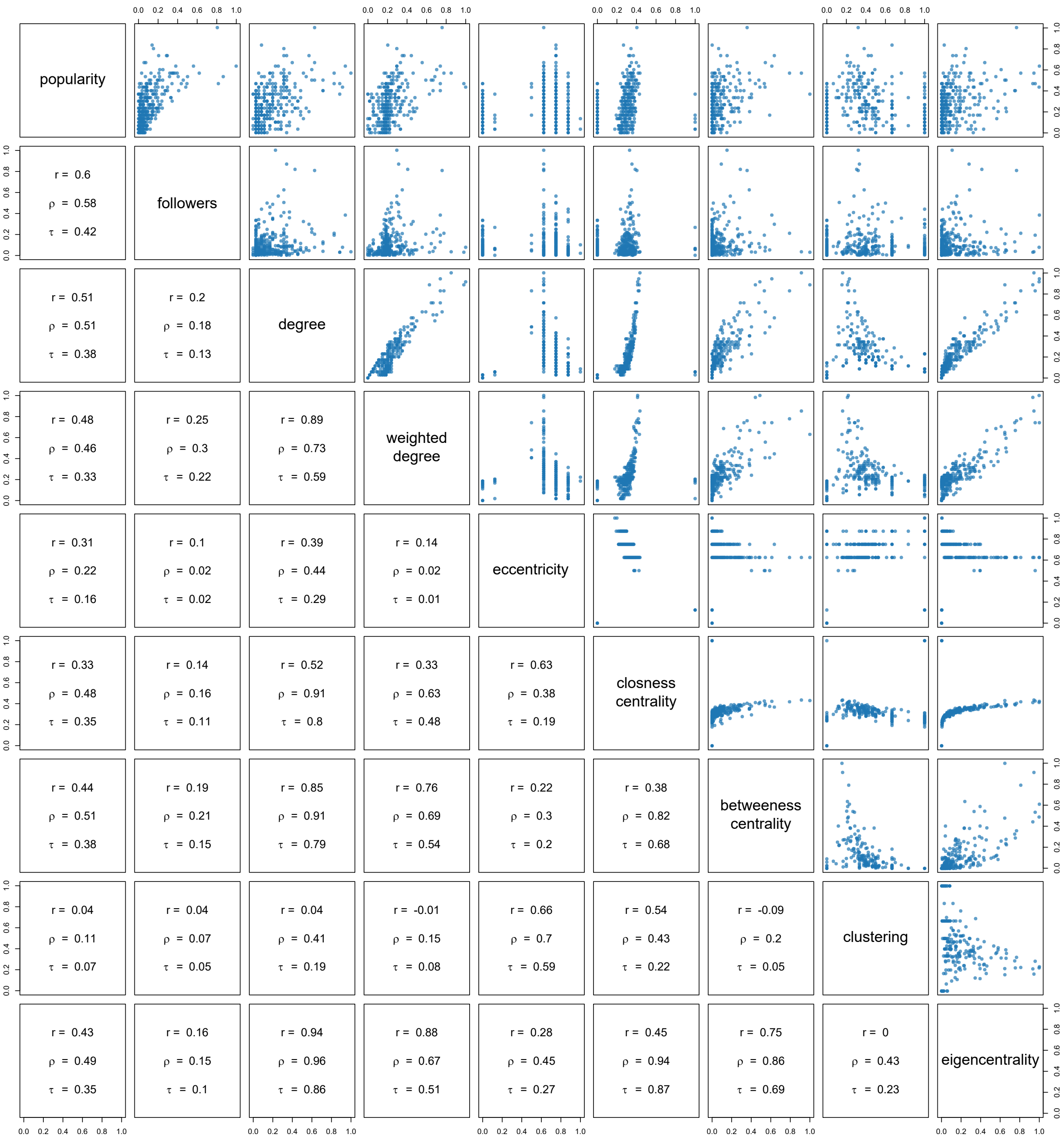


Figure A.3: Scatterplot matrix of topological metrics and success measures (r: Pearson, ρ: Spearman, and τ: Kendall).

A.3 Music Features Description

We now explain some features used to build the modalities considered in our methodology. The dataset used in this work contains acoustic fingerprints collected directly from Spotify[1]. Some of them are objective, while others are more subjective.

Objective Features

[1]Spotify API Doc: https://developer.spotify.com/documentation/web-api/reference/

Key: the estimated overall key of a song, mapped as an integer number using standard Pitch Class notation.

Loudness: the general loudness measured in decibels (dB).

Mode: the general modality of a song (i.e., major or minor).

Time Signature: the amount of beats in each bar (measure).

Tempo: the speed of the song, measured in beats per minute (BPM).

Subjective Features

Acousticness: informs the probability of a song to be acoustic or not.

Danceability: combines *tempo*, rhythm stability, beat strength, and other elements to describes how suitable a song is for dancing.

Energy: represents the intensity and activity of a song by combining information such as dynamic range, perceived loudness, timbre, onset rate, and general entropy.

Instrumentalness: measures the probability of a song to be instrumental, that is, not contain vocals. For example, "ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal".

Liveness: detects the presence of an audience in a song. The higher the liviness value, the higher the probability of a song being performed live.

Speechiness: measures the probability of a given song to have spoken words in it.

Valence: describes the positiveness within a song. High valence values represent happier songs, whereas low values characterize the opposite.

A.4 Quantile-Quantile (Q-Q) Plots

Q-Q plots are a handy tool for comparing two distributions. For each month, we compare the deviations between true and predicted distributions by plot them against each other. The closer the values form a straight line, the higher chance to come from a similar distribution. Best performing models are near the $x = y$ line (at 45°), and we study those deviating from it. If distributions are linearly related, the Q-Q points form one line, but not necessarily on $x = y$.

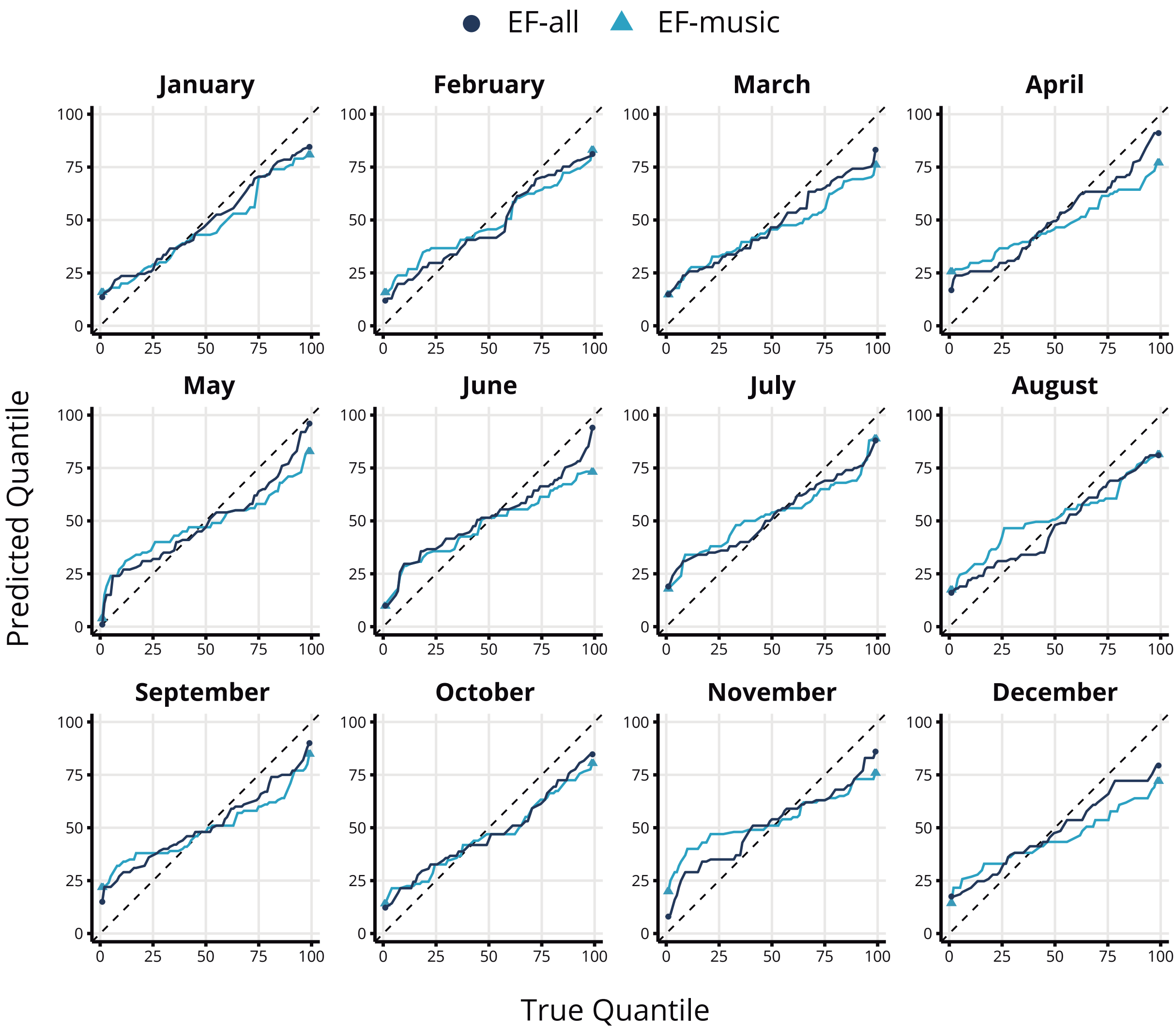


Figure A.4: Full Quantile-Quantile (Q-Q) plots for all months in 2018, comparing the predicted outcomes with the true distribution for EF-all and EF-music. Each model is indicated with a different color and shape. The dashed line indicates identity.

A.5 Ternary Plots

Given the three musical modalities (music, artist, album), we asses the relative importance of each specific multimodal feature by training three individual models. We predict the *rank_score* of each song using all three models separately in L2P. We compare the actual *rank_score* with the predicted value and normalize the absolute errors E_{music}, E_{artist} and E_{album}, so that they sum up to one (100%). We use a ternary plot to inspect the source of errors for different songs, coloring each point and set different shapes based on actual *rank_scores*.

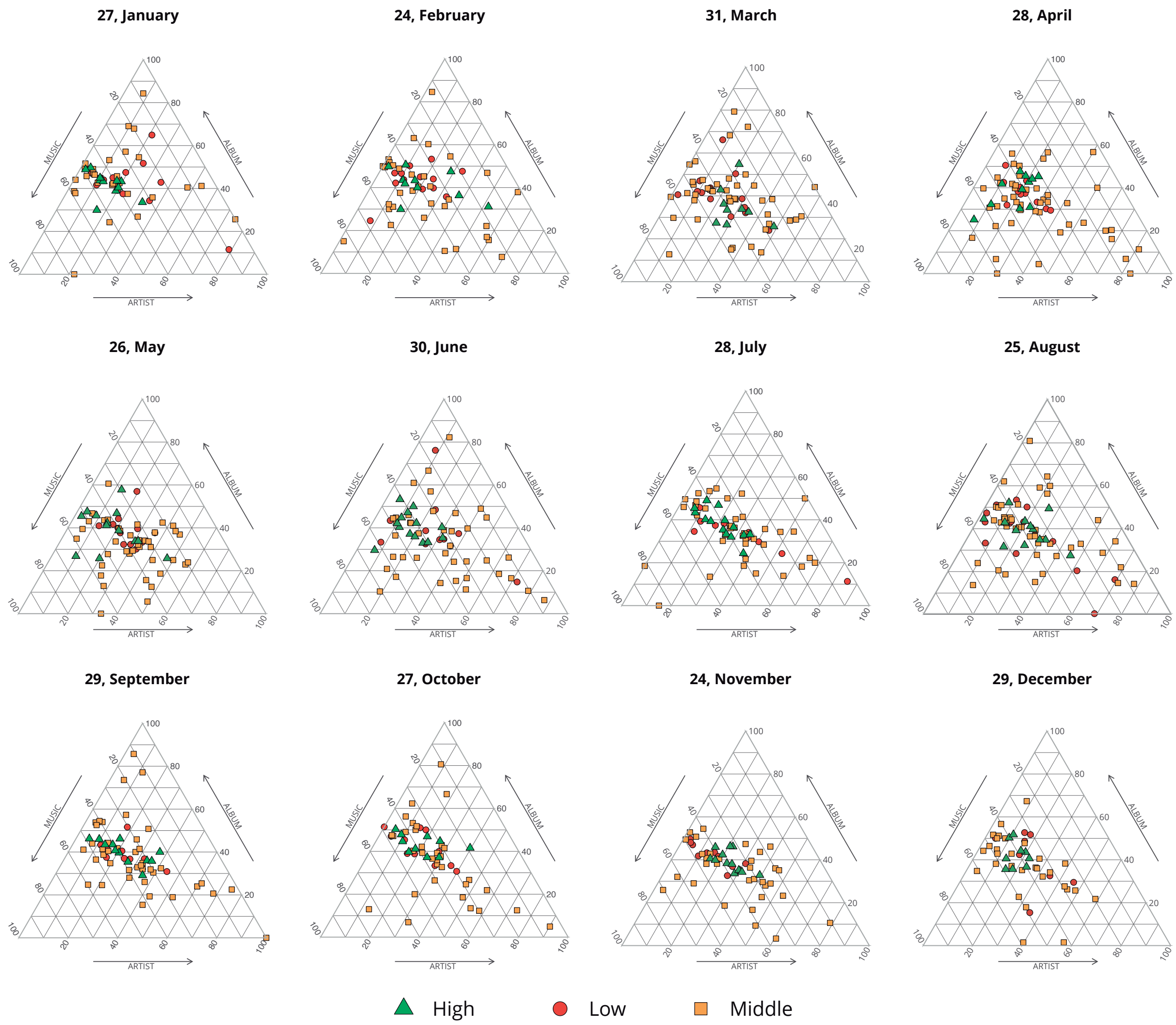


Figure A.5: Ternary diagram plots for feature importance. (a, b, c) normalized absolute error for feature group importance per month: a point shows the three L2P normalized values $(E_{music}, E_{artist}, E_{album})$.

A.6 Experimental Setup Details

Here, we give further information on our experimental evaluation by focusing on the parameters of the classifiers used in our models. We developed all experiments using the Python package Scikit-Learn[2]. This section is divided according to the two tasks assessed in this work: classification (Section A.6.1) and ranking placement (Section A.6.2).

A.6.1 Binary Classification

In this task, one key step is selecting the best classifier for our three proposed models. In the EF-music and EF-all strategies, this comparison was conducted by evaluating 19 classifiers implemented on Scikit-Learn and set with their default parameters. Furthermore, in the LF model we also compare these classifiers for each modality (music, artist and album) and select the best one for each model. Then, we combine their results using a VotingClassifier, also implemented in Scikit-Learn. In all three models, after selecting the best classifier for the model, we run a grid search for finding the best hyperparameters for each model. We do not perform cross-validation on this search, as our data need to be split in chronological order. In the LF model, the parameters of the VotingClassifier were also tuned. Table A.2 presents the parameters tuned for each classifier in each model as well as the considered search space.

A.6.2 Ranking Placement

In this task, we employ the Learning to Place (L2P) algorithm using its default parameters and classifier (Random Forest, also set with default parameters). Figure A.6 graphically describes the main steps for predicting the place of a song in a hit song ranking. The original Python implementation of L2P (by its authors) is available on https://github.com/xindi-dumbledore/L2P.

[2]Scikit-Learn: https://scikit-learn.org/.

Table A.2: Parameter grid for tuning models' hyperparameters, with the best values underlined

	Classifier	Hyperparameter	Search space
EF-music	*RandomForest*	n_estimators max_features max_depth	[100, 200, 500] ['sqrt', 'log2'] [10, 50, 100, None]
EF-all	*MLP*	max_iter alpha	[100, 200, 500, 1000, 2000] [10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}, 10^{-9}]
LF	*RandomForest* (music modality)	n_estimators max_features max_depth	[100, 200, 500] ['sqrt', 'log2'] [10, 50, 100, None]
	RandomForest (artist modality)	n_estimators max_features max_depth	[100, 200, 500] ['sqrt', 'log2'] [10, 50, 100, None]
	RandomForest (album modality)	n_estimators max_features max_depth	[100, 200, 500] ['sqrt', 'log2'] [10, 50, 100, None]
	Voting	weights	[[0,0,1], [0,1,0], [0,1,1], [1,0,0], [1,0,1], [1,1,0], [1,1,1]]

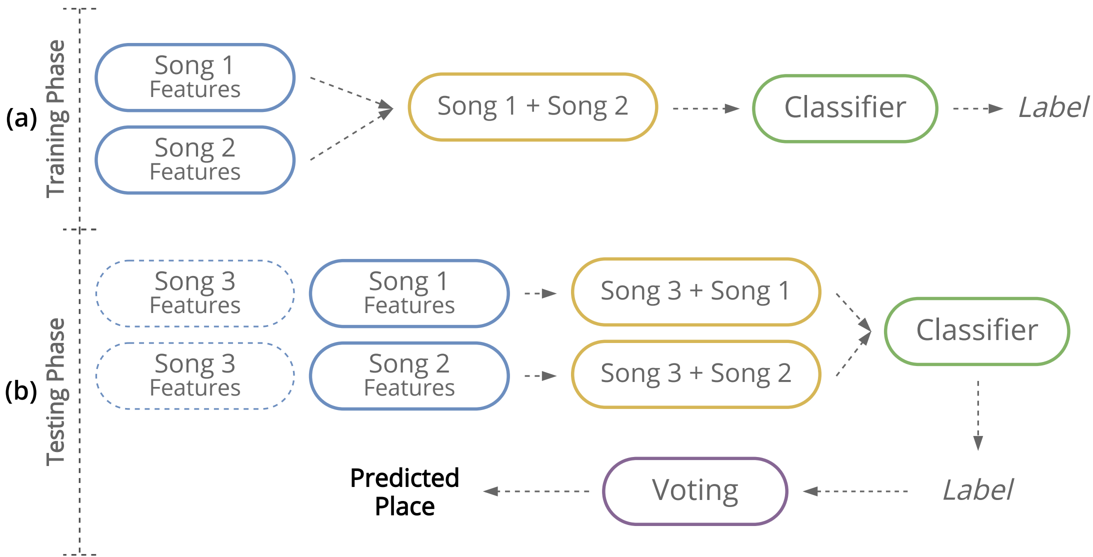


Figure A.6: Learning to Place (L2P) flowchart. (a) *Training:* train a classifier on the pairwise relationship between each pair of train songs. (b) *Testing:* predict pairwise preferences between a new song, *Song* 3, and all train songs using the trained classifier; place *Song* 3 in the given sequence of trained songs ranked by *rank_score* through voting.

Bibliography

[1] Fabian Abel, Ernesto Diaz-Aviles, Nicola Henze, Daniel Krause, and Patrick Siehndel. Analyzing the blogosphere for predicting the success of music and movie products. In *International Conference on Advances in Social Networks Analysis and Mining, ASONAM*, pages 276–280, Odense, Denmark, 2010.

[2] Carlos V. S. Araújo, Rayol M. Neto, Fabíola Guerra Nakamura, and Eduardo Freire Nakamur. Predicting music success based on users' comments on online social networks. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web, WebMedia*, pages 149–156, Gramado, RS, Brazil, 2017.

[3] Carlos Vicente Soares Araujo, Marco Antônio Pinheiro de Cristo, and Rafael Giusti. Predicting music popularity using music charts. In *18th IEEE International Conference On Machine Learning And Applications, ICMLA*, pages 859–864, Boca Raton, FL, USA, 2019.

[4] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, 2010.

[5] Frank B Baker and Lawrence J Hubert. Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, 70(349):31–38, 1975.

[6] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443, 2019.

[7] Kerstin Bischoff, Claudiu S. Firan, Mihai Georgescu, Wolfgang Nejdl, and Raluca Paiu. Social knowledge-driven music hit prediction. In *Advanced Data Mining and Applications, ADMA*, pages 43–54, Beijing, China, 2009.

[8] Fabio Calefato, Giuseppe Iaffaldano, and Filippo Lanubile. Collaboration success factors in an online music community. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork, GROUP*, pages 61–70, Sanibel Island, Florida, USA, 2018.

[9] Òscar Celma and Pedro Cano. From hits to niches? or how popular artists can bias music recommendation and discovery. In *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, pages 1–8, Las Vegas, Nevada, 2008.

[10] Malika Charrad, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. Nbclust: An r package for determining the relevant number of clusters in a data set. *Journal of Statistical Software, Articles*, 61(6):1–36, 2014.

[11] Song Hui Chon, Malcolm Slaney, and Jonathan Berger. Predicting success from music sales data: a statistical and adaptive approach. In *Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia, ACM*, pages 83–88, California, USA, 2006.

[12] Alan Collins, Chris Hand, and Martin C Snell. What makes a blockbuster? economic analysis of film success in the united kingdom. *Managerial and Decision Economics*, 23(6):343–354, 2002.

[13] Ruth Dhanaraj and Beth Logan. Automatic prediction of hit songs. In *Proceedings of the International Conference on Music Information Retrieval, ISMIR*, pages 488–491, London, UK, 2005.

[14] Walter Enders. *Applied Econometric Time Series, 4 Ed.* John Wiley & Sons, 2014.

[15] Slim Essid and Gaël Richard. Fusion of multimodal information in music content analysis. In *Multimodal Music Processing*, volume 3 of *Dagstuhl Follow-Ups*, pages 37–52. Schloss Dagstuhl, Germany, 2012.

[16] Samuel P. Fraiberger, Roberta Sinatra, Magnus Resch, Christoph Riedl, and Albert-László Barabási. Quantifying reputation and success in art. *Science*, 362(6416):825–829, 2018.

[17] Wayne A Fuller. *Introduction to statistical time series*, volume 428. John Wiley & Sons, 2009.

[18] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* O'Reilly Media, 2019.

[19] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 37(3):424–438, 1969.

[20] Mark S. Granovetter. The strength of weak ties. In Samuel Leinhardt, editor, *Social Networks: A Developing Paradigm*, pages 347 – 367. Elsevier, 1977.

[21] Leandro Guedes and Carla M. D. S. Freitas. Exploring music rankings with interactive visualization. In *Proceedings of the Symposium on Applied Computing, SAC*, pages 214–219, Marrakech, Morocco, 2017.

[22] Edward J Hannan and Barry G Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):190–195, 1979.

[23] Dorien Herremans, David Martens, and Kenneth Sörensen. Dance hit song prediction. *Journal of New Music Research*, 43(3):291–302, 2014.

[24] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17:543–582, 2016.

[25] Anthony Hu and Seth Flaxman. Multimodal sentiment analysis to explore the structure of emotions. In *Proceedings of the 24th International Conference on Knowledge Discovery & Data Mining, ACM SIGKDD*, pages 350–358, London, UK, 2018.

[26] Xiao Hu, Ying Que, Noriko Kando, and Wenwei Lian. Analyzing user interactions with music information retrieval system: An eye-tracking approach. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR*, pages 415–422, Delft, The Netherlands, 2019.

[27] Giuseppe Iaffaldano. Investigating collaboration within online communities: Software development vs. artistic creation. In *Proceedings of the 2018 ACM Conference on Supporting Groupwork, GROUP*, pages 384–387, Sanibel Island, FL, USA, 2018.

[28] Myra Interiano, Kamyar Kazemi, Lijia Wang, Jienian Yang, Zhaoxia Yu, and Natalia L Komarova. Musical trends and predictability of success in contemporary songs in and out of the top charts. *Royal Society open science*, 5(5):171274, 2018.

[29] Ioannis Karydis, Aggelos Gkiokas, and Vassilis Katsouros. Musical track popularity mining dataset. In *Artificial Intelligence Applications and Innovations - 12th IFIP*, pages 562–572, Thessaloniki, Greece, 2016.

[30] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[31] Hyoung-Gook Kim, Gee Yeun Kim, and Jin Young Kim. Music recommendation system using human activity recognition from accelerometer data. *IEEE Trans. Consumer Electronics*, 65(3):349–358, 2019.

[32] Yekyung Kim, Bongwon Suh, and Kyogu Lee. # nowplaying the future billboard: mining music listening behaviors of twitter users for hit song prediction. In *SoMeRA'14, Proceedings of the First International Workshop on Social Media Retrieval and Analysis*, pages 51–56, Gold Coast, Queensland, Australia, 2014.

[33] Noam Koenigstein, Yuval Shavitt, and Noa Zilberman. Predicting billboard success using data-mining in p2p networks. In *IEEE International Symposium on Multimedia, ISM*, pages 465–470, California, USA, 2009.

[34] Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th International Conference on Knowledge Discovery & Data Mining, ACM SIGKDD*, pages 1269–1278, Anchorage, AK, USA, 2019.

[35] Denis Kwiatkowski, Peter CB Phillips, Peter Schmidt, Yongcheol Shin, et al. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1-3):159–178, 1992.

[36] Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[37] Junghyuk Lee and Jong-Seok Lee. Predicting music popularity patterns based on musical complexity and early stage popularity. In *Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia, ACM SLAM*, pages 3–6, Brisbane, Australia, 2015.

[38] Junghyuk Lee and Jong-Seok Lee. Music popularity: Metrics, characteristics, and audio-based prediction. *IEEE Trans. Multimedia*, 20(11):3173–3182, 2018.

[39] Tao Li, Mitsunori Ogihara, and George Tzanetakis. *Music data mining*. CRC Press, Inc., 1st edition, 2011.

[40] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS*, page 4768–4777, Long Beach, California, USA, 2017.

[41] Corey Lynch, Kamelia Aryafar, and Josh Attenberg. Images don't lie: Transferring deep visual semantic features to large-scale multimodal learning to rank. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining, ACM SIGKDD*, pages 541–548, San Francisco, CA, USA, 2016.

[42] Munir Makhmutov, Joseph Alexander Brown, Manuel Mazzara, and Leonard Johard. MOMOS-MT: mobile monophonic system for music transcription: Sheet music generation on mobile devices. In *Proceedings of the Symposium on Applied Computing, SAC*, pages 543–549, Marrakech, Morocco, 2017.

[43] D. Martín-Gutiérrez, G. Hernández Peñaloza, A. Belmonte-Hernández, and F. Álvarez García. A multimodal end-to-end deep learning architecture for music popularity prediction. *IEEE Access*, 8:39361–39374, 2020.

[44] Siobhan McAndrew and Martin Everett. Music as collective invention: A social network analysis of composers. *Cultural Sociology*, 9(1):56–80, 2015.

[45] Cory McKay, Ichiro Fujinaga, and Philippe Depalle. jaudio: A feature extraction library. In *Proceedings of the International Conference on Music Information Retrieval*, pages 600–3, London, UK, 2005.

[46] Kevin P. Murphy. *Machine learning - a probabilistic perspective.* Adaptive computation and machine learning series. MIT Press, 2012.

[47] Y. V. Srinivasa Murthy and Shashidhar G. Koolagudi. Content-based music information retrieval (CB-MIR) and its applications toward the music industry: A review. *ACM Computing Surveys, CSUR*, 51(3):45:1–45:46, 2018.

[48] M. E. J. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, 64:016132, Jun 2001.

[49] Mark E. J. Newman. *Networks: An Introduction.* Oxford University Press, 2010.

[50] Yizhao Ni, Raul Santos-Rodriguez, Matt Mcvicar, and Tijl De Bie. Hit song science once again a science? In *Proceedings of the 4th International Workshop on Machine Learning and Music*, pages 355–360, Sierra Nevada, Spain, 2011.

[51] François Pachet. Hit song science. In Tao Li, Mitsunori Ogihara, and George Tzanetakis, editors, *Music Data Mining*, chapter 10, pages 305–326. CRC Press, 2011.

[52] François Pachet and Pierre Roy. Hit song science is not yet a science. In *Proceedings of the International Conference on Music Information Retrieval, ISMIR*, pages 355–360, Philadelphia, USA, 2008.

[53] Judea Pearl. *Causality: models, reasoning and inference, second edition*, volume 29. Cambridge University Press, 01 2000.

[54] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[55] R Rajyashree, Anmol Anand, Yash Soni, and Harshitaa Mahajan. Predicting hit music using midi features and machine learning. In *2018 3rd International Conference on Communication and Electronics Systems, ICCES*, pages 94–98, Coimbatore, India, 2018.

[56] Jing Ren and Robert J. Kauffman. Understanding music track popularity in a social network. In *25th European Conference on Information Systems, ECIS*, page 25, Guimarães, Portugal, 2017.

[57] Jing Ren, Jialie Shen, and Robert J. Kauffman. What makes a music track popular in online social networks? In *Proceedings of the 25th International Conference Companion on World Wide Web, WWW*, pages 95–96, Montreal, Canada, 2016. International World Wide Web Conferences Steering Committee.

[58] Peter J Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[59] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856, 2006.

[60] Giovanni Scardoni and Carlo Laudanna. Centralities based analysis of complex networks. In *New Frontiers in Graph Theory*, chapter 16. IntechOpen, Rijeka, 2012.

[61] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[62] Burr Settles and Steven Dow. Let's get together: The formation and success of online creative collaborations. In *Conference on Human Factors in Computing Systems, ACM SIGCHI*, pages 2009–2018, Paris, France, 2013.

[63] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[64] Benjamin Shulman, Amit Sharma, and Dan Cosley. Predictability of popularity: Gaps between prediction and understanding. In *Proceedings of the Tenth International Conference on Web and Social Media*, pages 348–357, Cologne, Germany, 2016.

[65] Mariana O. Silva and Mirella M. Moro. Causality Analysis Between Collaboration Profiles and Musical Success. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web, WebMedia*, page 369–376, Rio de Janeiro, Brazil, 2019.

[66] Mariana O. Silva, Anísio Oliveira, Gabriel P. Lacerda, and Mirella M. Moro. Collaboration-aware multimodal hit song prediction. *Future Generation Computer Systems*, 2020. [under review].

[67] Mariana O. Silva, Laís M. Rocha, and Mirella M. Moro. Collaboration Profiles and Their Impact on Musical Success. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, SAC*, pages 2070–2077, Limassol, Cyprus, 2019.

[68] Mariana O. Silva, Laís M. Rocha, and Mirella M. Moro. MusicOSet: An Enhanced Open Dataset for Music Data Mining. In *XXXIV Simpósio Brasileiro de Banco de*

Dados: Dataset Showcase Workshop, SBBD 2019 Companion, Fortaleza, CE, Brazil, 2019.

[69] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441–471, 1987.

[70] Alistair Sutcliffe, Robin Dunbar, Jens Binder, and Holly Arrow. Relationships and the social brain: integrating psychological and evolutionary perspectives. *British journal of psychology*, 103(2):149–168, 2012.

[71] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[72] Hiro Y Toda and Taku Yamamoto. Statistical inference in vector autoregressions with possibly integrated processes. *Journal of econometrics*, 66(1-2):225–250, 1995.

[73] Brian Uzzi and Jarrett Spiro. Collaboration and creativity: The small world problem. *American Journal of Sociology*, 111(2):447–504, 2005.

[74] Andreu Vall, Matthias Dorfer, Markus Schedl, and Gerhard Widmer. A hybrid approach to music playlist continuation based on playlist-song membership. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing, SAC*, pages 1374–1382, Pau, France, 2018.

[75] Gabriel Vigliensoni and Ichiro Fujinaga. Automatic music recommendation systems: Do demographic, profiling, and contextual features improve their performance? In *Proceedings of the International Conference on Music Information Retrieval, ISMIR*, pages 94–100, New York, United States, 2016.

[76] Qi Wang, Mengying Sun, Liang Zhan, Paul Thompson, Shuiwang Ji, and Jiayu Zhou. Multi-modality disease modeling via collective deep matrix factorization. In *Proceedings of the 23rd International Conference on Knowledge Discovery and Data Mining, ACM SIGKDD*, pages 1155–1164, Halifax, NS, Canada, 2017.

[77] Ranran Wang, Xiao Ma, Chi Jiang, Yi Ye, and Yin Zhang. Heterogeneous information network-based music recommendation system in mobile networks. *Comput. Commun.*, 150:429–437, 2020.

[78] Xindi Wang, Burcu Yucesoy, Onur Varol, Tina Eliassi-Rad, and Albert-László Barabási. Success in books: predicting book sales before publication. *EPJ Data Science*, 8(1):31, 2019.

[79] Li-Chia Yang, Szu-Yu Chou, Jen-Yu Liu, Yi-Hsuan Yang, and Yi-An Chen. Revisiting the problem of audio-based hit song prediction using convolutional neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 621–625, New Orleans, LA, USA, 2017.

[80] Haiqing Yu, Yanling Li, Shujun Zhang, and Chunyan Liang. Popularity prediction for artists based on user songs dataset. In *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence, ICCAI*, pages 17–24, Bali, Indonesia, 2019.

[81] Eva Zangerle, Michael Vötter, Ramona Huber, and Yi-Hsuan Yang. Hit song prediction: Leveraging low- and high-level audio features. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR*, pages 319–326, Delft, The Netherlands, 2019.