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

Gabriel Pereira de Oliveira

Analyses of Musical Success based on Time, Genre and Collaboration

Belo Horizonte
2021

Gabriel Pereira de Oliveira

Analyses of Musical Success based on Time, Genre and Collaboration

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
Co-Advisor: Anisio Mendes Lacerda

Belo Horizonte
2021

Oliveira, Gabriel Pereira de.

O48m Analyses of musical success based on time, genre and collaboration [manuscrito] / Gabriel Pereira de Oliveira — 2021.
139 f. il.; 29 cm.

Orientadora: Mirella Moura Moro.
Coorientador: Anisio Mendes Lacerda.
Dissertação (mestrado) - Universidade Federal de Minas Gerais – Departamento de Ciência da Computação
Referências: f. 101-112

1. Computação – Teses. 2. Sistemas de recuperação da informação – Música - Teses. 3. Redes complexas – Teses. 4. Mineração de dados (Computação) – Teses. I. Mirella Moura Moro. II. Anisio Mendes Lacerda. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Computação. IV. Título.

CDU 519.6*85 (043)

Ficha Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6/1510 Universidade Federal de Minas Gerais - ICEx

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

FOLHA DE APROVAÇÃO

Analyses of Musical Success based on Time, Genre and Collaboration

GABRIEL PEREIRA DE OLIVEIRA

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 - Coorientador
Departamento de Ciência da Computação - UFMG

Profa. Renata de Matos Galante
Instituto de Informática - UFRGS

PROF. DENILSON BARBOSA
Department of Computing Science - University of Alberta

PROFA. MICHELE AMARAL BRANDÃO
Instituto Federal de Educação Ciência e Tecnologia de Minas Gerais

Belo Horizonte, 1 de Junho de 2021.

To my parents Mariléia and Gerson, and my sister Nikolle.

Acknowledgments

Agradeço primeiramente a Deus pelo discernimento e por ter me abençoado até aqui. Também agradeço imensamente à minha família. À minha mãe Marileia, pelo amor incondicional e por ser exemplo de força e sabedoria, além de sempre me incentivar a seguir o caminho dos estudos. Ao meu pai Gerson, que me ensinou desde cedo a importância de fazer minhas tarefas com empenho e dedicação. À minha irmã Nikolle, por ser exemplo de resiliência e por todo carinho de sempre.

Gostaria também de agradecer a todas as pessoas que compartilharam um pouco dessa trajetória comigo. Aos meus amigos Felippe Ribas, Gabriela Scarabelli e Luíza Trindade, pelos muitos momentos juntos na UFMG e pelo carinho e cuidado em meio à pressão do mestrado. Agradeço também ao Guilherme Dutra, Hallini Jardim, Larissa Gomes, Laura Parreiras, Lucas Lima, Matheus Marinho e a todas as minhas amizades pelo enorme apoio desde sempre. Amo muito vocês.

Sem dúvidas, minha trajetória não seria a mesma sem os conselhos e ensinamentos valiosos da minha orientadora, Professora Mirella Moro. Muito obrigado por enxergar meu potencial desde a graduação e por nunca me desamparar nesta caminhada. Agradeço também ao meu coorientador, Professor Anisio Lacerda, que sempre foi muito solícito e paciente, e com quem aprendi bastante durante o mestrado.

Eu também não poderia deixar de agradecer aos meus amigos Mariana Silva e Danilo Seufitelli, com quem tive a honra de trabalhar junto, por toda colaboração e aprendizado durante esse período. Com vocês aprendi o poder de se trabalhar em grupo e que fazer pesquisa também pode ser leve e divertido. Este trabalho definitivamente não teria sido o mesmo sem a colaboração de vocês.

Aproveito para agradecer grandemente aos colegas do Laboratório CS+X. Quero deixar um agradecimento muito especial a Natércia Aguilar e Michele Brandão por serem grandes referências e por terem me acompanhado nos primeiros passos na pesquisa, ainda na graduação, sempre com muita disposição em ensinar e ajudar. Agradeço também a Michele Brito pela amizade e pelo suporte no laboratório.

Além disso, minha enorme gratidão às amizades que fiz durante todos esses anos no DCC. Em especial ao Pedro Brum, por ser minha dupla desde o primeiro semestre da graduação e por compartilhar grande parte da trajetória acadêmica. Agradeço também a todas as pessoas com quem trabalhei e que contribuíram para a realização deste trabalho: Bruna Campos, Clarisse Scofield, Gabriel Reis e Iago Domingues.

Por fim, agradeço à UFMG e ao DCC, por terem me proporcionado uma educação

pública de qualidade e vivências que contribuíram muito para o meu crescimento pessoal e profissional. Agradeço também ao CNPq pelo financiamento deste trabalho através de bolsa, permitindo que eu me dedicasse em tempo integral à pesquisa.

Em resumo, agradeço a cada pessoa que compartilhou comigo um pouco dessa trajetória, pois carrego comigo um pouco de cada um. Com isso, espero cada vez mais poder ser uma pessoa e um profissional cada vez melhor, retribuindo para a sociedade as oportunidades que recebi. Mais uma vez, meu muito obrigado.

"*Não se pode falar de educação sem amor.*"
(Paulo Freire)

Resumo

A música é uma das formas culturais mais importantes do mundo, como também uma das mais dinâmicas. Essa natureza dinâmica pode influenciar diretamente a carreira de artistas e refletir em seu sucesso. Neste trabalho, analisamos o sucesso musical através da perspectiva de gêneros musicais. Especificamente, modelamos as linhas do tempo de sucesso de artistas e gêneros para detectar e prever períodos contínuos de maior impacto, i.e., *hot streaks*. À medida em que a colaboração entre artistas se torna uma das principais estratégias para promover novas músicas, nós construímos e caracterizamos redes de colaboração de gêneros baseadas em sucesso para nove mercados em todo o mundo. A partir de tais redes, detectamos perfis de colaboração diretamente relacionados ao sucesso musical. Em seguida, exploramos comportamentos de gênero excepcionais nas redes onde o sucesso se desvia do padrão. Os resultados mostram que o estudo da colaboração entre gêneros é uma maneira poderosa de avaliar o sucesso musical, descrevendo comportamentos semelhantes em músicas colaborativas de várias formas. Ademais, considerar os mercados globais e regionais é fundamental, pois cada país possui sua dinâmica de sucesso e preferências de gêneros. Complementando, a abordagem regional revela padrões locais que moldam o ambiente global. De modo geral, nosso trabalho contribui tanto para a academia quanto para a indústria musical, à medida que investigamos fatores implícitos da ciência por trás do sucesso musical.

Palavras-chave: *hit song science*, recuperação de informações musicais, gêneros musicais, redes complexas, dados, mineração de dados.

Abstract

Music is one of the world's most important cultural forms and one of the most dynamic. Such a dynamic nature can directly influence artists' careers and reflect their success. In this work, we analyze musical success from a genre-oriented perspective. Specifically, we model both artist and genre success timelines to detect and predict continuous periods with higher impact, i.e., *hot streaks*. As artist collaboration becomes one of the main strategies to promote new songs, we build and characterize success-based genre collaboration networks for nine markets worldwide. From such networks, we detect collaboration profiles directly related to musical success. Furthermore, we mine exceptional genre patterns in the networks where the success deviates from the average. Our findings show that studying genre collaboration is a powerful way to assess musical success by describing similar behaviors within collaborative songs from multiple perspectives. In addition, considering both global and regional markets is fundamental, as each country has its success dynamics and genre preferences. Such a regional approach also reveals local patterns that shape the global environment. Overall, our work contributes to both the academy and the music industry, as we shed light on the underlying factors of the science behind musical success.

Keywords: hit song science, music information retrieval, musical genres, complex networks, data science, data mining

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Music is not only one of the world's most important cultural industries, but also one of the most dynamic. Over the last few decades, the world has seen a dramatic change in the way people consume music, moving from physical records to streaming services. Few years ago, songs and their videos needed to be played on the radio and TV to be successful; but today, they can be easily accessed on digital platforms such as Spotify and YouTube. Streaming services combined with social networks have become the main form of disseminating the work of artists, bringing universal access to these contents alongside brand new data about music itself and its social impact. Since 2017, such services have become the main source of revenue within the global recorded music market, mainly due to the fans' engagement and adoption of these platforms. In fact, their revenues more than doubled from then, reaching US$ 13.4 billion as of 2020.[1]

Indeed, the dynamic nature of the music industry can directly influence the behavior of artists' careers. That is, an artist's career can suffer ups and downs depending on the current market moment. At a higher level of abstraction, the same fluctuating behavior happens for musical genres. Figure 1.1 shows the weekly evolution of five popular genres in the United States, measured by the total number of songs featured on the Billboard Hot 100.[2] From the 1960s to the 1980s, *soul* and *rock* genres dominated the music scene, with Stevie Wonder, Aretha Franklin, The Beatles, and Queen being some of the greatest artists of this period. A substantial change in musical genre preferences marked the 1990s, mainly due to technological advances, such as Internet popularization. From then on, *pop* and *rap* conquered space on the charts and became protagonists at the beginning of the 2000s. Britney Spears, Eminem, Beyoncé, and Drake are examples of artists of such genres. Finally, *country* remained stable throughout the period but increased its participation between 2000 and 2016.

As the music industry becomes more complex and competitive, artists are encouraged to reinvent strategies to maintain their presence in the market and reach new audiences. Thus, artist collaboration has grown into one of the main tactics to promote new songs. This widely adopted strategy is a strong force driving music nowadays, main-

[1] IFPI Global Music Report 2021: https://gmr.ifpi.org/

[2] The Billboard Hot 100 is the main weekly song chart within the United States. A song's position in the chart is calculated by considering sales, radio plays, and streaming count.

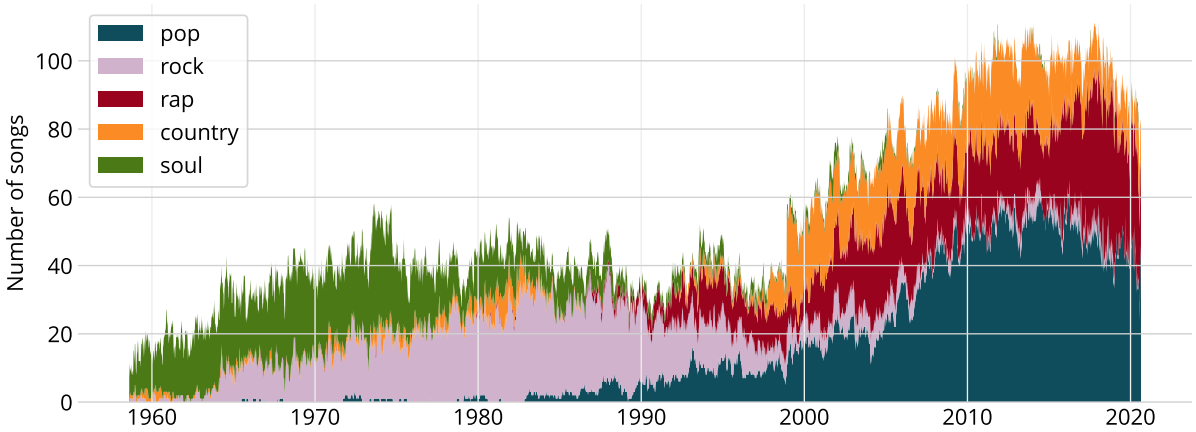


Figure 1.1: Evolution of popular genres in the United States, measured by the total number of songs featured on the weekly Billboard Hot 100 Chart (1958 - 2020).

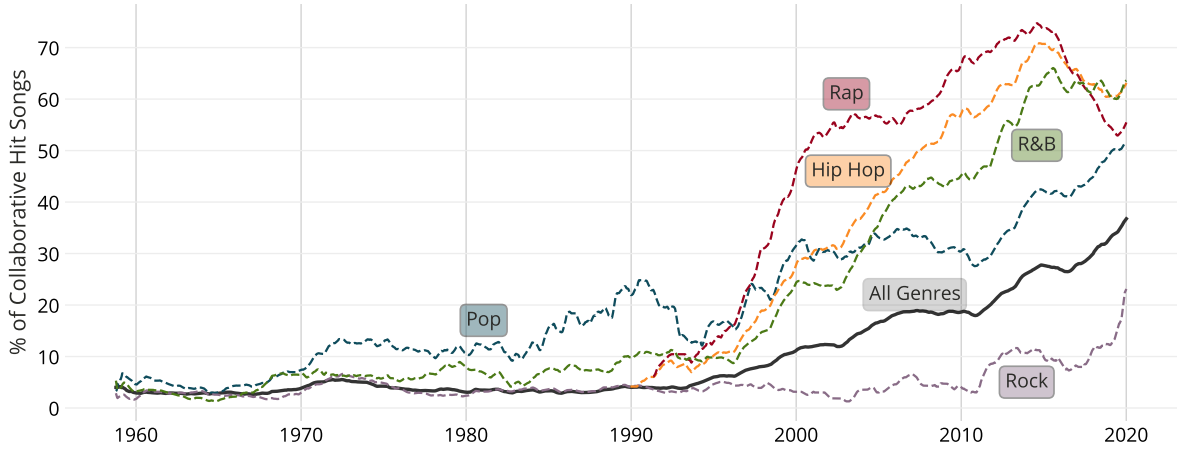


Figure 1.2: Historical frequency of collaborative hit songs for selected genres on Billboard Hot 100 Chart (1958 - 2020).

taining artists' relevance in the market. Such connections usually help artists bridge the gap between styles and genres, overlapping new fan bases and consequently increasing their numbers. In such a way, several studies approach the factors behind musical success, creating an emerging field within computer science called Hit Song Science (HSS). Collaboration-aware studies then become promising, as successful artists are more likely to have a high degree of collaboration in success-based networks [100]. In fact, there is strong evidence that factors leading to an ideal musical partnership can be understood by exploring collaboration patterns that directly impact its success [21].

The genre perspective is very important when analyzing the impact of collaborations in musical success, as each genre has a distinct audience that behaves in its own way. Figure 1.2 shows this phenomenon and highlights the growing trend in the number of collaborations within Billboard Hot 100 Charts. Although the general curve increases over time, genres such as *pop* and *R&B* present a collaboration rate higher than others (e.g., *rock*). This contrast can be explained by the intrinsic nature of each music genre. For

instance, *pop* and *R&B* artists frequently collaborate with the *rap* community, mainly as featured artists. Also, partnerships involving *pop* music may take place not only through intra-genre collaborations but also through inter-genres, bringing an additional dimension to their songs.

For example, in April 2019, the collaboration between the American *pop* singer Halsey and the *k-pop* group BTS in the song *Boy With Luv* became the most viewed YouTube music video in 24 hours and reached #8 on Billboard Hot 100 Chart. This and other collaborations with other American artists increased BTS' popularity in the United States and paved the way for the South Korean group to win their first #1 on Hot 100 with the single *Dynamite* in August 2020. The success achieved by BTS also shed light on other *k-pop* acts that became widely popular in the US and other Western countries. In 2020, the girl group Blackpink appeared on the Hot 100 with four different singles, being two collaborations (*Sour Candy* with Lady Gaga and *Ice Cream* with Selena Gomez) and two solos (*How You Like That* and *Lovesick Girls*).

The *k-pop* popularization is only an example of the power of regional genres in the music industry. In the past few years, the collaborations between *pop* and *reggaeton* artists have become more frequent and successful, mostly due to the stardom of the hit *Despacito* by Luis Fonsi and Daddy Yankee in 2017. This song gained a remix with the Canadian pop singer Justin Bieber, reaching the #1 position in the Hot 100 for 16 consecutive weeks. Therefore, record companies are now working to develop local music ecosystems to promote regional cultures across the world. The advance of the Internet and the continuous dissemination of streaming services provide a global platform for artists to engage with their fans. Thus, local genres that were once popular in specific contexts are now globally consumed. Moreover, research on musical success must be aware that local engagement shapes the global music environment, expanding the analysis for markets other than the United States, which is the biggest music market in the world, but not the only one. Hence, as this creative industry changes, it becomes more unpredictable; and doing both predictive and diagnostic analyses in such a context remains challenging.

1.1 Research Goals

Remaining an industry of creative growth, it is only natural for music (i.e., all musical scene members) adapting to new conditions and redefining its layout. Not surprisingly, the Grammy[3] categories were tightened (from 109 to 78, in 2012) as a result of music's dynamic nature. Categories and genres are constantly changing, but they remain

[3]Grammy Awards: https://en.wikipedia.org/wiki/Grammy_Award

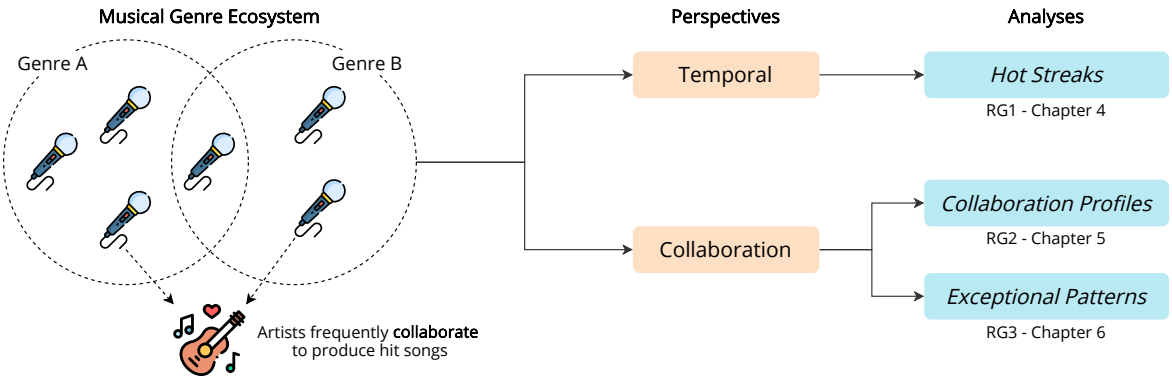


Figure 1.3: Analyses conducted in this work, according to the Research Goals (RGs).

relevant to comprehend the music context. In addition, as the collaboration phenomenon becomes stronger over the years, it is necessary to explore all factors that make it so relevant nowadays. Therefore, this work aims to *analyze artist collaboration under a genre perspective to better understand how the genre connections impact musical success.* We do so by exploring the musical genre ecosystem in temporal and collaboration perspectives (Figure 1.3). Therefore, we assess such an objective through three Research Goals (RGs):

RG1. Understand the temporal evolution of both artist and genre careers, by identifying and predicting periods of high impact in such careers (i.e., hot streaks);

RG2. Analyze the dynamics of cross-genre connections by detecting collaboration profiles in success-based networks (i.e., connections formed by genres of artists who cooperate and create hit songs);

RG3. Mine frequent genre patterns within hit songs in recent years, i.e., investigating the relationship between combining different music genres and musical success.

1.2 Main Contributions

Overall, the main contributions of this work on the relation between musical genres and success are described as follows. The topics are organized according to the Research Goal (RG) they are related to.

RG1. Hot Streaks in Musical Careers (Chapter 4).

1. Based on data from the Billboard Hot 100, we model the time series for both artists and successful genres. From these series, we found that the most successful weeks are grouped in time;

2. We detect hot streaks in musical careers using Piecewise Aggregate Approximation (PAA), which is a method for reducing the dimensionality of time series;

3. Through characterization analysis, we reveal general and specific patterns of hot streaks for artists of different genres. We evaluated characteristics such as quantity, duration, and appearance of the first hot streak periods;

4. We find that artists have more songs on the charts during periods of hot streaks and that the career peaks for artists appear and disappear gradually over time;

5. We assess the hot streak prediction problem as a binary classification task, and our findings reveal that our model was successful in anticipating successful periods for popular music genres;

6. We detect that factors including the number of songs present in the charts and the artists' career time are relevant to increase the predictive power of our model, as well as acoustic features such as *time signature* and *energy*.

RG2. Collaboration Profiles in Genre Networks (Chapter 5).

1. We collect and build a unique dataset on musical success in global and regional markets. We focus on genre collaboration, but we also provide meaningful information about charts, songs, and artists;

2. We also build a success-based genre collaboration network for each considered market, by connecting genres from artists who team up to make hit songs;

3. We find that individually analyzing regional markets is fundamental, as local genres play a key role on determining hit songs and popular artists;

4. In general, our results reveal that genre collaborations are increasing, with emerging local genres hitting global success – despite the differences in the evolution of regional markets;

5. Our network-based analysis on genre collaborations describe three significant factors (*Attractiveness, Affinity* and *Influence*) that uncover four collaboration profiles (*Solid, Regular, Bridge* and *Emerging*) directly related to musical success.

RG3. Exceptional Genre Patterns on Hit Songs (Chapter 6).

1. We use data mining techniques to reveal frequent genre patterns in songs that made to the charts in each market;

2. We confirm that there are significant differences in the behavior of each market, with regional genres playing an important role in patterns;

3. We find that each regional market has specific patterns of genre connections in which success is above average;

4. Our experiments reveal that association rules can be an important tool to identify and recommend promising musical genres collaborations.

1.3 Text Organization

The rest of this work is organized as follows. We present and discuss the related work in Chapter 2. Then, Chapter 3 contains the background and fundamental concepts required to understand this work. In Chapter 4, we analyze the temporal evolution of musical careers to detect and predict hot streak periods. Next, in Chapter 5 we build success-based genre networks and identify collaboration profiles within it. Chapter 6 presents a data mining approach to mine exceptional genre collaboration patterns using such networks and profiles. Finally, in Chapter 7 we present our concluding remarks and discuss future work on this subject.

Chapter 2

Related Work

Besides helping society to evolve, Computer Science has also directly influenced the development of most sciences and industries. One clear example is Entertainment, which has evolved from films and music on tapes and long players to digital media [20, 36]. Specifically, when it comes to music content, Music Information Retrieval (MIR) emerges as an interdisciplinary research field based on musicology, psychology, and computer science to extract meaningful information from musical content [63, 72].

In this chapter, we provide an overview of the main research topics within MIR and other subjects that are related to this work. Specifically, we divide such related work into five sections. First, we outline the main data sources used in music-related studies (Section 2.1). In Section 2.2, we describe Hit Song Science, which aims to predict the success of a song before its release. Then, we briefly review studies regarding the use of the genre information in MIR (Section 2.3) and hot streaks in professional careers (Section 2.4). Next, we discuss research regarding collaboration and its relation with success (Section 2.5). Finally, we present our final considerations by emphasizing the relevance of this study in face of such related work (Section 2.6).

2.1 Music Data Sources

The first step of most music-oriented studies is to gather data regarding song characteristics. However, such features can be seen by many facets, as the definitions are open to different visions. For instance, data about a given song can be acoustic and/or lyric-based, while its popularity may be measured considering its position within a chart or its sales revenue. Besides, information concerning consumers' behavior may be aggregated to the analyses to enhance the results. Therefore, using data from multiple sources is necessary and useful to build better models for analyzing and predicting musical success. In this section, we describe and classify the main and most commonly used data sources in four categories according to their purpose: popularity; acoustic characteristics;

lyrics; and social behavior.

Regarding song **popularity**, research studies usually consider information such as position in charts to determine the level of success. The US-based magazine Billboard[1] is the most consolidated data source, providing many different types of rankings since the 1940s. The Hot 100 is the most commonly used, as it is a weekly list of the 100 most popular songs (regardless of music genre or style) in the US, considering data from radio airplay, sales, and streaming activity [18, 9, 100, 89, 107]. Billboard also aggregates the weekly rankings in a Year-End Hot 100 Chart, which is used in some music-related studies [102, 103]. However, there are several studies considering other specific Billboard charts in their analyses. For example, Chon et al. [25] focus on one specific genre by using the Top Jazz Chart, based only on the albums' sales. Also, Lee and Lee [60] obtain data from The Rock Songs Chart, a weekly list of the 50 most popular rock songs. Such authors believe that this choice may produce cleaner results and better insights when focusing on specific genres.

As the world becomes more connected, and the globalization process reaches most of the countries, local engagement shapes the global music environment. In such a way, some studies consider charts from outside the US in their analyses and predictions. The United Kingdom is the second most considered market, having its charts published by the Official Charts Company[2] (OCC) [75, 53]. Besides using British charts, Fan and Casey [34] also collect Chinese hit songs for comparison purposes. Moreover, there are studies considering other European countries (e.g., France, Belgium and Germany) [22, 49] and Asian markets such as South Korea [97] and Indonesia [35]. Other popularity approaches use YouTube views and likes [24] and sales data provided by platforms such as Amazon [1] and Nielsen SoundScan [9].

Now, changing the subject to features, **acoustic characteristics** of a song are important tools for describing its structure. Besides being better discussed in Section 2.2, it is important to note that they have been largely used since early music-oriented research studies, such as Dhanaraj and Logan [31], which use in-house databases as their data source. With the evolution of the Music Information Retrieval (MIR) field, new sources take place, as the EchoNest API, with more than a trillion data points on over 34 million songs in its database [50]. Several studies use this API for extracting features such as tempo, time signature, song duration, and loudness [75, 103, 49]. Nonetheless, with the expansion of music streaming services and the acquisition of EchoNest by Spotify in 2014, its Developer API[3] is now the main source of acoustic features, thus being used by most recent studies [68, 4, 69, 89]. Nonetheless, there are still other sources used, such as

[1]Billboard Charts: http://www.billboard.com/charts
[2]Official Charts Company: http://www.officialcharts.com/charts
[3]Spotify Developer API: http://developer.spotify.com/documentation/web-api

the Million Song Dataset[4] (MSD) [116] and AcousticBrainz[5] [53].

Data frequently used within music-related research also include song **lyrics**, mainly for generating features related to rhyme and text. From the early years of HSS, there is no consensus on the best and most reliable source for song lyrics, and each study considers a different lyric source. For example, websites such as Astraweb Lyric Search [31], MetroLyrics [103], MusicSongLyrics [92], and LyricsMania [24] are used as lyrics sources by several authors. However, more recent studies, such as Martín-Gutiérrez et al. [68], use Genius, which has an exclusive API for developers to collect data in a simple and fast way, with no need to use web crawlers or HTML pages.

Finally, a different kind of data source has recently emerged: **social media**, which is changing the way people share their opinions and impacting several areas, including the music industry. Therefore, the consumers' behavior plays a key role in musical success analysis, and online platforms such as Last.fm[6] are largely used to collect listener-based data and features [18, 49, 91]. Moreover, blogging platforms are also important sources of information about people's feelings on a given song, album, or artist. For example, Abel et al. [1] use Spinn3r[7] to collect more than 100 million blog posts in the music domain. More recent studies collect data from social networks such as Twitter, Instagram, and Facebook to analyze users' behavior related to a new musical release [7, 27, 107].

2.2 Hit Song Science

Hit Song Science (HSS) is an emerging field within MIR that tackles the problem of predicting the popularity of a given song. It involves acquiring and analyzing musical data from distinct data sources to study the relation between the intrinsic features of songs and their success. Reinforcing its multidisciplinary characteristic, studies in HSS combine Machine Learning and Data Mining techniques with musicology and psychology concepts to verify whether popular songs share similar feature patterns. Besides, the dynamic and volatile nature of the musical scenario converts HSS into a strategic research area, as predicting the popularity of songs and artists provides benefits for all involved parties in the global music industry. In fact, HSS has become a trending research topic in academia. An evidence for such a claim is the increasing volume in publications[8] about it, as illustrated in Figure 2.1.

[4]Million Song Dataset: http://millionsongdataset.com/
[5]AcousticBrainz: http://acousticbrainz.org/
[6]Last.fm API: http://www.last.fm/api/
[7]Spinn3r: http://docs.spinn3r.com/
[8]The considered HSS publications are further listed in Table 2.2.

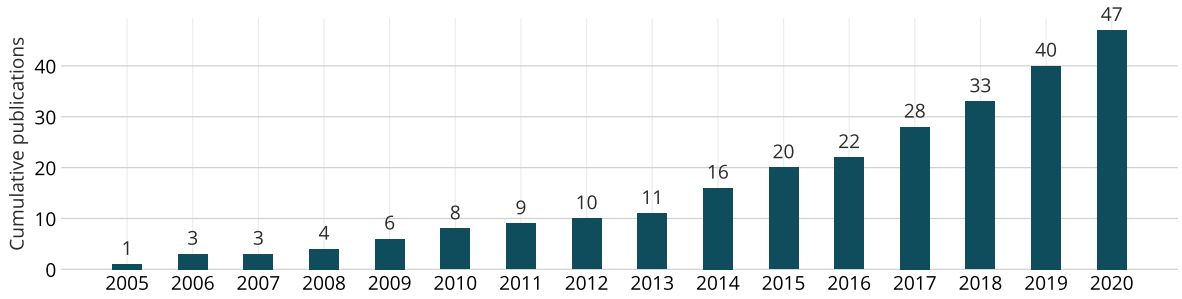


Figure 2.1: Hit Song Science publications (cumulative), 2005 – 2020.

The concept of *Hit Song Science* was first introduced in 2003 by Polyphonic HMI,[9] an artificial intelligence company focused on using mathematics and computer science to solve problems in the music industry. Such a company developed a commercial tool to predict, on a scale from 1 to 10, the success of a song in the current market based on its chart position. Such an achievement motivated researchers in the MIR community to develop the first scientific studies regarding hit song prediction. For instance, Dhanaraj and Logan [31] use acoustic and lyric-based features in a classification model to provide the first evidence that there is indeed a pattern connecting hit songs, as their model performs slightly better than random.

Following such a study, early works in HSS used mostly song-related features to predict popularity with distinct approaches. For instance, Chon et al. [25] analyze data at an album scale and find that the higher the starting position of an album is, the longer it is likely to stay on musical charts. However, Pachet and Roy [84] point out that the features commonly used at the time might not be enough to reveal relevant information about musical success. Their study considers audio and human-generated features and defines hit song prediction as a classification task, but it did not achieve significant results.

The advance of machine learning algorithms and the discovery of possible new features helped to overcome such obstacles. The emergence of blogs and online social networks in the early 2000s paved the way for improving prediction models. In HSS, Salganik et al. [94] are the first to study the impact of social influence on determining song popularity, revealing that its presence increases both inequality and unpredictability of success. In contrast, Abel et al. [1] use blogging behavior to predict music sales performance. In recent years, features extracted from social platforms such as Twitter, Facebook, and Instagram are also considered in prediction models [7, 27, 107].

Nonetheless, acoustic features are still used in most studies in HSS, as they act as descriptors of the core elements of a song, which include: pitch (melody and harmony), rhythm, dynamics, and the qualities of timbre and texture. However, the main sources of such features have changed over time from the early in-house datasets [31, 84] to data

[9]Polyphonic HMI, Hit Song Science: http://bit.ly/polyphonic-hmi

Table 2.1: Main acoustic features obtained from Spotify.

Feature	Description	Type	Value Range
acousticness	The probability of a song to be acoustic or not	Float	$[0, 1]$
danceability	Informs whether a song is suitable for dancing or not in terms of probability	Float	$[0, 1]$
duration_ms	The duration of a song in milliseconds	Integer	$[0, \text{inf})$
energy	The intensity and activity of a song considering information such as dynamic range, perceived loudness, timbre, onset rate, and general entropy	Float	$[0, 1]$
instrumentalness	The probability of a song to be instrumental, i.e., without vocals	Float	$[0, 1]$
key	The estimated overall key of a song, mapped as an integer number (e.g., $C = 0$, $C\# = 1$, and so on)	Integer	$[0, 11]$
liveness	The probability of a song being performed live, i.e., the presence of an audience in a song	Float	$[0, 1]$
loudness	The general loudness measured in decibels (dB)	Float	Typically $[-60, 0]$
mode	The general modality of a song (i.e., major= 1 or minor= 0)	Integer	$[0, 1]$
speechiness	The probability of a given song to have spoken words in it	Float	$[0, 1]$
tempo	The speed of the song, measured in beats per minute (BPM)	Float	N/A
time_signature	The amount of beats in each bar (measure)	Integer	N/A
valence	The positiveness of a song, in which high valence values represent happier songs, whereas low values means the opposite	Float	$[0, 1]$

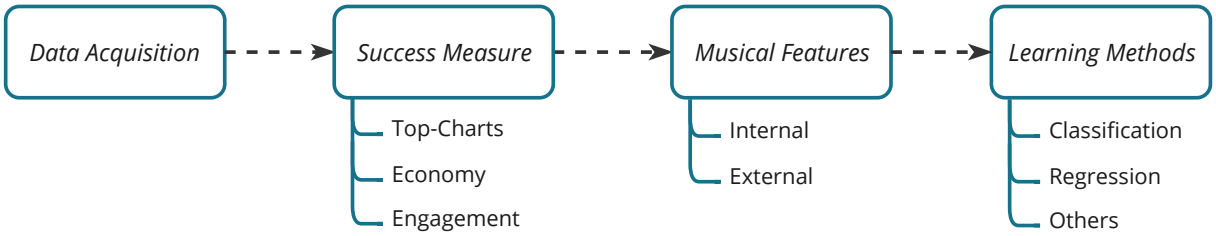


Figure 2.2: Generic workflow for the Hit Song Prediction problem.

extracted from digital streaming [27, 69]. In the latter category, Spotify[10] was founded in 2008, and today it is the world's most popular audio streaming service, with over 70 million songs (as of May 2021). Its Web API[11] provides detailed information about the tracks, including specific audio features detailed in Table 2.1. The ease in the data collection process and the availability of data from several markets make Spotify one of the popular sources for acoustic features [27, 68, 4, 69].

Overall, most of the studies in Hit Song Science follow a common sequence of steps, from data collection to model selection. Therefore, we generalize such steps into a generic workflow for the *Hit Song Prediction* problem, as shown in Figure 2.2. Here, we provide a brief description of each phase, as detailed next.

[10]Spotify: https://www.spotify.com/
[11]Spotify for Developers: https://developer.spotify.com/documentation/web-api/

Data Acquisition. Songs are complex and dynamic objects that can be analyzed in different ways, and Hit Song Science (HSS) emerges as a field where studies try to use many of these facets in their models. Therefore, a convenient approach is to collect information about songs through multiple sources to complement audio-based musical success prediction, as previously shown in Section 2.1.

Success Measure. Musical success is usually associated with fame and power. However, in science, defining and measuring success remains a topic of great interest not only for the music industry but also for the MIR community. There are different measures of musical success in the HSS literature, but we can generalize them into three distinct classes: *Top-charts*, in which success is defined based on the song position in musical charts; *Economy*, relating success to economic indicators, including sales revenue; and *Engagement*, which includes the social interactions with musical content.

Musical Features. The success of a given song may be associated with a set of characteristics from the musical scenario. Such features are used to describe songs in several aspects, serving as input for learning models. In HSS, there are two main groups of features: *Internal*, which are directly extracted from the audio (i.e., acoustic fingerprints, lyrics, and metadata); and *External*, obtained from agents or objects that may influence the musical success (i.e., artist popularity, album sales, and streams).

Learning Methods. One of the main goals of Hit Song Science is to discover the set of predictors that contribute to the success of a song. In general, most works use machine learning on such a task. Thus, we list the main learning approaches used within HSS studies: *Classification*, in which the goal is to predict labels for a set of instances; *Regression*, in which the prediction output is a continuous value; and *Other* methods, such as clustering and statistical analysis.

Table 2.2 summarizes the existing research in Hit Song Science according to the predefined phases of our workflow: data sources, success perspective, considered features, and machine learning tasks. Its last line is a contribution of this dissertation, as presented in Chapter 5.

Table 2.2: Research works in Hit Song Science, with corresponding data sources, success perspectives, considered features, and the machine learning task.

Year	Reference	Data Sources	Success	Features	ML Task
2005	Dhanaraj and Logan [31]	Oz Net Music Chart, In-house database, Astraweb Lyrics	Top-Charts	Internal	Classification
2006	Chon et al. [25]	Billboard	Top-Charts	External	Other
2006	Salganik et al. [94]	purevolume	Engagement	External	Other
2008	Pachet and Roy [84]	HiFind Database	Top-Charts	Internal, External	Classification
2009	Bischoff et al. [18]	Last.fm, Billboard	Top-Charts	Internal, External	Classification

Continued on next page

Table 2.2: (continued from previous page)

Year	Reference	Data Sources	Success	Features	ML Task
2009	Koenigstein et al. [58]	Gnutella, Billboard	Top-Charts	External	Classification, Regression
2010	Abel et al. [1]	Spinn3r, Amazon	Economy	Internal, External	Classification, Regression
2010	Yoo and Kim [113]	N/A	Top-Charts	External	Other
2011	Ni et al. [75]	OCC, EchoNest	Top-Charts	Internal	Classification
2012	Berns and Moore [17]	MySpace, Nielsen SoundScan	Economy	External	Other
2013	Fan and Casey [34]	OCC, ZhongGuoGeQuPaiHangBang (China), EchoNest	Top-Charts	Internal	Regression
2014	Dewan and Ramaprasad [30]	Nielsen SoundScan, Google Blog Search, Last.fm, Amazon, Allmusic.com	Economy, Engagement	External	Other
2014	Herremans et al. [50]	OCC, Billboard, EchoNest	Top-Charts	Internal, External	Classification
2014	Kim et al. [56]	Twitter, Billboard	Top-Charts	External	Classification, Regression
2014	Nunes and Ordanini [76]	Billboard	Top-Charts	Internal	Classification
2014	Singhi and Brown [102]	Billboard	Top-Charts	Internal	Classification
2015	Buda and Jarynowski [22]	European Music Papers, radio, TV, Internet	Top-Charts	External	Other
2015	Frieler et al. [37]	Earwormery Database, Polyhex UK, Geerdes Database	Top-Charts	Internal	Classification
2015	Lee and Lee [60]	Billboard	Top-Charts	Internal, External	Classification
2015	Singhi and Brown [103]	EchoNest, Billboard, Metro Lyrics	Top-Charts	Internal	Classification
2016	Ren et al. [92]	Last.fm, Wikipedia, 7digital, Google Lyrics, MusicSong Lyrics	Engagement	Internal, External	Classification
2016	Shulman et al. [98]	Last.fm	Engagement	External	Classification
2017	Araujo et al. [7]	Twitter, Spotify, Billboard	Economy, Engagement	External	Regression
2017	Askin and Mauskapf [9]	Billboard, Discogs, Echo Nest, SoundScan	Top-Charts	Internal, External	Regression
2017	Chiru and Popescu [24]	YouTube	Engagement	Internal	Regression
2017	Herremans and Bergmans [49]	The Ultrapop 50, Last.fm, EchoNest	Top-Charts	Internal, External	Classification
2017	Ren and Kauffman [91]	Last.fm	Top-Charts	Internal, External	Classification, Regression
2017	Yang et al. [112]	KKBOX	Engagement	External	Regression
2018	Febirautami et al. [35]	Spotify	Engagement	Internal	Classification
2018	Interiano et al. [53]	OCC, MusicBrainz, InternalBrainz	Top-Charts	Internal	Classification
2018	Lee and Lee [61]	Billboard	Internal	Classification	
2018	Rajyashree et al. [88]	Million Song Dataset	Engagement	Internal	Classification
2018	Shin and Park [97]	Gaon Music Charts	Top-Charts	Internal, External	Other
2019	Araujo et al. [5]	Spotify	Top-Charts	Internal	Classification
2019	Cosimato et al. [27]	Billboard, iTunes, Spotify, Twitter, Instagram, Facebook, YouTube, Newspapers	Top-Charts	External	Classification
2019	Middlebrook and Sheik [70]	Spotify, Billboard	Top-Charts	Internal, External	Classification

Continued on next page

Table 2.2: (continued from previous page)

Year	Reference	Data Sources	Success	Features	ML Task
2019	Silva and Moro [99]	MusicOSet	Top-Charts	External	Other
2019	Silva et al. [100]	MusicOSet	Engagement	External	Other
2019	Yu et al. [114]	N/A	Engagement	Internal, External	Regression
2019	Zangerle et al. [116]	Million Song Dataset, Billboard	Top-Charts	Internal	Regression
2020	Al-Beitawi et al. [4]	Spotify	Top-Charts	Internal	Other
2020	Araujo et al. [6]	Spotify	Top-Charts	Internal	Classification
2020	Martín-Gutiérrez et al. [68]	SpotGenTrack	Engagement	Internal, External	Classification, Regression
2020	Matsumoto et al. [69]	Spotify	Engagement	Internal, External	Classification, Regression
2020	Raza and Nanath [89]	Billboard	Top-Charts	Internal	Classification
2020	Tsiara and Tjortjis [107]	Twitter, Billboard	Top-Charts	External	Classification, Regression
2020	**Oliveira et al. [79]**	**Spotify**	**Top-Charts**	**External**	**Other**

2.3 Genres in Music Information Retrieval

The musical genre is one of the most prominent high-level music descriptors, and it is fundamental within the musical scenario by aggregating songs that share common characteristics. Hence, they are frequently used in MIR to extract relevant information from music content as several tasks are genre-dependent or directly related to them. For instance, classification is regularly the first step in many MIR applications, thus being one of the core tasks in such a field. Indeed, the genre classification task (i.e., categorizing songs into different genres) has become widely studied in recent years [8, 109, 106]. Going further, Oramas et al. [83] use three distinct modalities (audio, text, and images) to categorize musical items into multiple labels, and Ghosal and Sarkar [40] apply deep learning techniques to enhance classification models, achieving an accuracy of 95.4%.

Although genre provides one of the most convenient categorizations of music and is widely used in music science, it does not necessarily mean that genre is easily categorized or recognized. In this sense, Prockup et al. [85] provide evidence that music genres can be modeled through a combination of several musical attributes. Nonetheless, there are also genre-aware studies assessing other MIR tasks, such as music source separation [59], genre preferences [13], disambiguation/translation [48, 32], new datasets [19] and ontologies [95]. Network science, one of the core topics of our methodology, has also been used to model genres into influence networks [21] and song communities [26].

Understanding musical aspects can be genre-dependent, and this also reflects in the musical success. Therefore, several studies in Hit Song Science (see Section 2.2) use

genre information in their models. For example, Shin and Park [97] consider genres to understand the life trajectory of songs in Gaon Chart,[12] one of the main Korean music rankings. Regarding prediction models, Ren and Kauffman [91] aggregate genres in a musical construct vector (MCV) to summarize the acoustic content of a song. Such MCVs are used as features in a regression model to estimate and predict the popularity duration of a track from a top-charts perspective.

Furthermore, predictive models also use genre as a high-level feature to complement the song description with abstract concepts. Interiano et al. [53] aggregate genres into more general classes (i.e., clusters) to assess success dynamics in UK charts. Besides, Zangerle et al. [116] combine such information with low-level features to produce a more holistic description of songs. Overall, there is strong evidence that music genre may influence musical success, and such information leads to improving the performance of success prediction models [1, 9].

2.4 Hot Streaks

Evaluating the impact of human performance is a common practice in many disciplines, and the term *hot streak* emerges to refer to a specific period within professional careers when success is significantly higher than the average. Such a phenomenon is noted by regular people and widely studied, mainly in the sports field [41, 14, 86]. For example, one may call a hot streak when a team wins several tournaments in a row or even when a specific player performs much better than expected. Although there is a discussion on the empirical nature of such phases in sports [10], the idea of hot streaks is present in studies in other fields, such as gambling [11, 87, 104] and financial markets [47].

Research assessing the impact on individual and creative careers is much more recent. Sinatra et al. [101] aim to uncover temporal patterns in scientific careers, concluding that the most impactful work is randomly distributed over scientists' body of work. They also propose a stochastic model (Q-model) to describe success based on productivity, individual effort, and luck. Following such findings, Liu et al. [65] consider large-scale careers of artists, film directors, and scientists to demonstrate that hot streaks are remarkably universal across diverse domains, yet usually unique across different careers. They also find that impactful works show a high degree of temporal regularity, which can indeed be described by a hot streak model.

Such a result is also demonstrated in the social media domain, which has become extremely popular with the advance of the Internet across the world. In this sense,

[12] Gaon Chart: http://gaonchart.co.kr/

Garimella and West [38] use data from Twitter, one of the most popular online social networks,[13] and define users' impact as the reach of their content. Specifically, they consider several specific success metrics, such as the number of views, likes, retweets, and shares. Their findings point the existence of hot streaks within success phases. Besides, they demonstrate that such phases are driven by new retweeters who suddenly start following and retweeting a user, thus leading to an increase in the follower count.

Janosov et al. [54] also consider luck as a crucial ingredient to achieve impact in creative domains. Therefore, they analyze data from science, art, movie, and music fields to apply the aforementioned Q-model to quantify luck in such contexts. Regarding music, they model the historical artist timelines based on the release year of songs and measure success by the total play counts obtained from Last.fm. Besides showing luck is generally more relevant to the impact of a song, some genres are less influenced by randomness than others. Nonetheless, such a study does not investigate the clustered effect of a set of songs on musical careers (i.e., hot streaks), and thus further investigation is required for such a domain.

2.5 Collaboration and Success

As the world becomes more interconnected with the Internet and other technological advances, people are frequently in touch with their peers, thus reducing the previously existing distance and communication barriers. Therefore, general content production has become increasingly collaborative. Following the popular saying *"Many hands make light work"*, the impact of collaboration on content popularity has become a hot research topic. For instance, content created by social platform users is the subject of studies that apply several techniques ranging from network science [111] to deep learning with neural networks [74].

In general, collaboration occurs when two or more actors participate in the creation, execution, and/or production of the same object. Figure 2.3 presents a generic example where four individuals A_1, A_2, A_3 and A_4 participate in the production of three distinct contents C_1, C_2 and C_3. Such a scenario can be modeled using graph theory and network science, primarily as a bipartite graph (left), with edges connecting actors to the objects they produce. Thus, it is possible to visualize which individuals participate in the production of each content. However, as the bipartite graph does not allow easy analysis of the collaboration between actors, it is necessary to project it into a new network

[13] About Twitter: https://about.twitter.com/

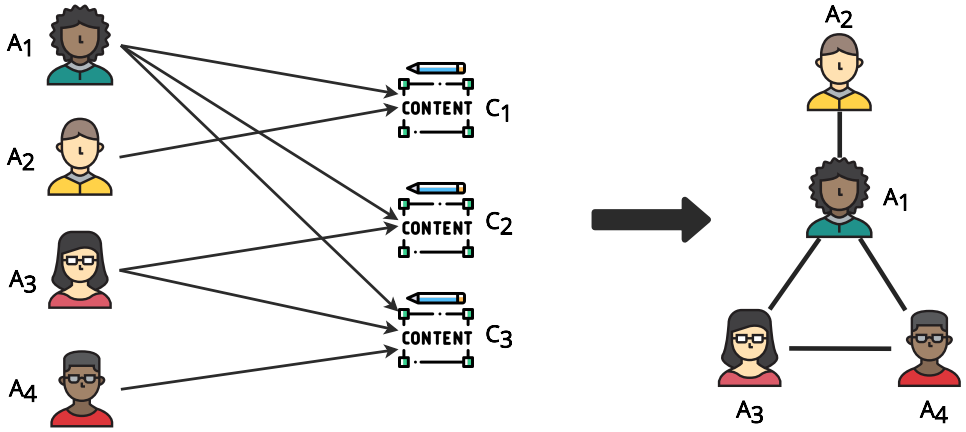


Figure 2.3: Collaboration between distinct actors on content creation.

(right) connecting actors who collaborate at least once. Such a modeling is used in diverse domains, including article editing [29], movie production [110], and social coding [16].

Regarding music content, Bischoff et al. [18] predict the potential of tracks for becoming hits by analyzing the relationships between tracks, artists, and albums. Moreover, Silva et al. [100] address collaboration as a key factor in success, using topological properties to detect relevant profiles in artist networks. In a later study, the causality between collaboration and success is addressed [99], reinforcing the relevance of the collaboration phenomenon in the musical scenario. In fact, such an approach is novel and promising in HSS, but it is restricted to the artist and song levels. In addition, these and most of the aforementioned studies regarding musical success only consider data from American charts, mainly Billboard Hot 100. This may be due to the ease of obtaining data but it may not reflect the whole global scenario, as each country has its own distinct behavior when consuming music, which includes preferred artists and genres.

2.6 Overall Considerations

As the music industry becomes more complex and competitive, developing strategies to maximize the expected musical success becomes increasingly relevant. Therefore, one of the contributions of this work is to introduce a model to describe artists' and genres' success timelines based on music charts. For each musical career, we focus on identifying and characterizing periods with success above the average (i.e., hot streaks). Our model is customized for each timeline since success remains a relative concept. Hence, we can distinguish success between independent artists and music superstars and also between different music genres. The genre-aware model may help the understand the underlying

factors that lead to success, as well as guide artists and record labels to better plan and manage their future single and album releases.

In addition, studying collaboration from a genre perspective may reveal important information on how artists from different communities team up to make a new hit song. To the best of our knowledge, we are the first to build a success-based genre network, investigating collaboration profiles over time and mining exceptional patterns within it, going deeper into the potential intrinsic factors that make up a successful collaboration. Likewise, our approach considering several regional markets makes this work more realistic, as local engagement shapes the global environment. We combine a precise heterogeneous data collection with proper modeling to enhance further data analysis by scientists and record labels CEOs.

Moreover, this work follows the Hit Song Science (HSS) workflow from Section 2.2. In short, we use data collected from Billboard and Spotify and measure success from a Top-chart perspective. In our analyses, we consider both internal and external features, including song, artist, and genre characteristics. We then use such features as the input of classification (Chapter 4), profiling (Chapter 5), and data mining (Chapter 6) tasks. Therefore, we shed light on the science behind the collaboration phenomenon, providing new knowledge to both the academic community and the music industry.

Chapter 3

Background and Fundamental Concepts

This chapter provides the fundamental concepts necessary to understand the challenges and solutions addressed in this work. First, Section 3.1 summarizes key concepts of basic statistics and machine learning used throughout this work. Then, Section 3.2 overviews concepts from Network Science, which uses networks (commonly modeled as graphs) to represent complex systems, such as telecommunication systems, biological structures, and social connections. Finally, Section 3.3 describes a brief background on Data Mining, which aims to extract meaningful knowledge from data. Here, we focus on the main concepts necessary to understand the next chapters and recommend specific references for more advanced ones [15, 73, 96, 115].

3.1 Statistics and Machine Learning

Statistics and learning techniques are important tools within any data science or data-driven research framework. In fact, machine learning may be defined as a subfield of Computer Science that aims to detect patterns in data based on statistical models, using such findings to predict future information [71]. In this section, we present an overview of the main concepts and methods used in this work. We focus on correlation analysis (Section 3.1.1), classification algorithms (Section 3.1.2), and model performance assessment (Section 3.1.3).

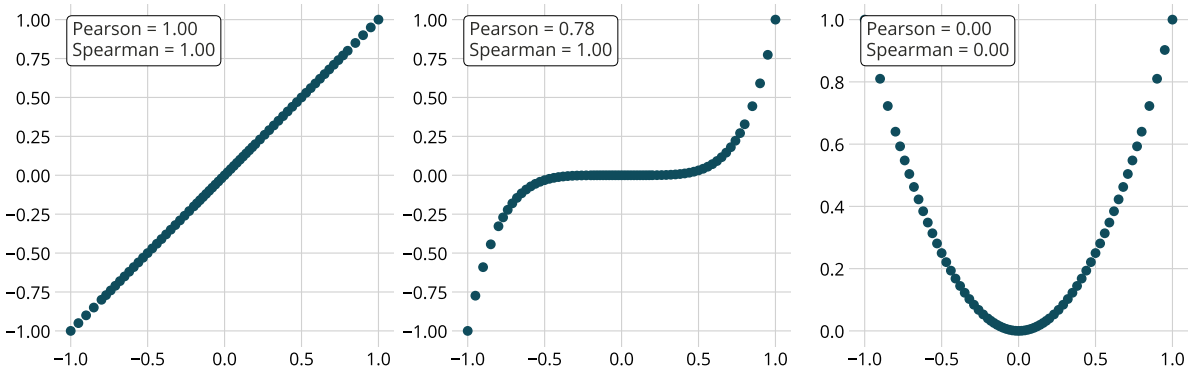


Figure 3.1: Examples of correlation between variables: perfect linear correlation (left), perfect monotonic correlation (center), and no linear and monotonic correlation (right).

3.1.1 Correlation Analysis

In statistics, correlation is a metric that reports the degree to which two or more variables are related. It can be represented by a numeric coefficient, whose value varies between -1 and 1. A correlation is perfectly positive when the coefficient is equal to 1, and perfectly negative when equal to -1 [52]. Coefficients equal to zero indicate no explicit correlation between the variables considered. There are several ways to calculate correlation coefficients, but we list the two main ones.

Pearson Coefficient (r). It measures the linear relationship (first-order) between variables. A relationship is linear when a variation in one variable is associated with a proportional variation in the other variable. Thus, if two variables are linearly correlated, one can be used to predict the other.

Spearman Coefficient (ρ). It measures the monotonic correlation between variables, i.e., uses the order of the data (rank) instead of the values themselves. Thus, if two variables are monotonically correlated, they tend to vary together, but not necessarily at a constant rate.

Figure 3.1 shows examples of correlation between variables. The left plot presents a perfect positive linear correlation between two variables ($r = 1$). Note that when there is a total linear correlation, there is also a total monotonic correlation ($\rho = 1$). In contrast, the center plot contains a perfect monotonic correlation ($\rho = 1$). However, the linear correlation is not perfect ($r = 0.78$), as there is no constant rate in the increase of the variables. Finally, the right plot is a case in which there is no linear nor monotonic correlation ($r = 0$ and $\rho = 0$). This fact does not mean that the variables are not correlated. Indeed, there is a quadratic relation between them, which is not captured by such coefficients.

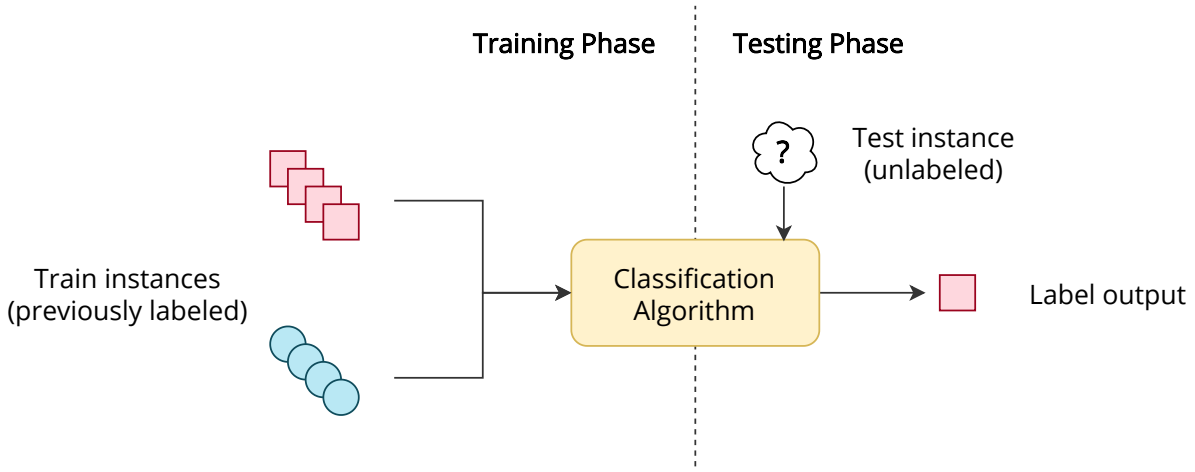


Figure 3.2: A generic workflow for a classification approach.

3.1.2 Classification Algorithms

Classification is a machine learning task whose goal is to automatically assign (i.e., predict) labels to a set of unlabeled instances by considering each instance's features. Such an assignment is made based on a previously known set of instances and a mathematical model. Figure 3.2 presents a generic workflow for classification, which is commonly divided into two phases: training and testing. In the former, the classifier receives a set of instances previously labeled (i.e., training set), representing the actual knowledge about the classes. In the latter, the classifier is tested with a set of unlabeled instances that represent unseen data. Then, it must assign a label for such test instances based on the knowledge from the training phase.

There are several algorithms used in classification tasks. In this section, we present four methods used in this work: Linear Regression, Support Vector Machines (SVM), Perceptron, and Stochastic Gradient Descent (SGD). For formal definitions and other classification algorithms, we refer to further references [71, 96].

Linear Regression. Despite having "regression" in its name, Logistic Regression is a classification learning algorithm. It is mainly used for binary classification tasks, in which there are only two options (classes) for the target to be assigned. Such an algorithm is based in the standard logistic function (also known as sigmoid) to assign probabilities for instances to belong to each class.

Support Vector Machines. The goal of this class of algorithms is to separate the training data points with the larger gap possible in the space with the so-called support vectors. In the testing phase, the new instances are put in the same space, and the predicted labels are defined based on which side of the gap they are.

Perceptron. It is a simple neural network model and works on a single node (neuron) that receives the input data to predict a class label. The prediction is made based on the weighted sum of the inputs (i.e., activation). Such a result is compared to a threshold, which determines the output label for the given instance.

Stochastic Gradient Descent. It is a simple approach to fit linear classifiers (e.g., Logistic Regression and Support Vector Machines) into convex loss functions. Therefore, it is not precisely a classification algorithm but an optimization method.

3.1.3 Performance Assessment

Once the classification model is set and ready to be run on the problem instances, it is necessary to evaluate its performance. Nonetheless, performance metrics are essential for comparing classifiers (and other machine learning algorithms). Due to the specific nature of each problem, there are distinct metrics for the classification task. Here, we list the most widely used of such metrics. Except for the confusion matrix, all metric values are in a range from 0 to 1, in which the higher the value, the better the classifier.

Confusion Matrix. It is a table that compares the predicted with the real target values. For binary classification problems, it is composed of four values: *True Positives* (TP), *False Positives* (FP), *False Negatives* (FN), and *True Negatives* (TN).

Accuracy. It is the ratio of values correctly predicted and the total of predictions. This metric is useful when prediction errors for all classes are equally important (i.e., false positives and false negatives). Its value is given by Equation 3.1.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.1}$$

Precision. It is the ratio of correct positive predictions and the total of positive predictions (including TP and FP), as given by Equation 3.2.

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

Recall. It is the ratio of the correct positive predictions and the real positive values (i.e., TP and FN), as given by Equation 3.3.

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

F1-Score. It is the harmonic mean of Precision and Recall, and its value is given by Equation 3.4.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.4}$$

Area under the ROC curve (AUC). The Receiver Operation Characteristic (ROC) curve is another method used to assess classification performance. It is built based on the *True Positive Rate* (TPR, Equation 3.5) and the False Positive Rate (FPR, Equation 3.6). Therefore, the area under the ROC curve (AUC) is frequently used to summarize such a curve in a numeric value.

$$TPR = \frac{TP}{TP + FN} \tag{3.5}$$

$$TPR = \frac{FP}{FP + TN} \tag{3.6}$$

3.2 Network Science

Network Science is a knowledge field defined by its study object and its methodology to model interconnected systems. Such a field offers a common language in which scientists from the most diverse research areas may analyze and get relevant information about their complex systems [15]. As many aspects in the real world are connected, Network Science emerges as an interdisciplinary field that provides simple yet powerful tools to model protein interactions, power grid transmission lines, and human social interactions. For instance, early works on social networks reveal the importance of human connections (i.e., *networking*) to get a job [42]. Indeed, Social Network Analysis has become a widely studied area, primarily due to the advances in technology and online platforms. Therefore, in this section, we briefly define some fundamental concepts from the network theory used throughout our work.

Formally, a social network is modeled as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, in which $\mathcal{V}$ is the set of vertices (nodes) that represent individuals (e.g., friends, artists), and $\mathcal{E}$ is the set of non-directed edges that connect vertices of individuals who share a relationship. To qualify such relationships, there are metrics for the weight of the edges (also known as strength and tie strength), which can be topological (given by the network structure) or semantic (given by the relation meaning). Research on Music Information Retrieval also benefits from the Network Science framework. For example, modeling collabora-

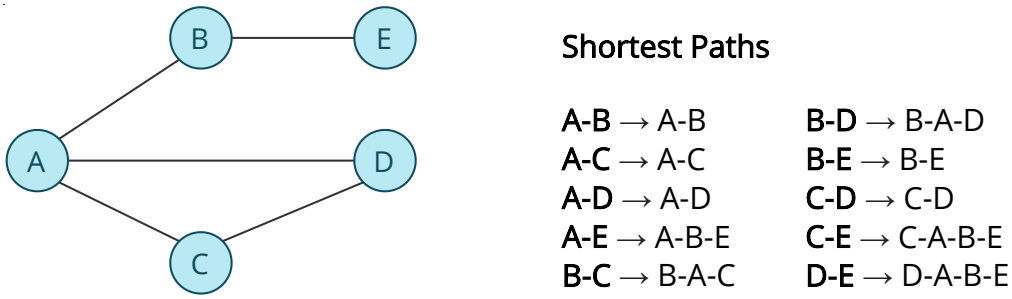


Figure 3.3: A generic network with shortest paths between nodes.

tion between artists who team up in songs allows understanding several aspects of such relationships, such as artist popularity [69] and success-based collaboration profiles [100].

Connectivity is a key concept for studying social networks since it allows to uncover how information flows throughout the network. In this work, we are interested in modeling genre connections to study how they affect success, and therefore we use well-established Network Science metrics to analyze musical collaboration. Such metrics consider the network topological features, i.e., they relate to the network structure (nodes and edges) as follows.[1] Let u and v be nodes in a network. For a given node u, let k_u be its degree (see definition ahead) and $\mathcal{N}(u)$ be its set of neighbors.

Degree and Weighted Degree. These metrics refer to the connectivity of each node in the network. The degree of a node is its amount of incident edges, and the weighted degree is the sum of the edges' weight. For example, in the network of Figure 3.3, the node A has a degree equal to 3, as there are edges connecting A to nodes B, C, and D. The edges in such a network are not weighted, and therefore the weighted degree values are equal to the degree (i.e., the edges' weight is 1).

Clustering Coefficient (CC). Measures the tendency of neighbors of a node to be connected themselves. The higher its value, the more interconnected the node neighborhood. Considering $L(v)$ as the number of links between the neighbors of v, the clustering coefficient of v is given by Equation 3.7:

$$CC(v) = \frac{2L(v)}{k_v(k_v - 1)} \tag{3.7}$$

Calculating the CC value for node A in Figure 3.3 requires the number of links (L) between the neighbors of A (i.e., B, C, and D). Since there is only one edge between C and D, $L(A) = 1$. In addition, the degree of A is 3 (i.e., the number of neighbors of A). Therefore, $CC(A) = \frac{2\times1}{3\times2} = \frac{1}{3} \approx 0.333$.

Common Neighbors (CN). The number of neighbors that a given pair of nodes have in common in a network, i.e., the intersection of their neighbor set, as formalized by Equation 3.8.

[1] For more information on such connectivity metrics, see references [64, 73, 66]

$$CN(u, v) = |\mathcal{N}(u) \cap \mathcal{N}(v)| \tag{3.8}$$

In the example from Figure 3.3, the only node that is neighbor of both A and D is the node C, and therefore $CN(A, D) = |\{C\}| = 1$.

Neighborhood Overlap (NO). The ratio between the common neighbors of a given pair of nodes and the union set of their neighbors. Edges with low NO reveal local bridges in the network, i.e., nodes traveling in "social circles", having almost no common connection. Furthermore, the removal of such an edge may completely disconnect the graph (if $NO = 0$) or difficult the access to other network components ($NO > 0$). Its value is given by Equation 3.9.

$$NO(u, v) = \frac{|\mathcal{N}(u) \cap \mathcal{N}(v)|}{|\mathcal{N}(u) \cup \mathcal{N}(v) - \{u, v\}|} \tag{3.9}$$

For example, in Figure 3.3, the edge connecting A and B is a bridge, since removing it would disconnect the network. Therefore, $NO(A, B) = \frac{|\emptyset|}{|\{C,D,E\}|} = 0$.

Preferential Attachment (PA). The probability of a given pair of nodes connecting in the future. The intuition behind this index is that, if a node has a high degree, it attracts more neighbors. Thus, when analyzing two nodes, the more neighbors they have, the more likely they are to connect in the future. Its value is given by Equation 3.10.

$$PA(u, v) = |\mathcal{N}(u)||\mathcal{N}(v)| \tag{3.10}$$

In Figure 3.3, the Preferential Attachment value for the edge connecting A and C is $PA = |\{B, C, D\}||\{A, D\}| = 3 \times 2 = 6$.

Edge Betweenness (EB). The fraction of shortest paths that go through an edge in the network. Edges with a high score represent a bridge-like connector between two parts of the network, and their removal may affect the communication between others due to the lost common shortest paths. The betweenness centrality c_B of an edge $e = (u, v)$ is given by Equation 3.11.

$$c_B(e) = \sum_{s,t \in V} \frac{\sigma(s, t|e)}{\sigma(s, t)} \tag{3.11}$$

For example, getting the EB value for the edge $\{A,B\}$ in Figure 3.3 requires counting the number of shortest paths passing through such an edge. From all ten paths, only six contain the edge (A, B). Therefore, $c_B(A, B) = \frac{6}{10} = 0.6$.

Resource Allocation (RA). For a pair (u, v) of nodes, it represents the fraction of a resource (e.g., information) that a node can send to another through its common neighbors, as given by Equation 3.12. If both nodes have a large number of common

neighbors, their RA Index tends to be high. Such an index is even higher if their neighbors have a low degree, as the resource is more likely to travel from u to v.

$$RA(u,v) = \sum_{w \in \mathcal{N}(u) \cap \mathcal{N}(v)} \frac{1}{|\mathcal{N}(w)|} \tag{3.12}$$

For example, calculating RA for A and D in Figure 3.3 requires to consider the common neighbors of both nodes (in this case, only the node C). Therefore, the RA value is given by $RA(A,D) = \frac{1}{|\mathcal{N}(C)|} = \frac{1}{|\{A,D\}|} = \frac{1}{2} = 0.5$.

3.3 Data Mining

As a novel and dynamic environment, the musical scenario brings high volumes of data about songs, their characteristics, and the social interactions about them. The popularization of digital platforms allows people worldwide to access and interact with content in real-time [44], increasing the cultural connection between distinct parts of the globe, while each market maintains its unique characteristics. From this context, Data Mining emerges as a research field aiming to discover relevant insights and patterns, as well as to build models to describe and understand such data. In this section, we present the main Data Mining concepts used in this work, following the definitions of Zaki and Meira Jr. [115]: Frequent Itemsets (Section 3.3.1), Association Rules (Section 3.3.2), and Subgroup Discovery (Section 3.3.3).

3.3.1 Frequent Itemset Mining

Frequent Itemset Mining (FIM) is a Data Mining approach to find groups of items that co-occur in the same transaction. Such a model is also known by the term *market basket analysis*, given that one of its classic applications is the analysis of customers' shopping patterns in supermarkets. The total set of *items* $\mathcal{I} = \{x_1, \cdots, x_m\}$ can describe all the items sold at the supermarket. A subset $X \subseteq \mathcal{I}$ is called an *itemset*. In the supermarket example, a transaction may represent a set of items bought by a specific customer, i.e., the shopping list. Formally, a transaction is a tuple (t, X), in which t is an unique identifier (i.e., *tid*), $t \in \mathcal{T}$ ($\mathcal{T}$ is the set of all *tids*, $\mathcal{T} = \{t_1, \cdots t_n\}$).

A *database* $\mathbf{D} = \mathcal{I} \times \mathcal{T}$ is a binary relation between the sets of items and *tids*,

and a *tid* $t \in \mathcal{T}$ contains an item $x \in \mathcal{I}$ if and only if $(t, x) \in \mathbf{D}$. In addition, let $\tau(X)$ be a function that returns all the transactions $t \in \mathcal{T}$ that contains the itemset X. From all these definitions, the sets of items frequently purchased by the same customer are analyzed, i.e., items that occur in the same basket. Note that the supermarket example facilitates the understanding, but these definitions can be generalized to any other domain. For example, as detailed in Chapter 6, we can model songs as transactions in which the items are the music genres of the artists who interpret them.

However, in FIM, the simple co-occurrence of these items in a single transaction may not be enough. In most cases, such co-occurrence must happen with a minimum frequency. For supermarkets, discovering items that are frequently bought together helps develop marketing strategies to increase their revenues. In contrast, for the music industry, discovering genre patterns may help record labels direct their promotion strategies. There are several algorithms for mining frequent itemsets (e.g., Apriori, Eclat, and FP-Growth), but here we focus only on the metrics used by all such methods to define the itemsets' frequency.

Support. This metric informs how many transactions contain a given itemset in absolute terms. Equation 3.13 presents the formula for calculating the itemset support.

$$sup(X, \mathbf{D}) = |\tau(X)| \tag{3.13}$$

Relative Support. Similarly, the relative support ($rsup$) informs the frequency in which an itemset appears on the transactions in a scale from 0 to 1. Then, $rsup = 1$ means that an itemset occurs in all transactions. Its value is given by Equation 3.14.

$$rsup(X, \mathbf{D}) = \frac{sup(X, \mathbf{D})}{|\mathbf{D}|} \tag{3.14}$$

3.3.2 Association Rules

Once mined, frequent itemsets can be used to generate Association Rules (AR). An AR is represented by the expression $X \rightarrow Y$ and is composed of an antecedent X and a consequent Y, two disjoint itemsets. It is important to highlight that an AR should not be interpreted as a sign of causality but of co-occurrence between items. Indeed, association rules allow to discover how itemsets are related, and there are several metrics to assess rule quality, as those described next.

Confidence. The rule confidence informs the probability of a consequent Y occurring in a transaction given the occurrence of an antecedent X. In other words, it is the frequency

in which Y occurs in transactions containing X, as given by Equation 3.15.

$$conf(X \rightarrow Y) = \frac{sup(XY)}{sup(X)} \tag{3.15}$$

Lift. It is defined as the ratio between the joint probability of X and Y co-occurring and the probability of these sets being independent, as shown by Equation 3.16. It may be used as a measure of surprise within a rule. Therefore, lift shows how much more frequently the consequent Y becomes after the occurrence of the antecedent X. Such a metric is symmetric, and values below 1 mean that the rule occurs less than expected, whereas values above 1 indicate the opposite.

$$lift(X \rightarrow Y) = \frac{rsup(XY)}{rsup(X) \cdot rsup(Y)} = \frac{conf(X \rightarrow Y)}{rsup(Y)} \tag{3.16}$$

3.3.3 Subgroup Discovery

Alongside Frequent Itemsets and Association Rules, we also use Subgroup Discovery (SD) in our work, a widely used technique in data mining to identify relevant patterns (subgroups) that deviate from the standard [57]. Here, let a dataset D be a collection of instances $x = (a_1, \cdots, a_m, t_1, \cdots, t_l)$, in which a_i is an attribute (from a set $\mathcal{A}$) and t_i is a target variable. Formally, a subgroup is induced by a p pattern, which is a function $p : \mathcal{A} \rightarrow \{0.1\}$. Thus, a subgroup S_p is defined as the set of instances covered by p, i.e., $S_p = \{x \in D \mid p(a_1, \cdots a_m) = 1\}$. Subgroups are then described according to their attributes, and they are relevant if the distribution of their target variable is very deviant from that observed in the whole dataset. The evaluation of the relevance of the subgroups is done with predefined quality metrics.

There are several search strategies to find relevant subgroups within a dataset, such as Beam, Exhaustive, and Evolutionary [46]. Subgroups are specified by a description language defined by domain experts and analysts. According to Rebelo de Sá et al. [90], such languages are frequently composed of conjunctions of attribute conditions. For example, consider a dataset in which the instances represent the movie preferences of streaming users. For each user, the attribute set comprises demographic information including age, country, and marital status. Besides, the target variable is defined as the user's favorite movie genre (e.g., action, drama, or sci-fi). Thus, in a scenario where the overall favorite genre is *drama*, a possible subgroup found by an SD algorithm is:

$$\text{Age} \geq 40 \quad \wedge \quad \text{Country} = \text{“Brazil”} \quad \Rightarrow \quad \text{Genre} = \text{“Action”}$$

This subgroup means that the people over 40 who live in Brazil have a distinct genre preference when compared to the whole dataset. That is, Brazilians over 40 prefer action movies, while people in general are more into drama. Therefore, using SD in descriptive analyses helps to reveal hidden groups with exceptional preferences that deviate from the average. In this work, we use an SD algorithm to find relevant subgroups within the genre collaboration network, as described in Chapter 6.

Chapter 4

Identifying and Predicting Hot Streaks in Musical Careers

Professional careers tend to have phases of high productivity, reaching the career peak. Hot streaks (HS) is the term commonly used for continuous periods of success above normal. Previous work reveals that hot streaks can arise at any time in a professional career [101, 65, 38]. In addition, such phases can occur in different ways in different areas. For instance, scientists can reach their peak when their publications achieve high citations. For athletes, such success may be based on rewards or victories in renowned competitions. We highlight Michael Phelps' career, the most decorated person of the Olympic Games history who has conquered 28 medals.[1] His career peaked at the 2008 Summer Olympics, winning gold in all eight competitions he disputed. However, this success did not come suddenly, as Phelps needed two editions to achieve such a milestone.

Such productivity peaks can also happen in creative careers, including cinema, art and music [54]. The latter is one of the most volatile industries, being considerably sensitive to external factors. According to IFPI Global Music Report 2021,[2] until 2016, physical media were the main form to consume music. After that, streaming platforms came to dominate the music market, moving around US$ 13.4 billion in 2020. This dynamic nature of the music industry can directly influence the behavior of artists' careers. That is, an artist's career can suffer ups and downs depending on the current market moment. For example, the singer Cher holds the record for the longest break between #1 hits on the Billboard Hot 100,[3] totaling almost 25 years between singles *Dark Lady* (March 23, 1974) and *Believe* (March 13, 1999).

Overall, the music industry is as dynamic as it is a crucial part of the entertainment world. Within so much uncertainty, a clear fact is: when an artist is on a hot streak, such an artist is also at the most profitable moment of a career. One hit wonders have just one peak and that is it, they are done (ergo the expression "one hit wonder"). Now, the real stars in the business are able to achieve many peaks and produce millions of dollars per week. Therefore, identifying hot streaks is one way of investing in the right artist at

[1] International Olympic Committee: https://www.olympic.org/athletes
[2] IFPI Global Music Report 2021: https://gmr.ifpi.org/
[3] Billboard Magazine: https://bit.ly/3hfzMJ0

the most relevant moments. Such identification is also useful for planning, adjusting, and even completely changing a marketing direction, for example.

In such a context, we identify and characterize hot streaks in the music scene, defined by high-impact bursts occurring in sequence within artist careers. Specifically, we aim to answer the following research questions (RQs):

RQ1. How do the most impactful weeks in musical careers are distributed over time?

RQ2. Does this behavior generalizes into continuous periods of high impact (i.e., hot streaks)?

RQ3. Are there specific hot streak patterns for distinct musical genres?

RQ4. What happens before, during, and after a hot streak period?

RQ5. Is it possible to predict whether a week belongs to a hot streak period?

RQ6. What are the factors that influence hot streak periods?

By answering all such questions, we uncover relevant insights on the temporal evolution of musical careers. Furthermore, we propose a model for predicting the most successful phases on the musical genres scale by grouping artists belonging to the same genre. The remainder of this chapter is organized as follows. First, in Section 4.1, we introduce the methodology to answer all research questions. Next, Section 4.2 presents and discusses the results and experimental evaluation for each RQ. Finally, we conclude this chapter with our overall considerations on the work in Section 4.3.

4.1 Methodology

To answer all research questions regarding hot streaks in musical careers, we propose a four-step methodology. We start by collecting data from Billboard and Spotify to build a success-based dataset (Section 4.1.1). From such data, we build chart-based success timelines for both artists and music genres and characterize them (*RQ1*, Section 4.1.2). Then, in Section 4.1.3, we identify hot streak periods in all such time series and perform exploratory analyses in order to answer *RQ2*, *RQ3* and *RQ4*. Finally, in Section 4.1.4, we model the hot streak prediction as a binary classification task (*RQ5*), and we use such a model to uncover the factors behind their occurrence (*RQ6*).

4.1.1 Data Collection

Billboard is an American-based music specialized magazine, which also operates in Canada, Brazil, Greece, Japan, South Korea, and Russia. Billboard is widely known for its exclusive charts on trends across all musical genres. According to IFPI,[4] the United States is the biggest music market in the world, and the *Hot 100*[5] is the main all-genre song ranking in the country. It is weekly published by Billboard since 1958, and is currently built by considering songs' sales, airplay, and streaming data. Thus, to model artist success over time, we collect all Hot 100 charts from August 11, 1958 to August 22, 2020 (data collection time) using the Python package *billboard.py*.[6] Each chart contains 100 entries, ranked from the most popular to the least popular song on that week.

However, Billboard charts do not offer all information necessary to answer our research questions, as chart entries are composed only of the song name and its artists. Therefore, we enrich our dataset by collecting data from Spotify, the world's most popular audio streaming service with more than 356 million users in 178 markets (as of March 2021). Using its API,[7] we are able to get extra information on artists and songs. Specifically, we obtain artist genres and debut date, as well as acoustic features for each song, such as key, mode, and energy. Our final dataset[8] is composed of 3,238 weekly charts containing 24,540 distinct songs from 6,248 artists belonging to 998 music genres. Such enriched and curated data allow us to build success-based time series to investigate the presence of hot streak periods in artist and genre careers. Figures 4.1 and 4.2 characterize our dataset by presenting the top 25 genres in the United States based on the number of artists and songs, respectively.

4.1.2 Time Series Modeling

Success in the music industry has a temporal structure, as the audience's tastes change over time. The dynamics of media platforms, the emergence of new music styles, and the artists' releases are some factors that shape what listeners consume. In this work, we use the Hot 100 charts as our basis to model success over time. For each artist, we build their time series from the debut date (i.e., the date of the first release obtained from

[4]IFPI Global Music Report: https://gmr.ifpi.org/
[5]Billboard Hot 100 Chart: https://www.billboard.com/charts/hot-100
[6]billboard.py: https://github.com/guoguo12/billboard-charts
[7]Spotify Developer API: https://developer.spotify.com/
[8]The dataset is publicly available on http://bit.ly/proj-bade

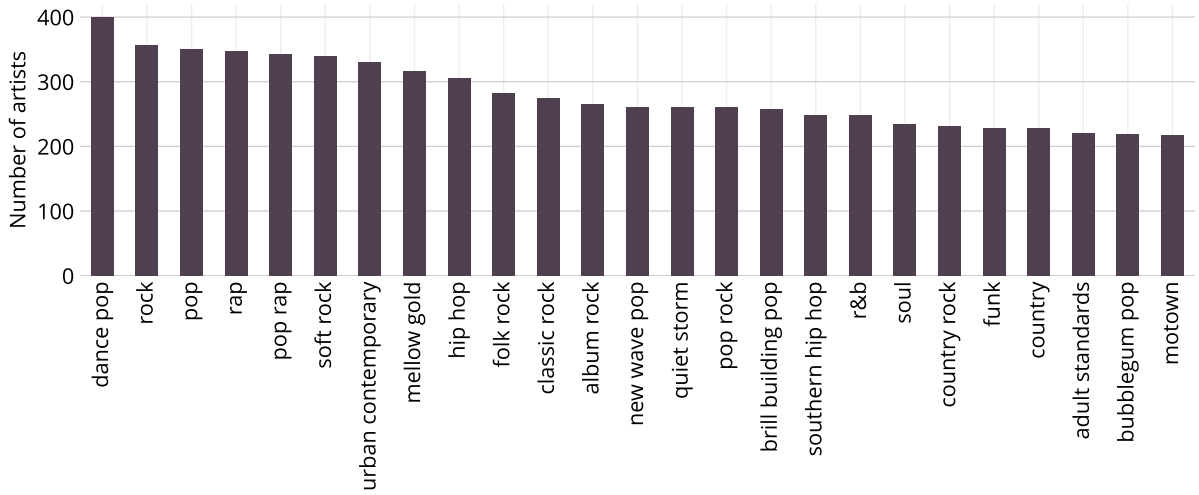


Figure 4.1: Top 25 music genres in the United States, sorted by the number of artists.

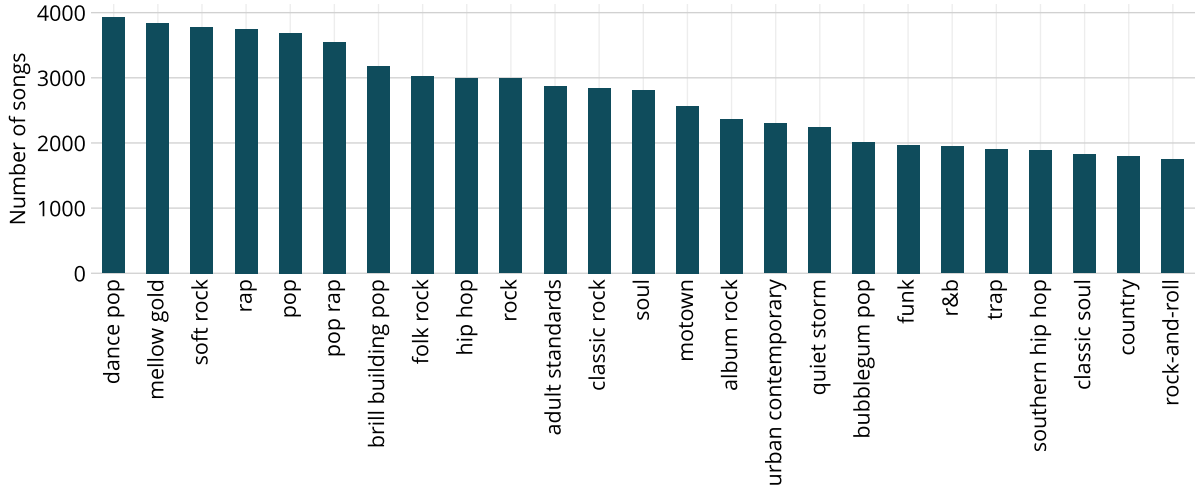


Figure 4.2: Top 25 music genres in the United States, sorted by the number of songs.

Spotify) to the last chart collected. Thus, each point in the time series represents the success of such an artist in a given week, according to the Hot 100 chart.

We measure the success of an artist by calculating the *rank scores* for all of their songs that appear on the week chart. The *rank_score* of a song i is $rank\_score(i) = max\_rank - rank(i) + 1$, where *max_rank* is the lowest possible rank (in our case, 100) and *rank(i)* is the position of the song on the chart. We then aggregate the rank scores of an artist using Discounted Cumulative Gain (DCG) [2], as this metric emphasizes the most relevant records (i.e., the highest ranked songs on the chart) and penalizes by a logarithmic factor songs that appear lower. The DCG value for an artist is given by Equation 4.1.

$$DCG = \sum_{i=1}^{n} \frac{rank\_score(i)}{log_2(i+1)} \tag{4.1}$$

Using such a metric is better than just summing the rank scores of the songs

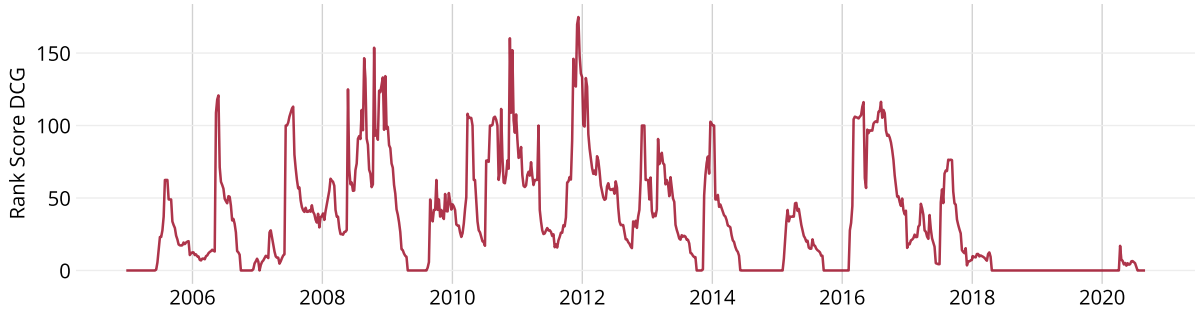


Figure 4.3: Rihanna's success time series (2005-2020). The success is measured on the rank score DCG obtained from weekly Hot 100 charts.

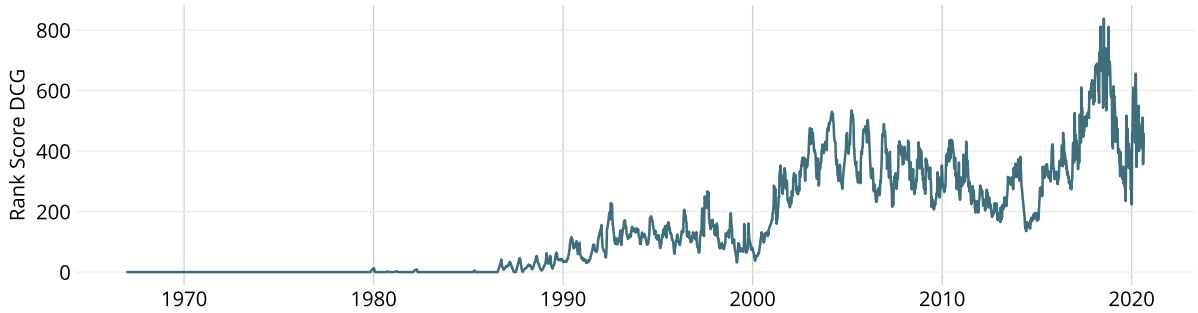


Figure 4.4: Rap success time series (1967-2020). The success is measured on the rank score DCG obtained from weekly Hot 100 charts.

because, in our case, it is preferable for an artist to have the #1 song than two songs in the middle of the chart (e.g., on positions #49 and #50). Therefore, DCG is more appropriate to measure success on the charts, and such metric is defined as the artist's success measure for each week. For example, Figure 4.3 shows Rihanna's success time series from her debut in 2005 to 2020. The highest success peak is observed by the end of 2011, when she released her sixth studio album *Talk That Talk*, which had *We Found Love* (in collaboration with Calvin Harris) as the lead single. The song stayed on the top of the Hot 100 for ten non-consecutive weeks, becoming the longest-running number-one single for both artists.

Analyzing genres' success over time is also one of the goals of this work, and we can build genre success time series based on data obtained from Spotify. First, we assign artists' genres to their songs, as the songs themselves do not have such information. Then, for each week, we aggregate all songs from artists belonging to a given genre that appear on that week's chart using DCG. Note that a song may be accounted for several time series, as artists often have more than one genre. For example, Michael Jackson's genre list includes *pop*, *r&b*, and *soul*, and thus his songs are included in all three time series.

Considering genres as collections of artists allows a high-level view of their success, as it becomes easier to identify and analyze genre popularity trends over time. Figure 4.4 presents the success time series for rap from 1967 to 2020. Artists from this genre

have been consolidating their presence on the charts since the late 1980s, but the first great success wave happened during the 2000s. Such an impactful era was mainly led by artists such as Jay-Z, Eminem, and Missy Elliott, who led the charts with successful solo hits and collaborations with pop and r&b artists, which became extremely popular in this period. In the late 2010s, rap achieved the highest popularity peak, with a new generation of rappers in the mainstream, such as Cardi B, Drake, and Kanye West.

4.1.3 Hot Streak Detection

After modeling success in musical careers and with the evidence that the most successful weeks tend to happen close to each other (see the previous section), we now assess *RQ2* by investigating if such a behavior generalizes into periods of higher impact (i.e., hot streaks), following recent research on this subject [38, 65]. First, we use a technique to reduce the dimensionality of the time series to delimiter periods within careers (Section 4.1.3.1). Then, we define a hot streak as the periods in which the success is above a certain threshold, obtained from the artist/genre career itself (Section 4.1.3.2).

4.1.3.1 Piecewise Aggregate Approximation

In this section, we present Piecewise Aggregate Approximation (from now on, PAA), a method to reduce the dimensionality of a time series proposed by Keogh and Pazzani [55]. Given a time series $X = x_1, x_2, \cdots, x_n$ of length n, PAA reduces it into a new series $\overline{X} = \overline{x_1}, \overline{x_2}, \cdots, \overline{x_N}$ with N dimensions, $1 \leq N \leq n$. The intuition behind this method is that the division of the original time series into N equal-sized segments would produce N new points. Their values are calculated by the average of the points within such frames, as given by Equation 4.2. Therefore, the approximation of each point on the original time series is made by simply assigning the PAA value of its corresponding segment.

$$\overline{x_i} = \frac{n}{N} \sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} x_j \tag{4.2}$$

We use such a method because highly impactful periods within artists' careers may contain weeks with low values for the success metric. Therefore, PAA is a helpful tool to smooth such differences and delimit periods in careers. In this work, we use the PAA implementation of *tslearn*[9] [105], a Python package for time series analysis. Running it requires defining the number of segments in which the series will be split, as this is

[9]tslearn: https://github.com/tslearn-team/tslearn/

the only parameter of the method. This is made individually for each time series, as artists' careers have different sizes. For example, Elvis Presley's time series begins in 1958, whereas Adele's starts in 2007. Thus, the number of segments for PAA differs according to the time series length. To allow comparison between different careers, we define a unique size of 52 weeks (i.e., one year) for each segment. Hence, the number of segments is calculated by dividing the time series length by this predefined size.

4.1.3.2 Defining Hot Streaks

The next step in our methodology is to define what makes a hot streak in artists' careers. Recalling the definition of hot streaks, such periods must present a success rate above the usual. Therefore, we identify hot streaks as the periods in which the success metric (approximated by PAA) is higher than a predefined threshold. Similar to the number of segments from PAA, we define this threshold individually for each artist's career. We base such a threshold on the artist's *activity rate* (AR) on Hot 100, which is the ratio between the number of weeks in which the artist appears on the chart and the total number of weeks of the time series. Hence, we define the hot streak threshold for each artist, as follows:

- **AR** $\geq$ **20**%: threshold is the *80th percentile* of the success metric;
- **15**% $\leq$ **AR** $<$ **20**%: threshold is the *85th percentile* of the success metric;
- **10**% $\leq$ **AR** $<$ **15**%: threshold is the *90th percentile* of the success metric;
- **AR** $<$ **10**%: threshold is the *95th percentile* of the success metric.

4.1.4 Definition and Metrics for Hot Streak Prediction

To predict the occurrence of hot streaks, we model such a problem as a binary classification task from a specific set of features (Section 4.1.4.1). Using such a definition, we build our experimental analysis through a preprocessing phase and selected learning methods evaluated by proper metrics (Section 4.1.4.2).

4.1.4.1 Problem Definition

We model the hot streak prediction as a *binary classification* task in which, for a given week, an algorithm predicts whether it belongs to a hot streak period in a time series or not. Formally, let X denote a set of weeks temporally sorted, and $Y = \{0, 1\}$ be the label space (i.e., 1 if a week is part of a hot streak and 0 otherwise). Thus, binary classification

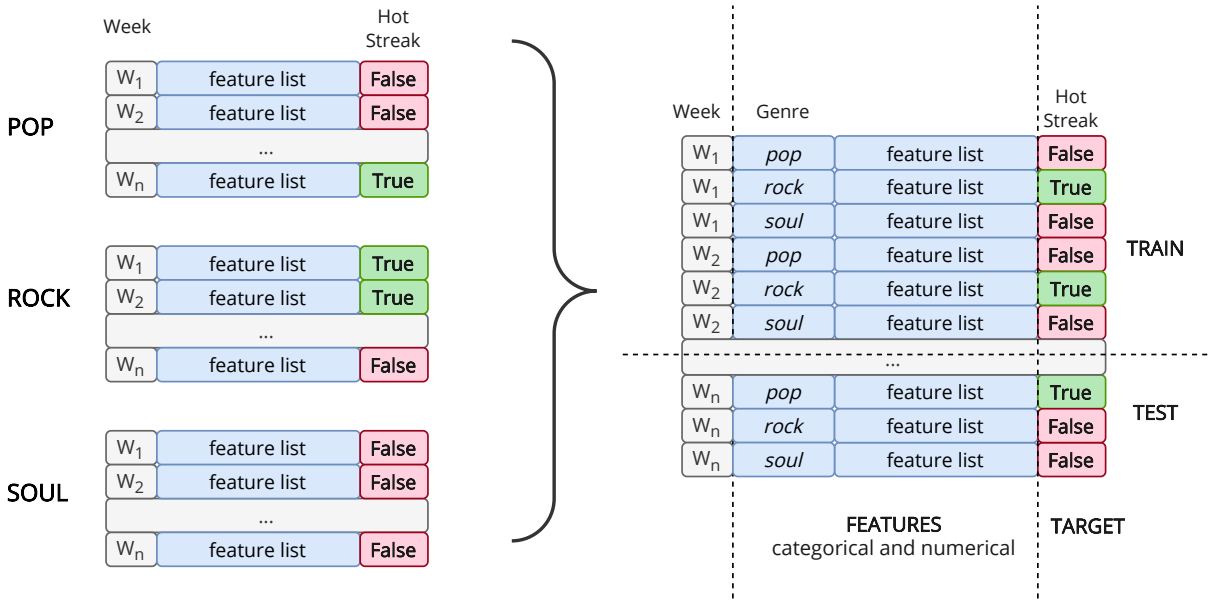


Figure 4.5: Music-oriented Hot Streak Binary Classification from genre time series.

aims to learn a function $f : X \rightarrow Y$ from the training set $\{(x_i, y_i) \mid 1 \leq i \leq m\}$, where $x_i \in X$ is an instance characterizing the features of a week, $y_i \in Y$ is the corresponding target value, and m is the number of instances. Then, the algorithm performs the predictions over a test set $\{(x_i, y_i) \mid 1 \leq i \leq n\}$ of n instances unseen by the model so far, representing real-world data.

Here, we consider genre time series for the prediction because artists' careers are very distinct from each other, and the majority of them have only one hot streak (see Section 4.2.3.1), which may affect the training process. In addition, previous works provide evidence that a luck component plays an important role within success in individual careers [54, 101]. On the other hand, genre careers are more stable and have well-established hot streak periods, providing examples of both hot streak and non-hot streak periods for the learning algorithms. Besides, it may be more useful for the music community (i.e., record label CEOs, producers, and artists themselves) to know hot streak periods for genres, as it sheds light on next investment targets and future partnerships.

Figure 4.5 illustrates our classification model, called *MHSBC – Music-oriented Hot Streak Binary Classification*. First, for each week in the genres' time series, MHSBC calculates a set of features (detailed ahead) describing all songs from one genre which are in the week's chart. It then runs PAA (see Section 4.1.3) to get the information about whether such a week is part of a hot streak period or not. Then, it combines genres' time series to get a unique set of instances for our model, which aims to be genre-aware. Thus, the genre information becomes a categorical feature in this final set. Note that we must respect the chronological order of the weeks, as it is extremely relevant in the train-test split phase (i.e., we can not make predictions knowing the "future"). Hence, in MHSBC

model, each instance describes the set of songs from a given genre that entered the Hot 100 chart in that week.

Feature list. To describe genre performance in a given week, we use a set of features obtained from the genre's songs within the charts in that week. We obtain such features from Spotify and divide them into three main groups: *(i) genre-related* – number of genre songs, number of genre distinct artists; *(ii) artist-related* – median artists per song, median career time, median genres per artist; and *(iii) song-related* – number of collaborative songs, number of explicit songs, median danceability, median energy, median key, mode key, median loudness, median mode, median speechiness, median acousticness, median instrumentalness, median liveness, median valence, median time signature, median duration. The description of each feature is given in Appendix A.

4.1.4.2 Setup and Metrics

As mentioned in the previous section, our proposed classification model (MHSBC) uses genre success time series. From the 998 genres of our dataset, we select only the genres with 50 or more artists. We do so to reduce noise in the data, as genres with few artists may be overspecialized (e.g., Texas Latin rap, NYC rap, and Nashville indie). Hence, our final set is composed of 87 genres, which time series are aggregated to become the model input. The train-test split is made chronologically, to keep the notion of training the model with observed data and testing it with the future. For that same reason, we do not perform cross-validation in our model. We split data in a 70-30% proportion for training and test sets, and the split date is defined as January 5, 2002.

Regarding the target label distribution, there is a disproportion in the number of hot streak and non-hot streak instances, which correspond to 20.7% and 79.3% of the total, respectively. In this case, the natural solution would be to resample the training set to obtain a 50-50% distribution. However, we are not able to perform such a technique in our data, otherwise, we lose the temporal information (i.e., order of weeks). This is the core of our modeling, and thus it can not be unconsidered.

Data Preprocessing. We also handle different ranges for both numeric and categorical features to correctly process data in our model. For each type of feature, we perform an appropriate transformation: *(i)* for each numeric attribute, values are normalized into a $[0, 1]$ range; *(ii)* as the genre is the only categorical attribute, it is binarized through the One-hot Encoding technique [39] to adjust it to the input format of most classifiers. No feature presented missing values; therefore, they do not impact our experiments.

Learning Methods and Metrics. The hot streak classification aims to predict if a week is part of a hot streak period for a given genre. Thus, we select four well-established and widely used classifiers: Logistic Regression (LR), Support Vector Classification with linear kernel (LinearSVC), Perceptron, and Stochastic Gradient Descent (SGD). We also

consider a dummy classifier as a baseline performance, which simply predicts the majority class. The best hyperparameter values for each classifier are obtained from Grid Search (see Appendix A), and all classifiers are evaluated using standard learning metrics for classification: accuracy, precision, recall, F1-Score, and the area under the Receiver Operating Characteristic curve (AUC).

4.2 Results and Evaluation

In this section, we present an experimental evaluation on hot streaks, following the methodology from the previous section. First, we verify if the weeks with higher success levels cluster over time (Section 4.2.1 answers *RQ1*). Next, we generalize the continuous success bursts into hot streak periods (Section 4.2.2 answers *RQ2*). We then perform a characterization analysis over them (Section 4.2.3 answers *RQ3* and *RQ4*). Finally, we present the evaluation of the hot streak prediction problem, as well as a study on the underlying factors behind them (Section 4.2.4 answers *RQ5* and *RQ6*).

4.2.1 Clustering Success Over Time

In order to answer *RQ1*, we follow the methodology used by Garimella and West [38] to investigate whether the most successful weeks occur close to each other in artists' careers. In this section, we present the results for artists grouped by genres to investigate cross-genre behavior differences. We define an artist's career as their success time series from their debut to the last date of our collection (August 22, 2020), as modeled in Section 4.1.2. Therefore, we set the position $P(w_i)$ of a week w_i within a time series as its index i. The k most successful weeks (i.e., with the highest Rank Score DCG values) are denoted by $W_1, W_2, \cdots, W_k$.

Here, our analyses are focused on two main points. First, we investigate the timing of the most successful weeks of an artist's career. Then, we look at the distribution of the difference between the positions of the two most successful weeks in artists' careers. Such analyses are all made in comparison with shuffled careers to check the robustness of our findings, that is, if the observed effects still happen.

Timing of most successful weeks. As a first step, we analyze the positions of the five

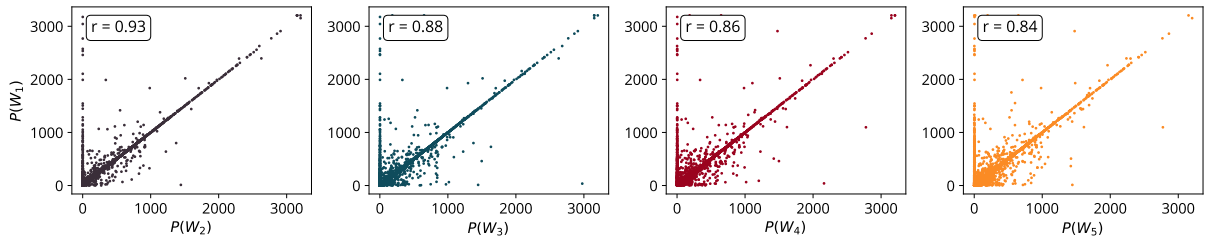
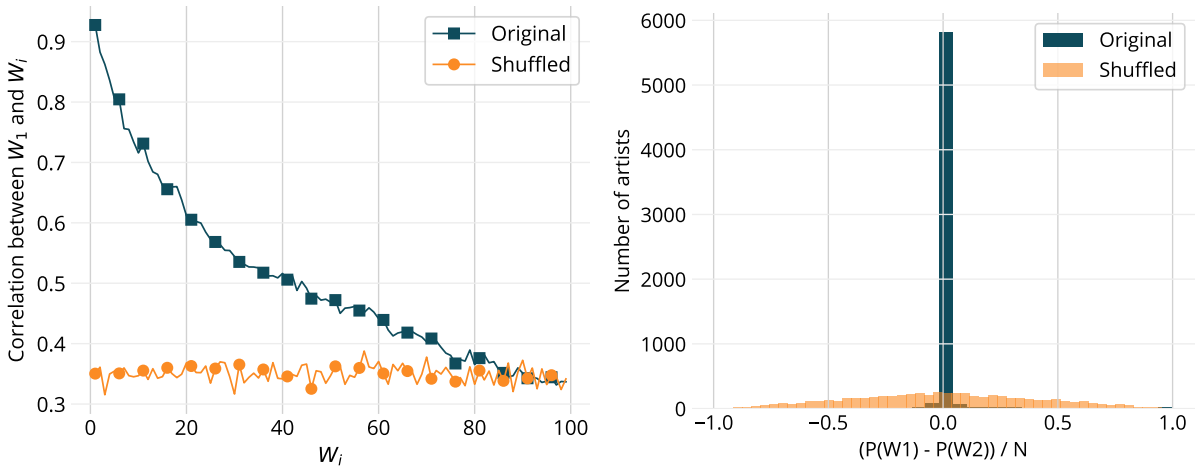


Figure 4.6: Scatter plots with Pearson correlation (r) of the position of the most successful week in artist careers (W_1) with W_2, W_3, W_4, and W_5, respectively. Each point represents an artist. All correlation values are statistically significant ($p < 0.05$).

most successful weeks within artists' careers. Figure 4.6 presents scatter plots of $P(W_1)$ versus $P(W_i)$ for $i \in [2, 5]$, as well as the Pearson correlation coefficient (r) for each plot. We consider all artists from our dataset. The results show a strong linear correlation for all considered genres, and also that the Pearson coefficient is higher when comparing the first and second most popular weeks. In the comparison with the third, fourth, and fifth weeks, the correlation decreases, even though its value remains high. Such a finding reinforces the hypothesis that the most impactful weeks within an artist's career are more likely to happen close to each other.

We then expand the investigation on the correlation values to compare the positions of W_1 to W_i for $i \in [2, 100]$. Figure 4.7(a) shows that there is indeed a decrease in the correlation in all considered genres, but this pattern is not observed in shuffled careers, in which the correlation is always between 0.3 and 0.4. Therefore, there is a general trend of clustering within the most successful weeks in artist careers, as such weeks tend to happen close to each other in the success time series.

Difference of the positions of the most successful weeks. We calculate the difference of the positions of the top two most successful weeks for artists ($P(W_1)$ and $P(W_2)$, respectively). We normalize such a difference by the number N of weeks of the artist time series. Figure 4.7(b) shows that for all considered genres, the distribution has a peak around zero, suggesting that these two weeks are close to each other on the timeline. Such a result agrees with the findings of the previous analysis. Further, when we shuffle artists' careers, the distribution of these differences is much different from the original, demonstrating that this behavior of musical careers is not random. Hence, there is strong evidence that artists may experience periods of outstanding success, or *hot streaks*, which we investigate in the next section.



(a) (b)

Figure 4.7: Correlation between the first and i-th most successful weeks (a) and the normalized difference between the positions of the first and second most successful weeks within artists' careers (b).

4.2.2 Identifying Hot Streaks

In the previous section, we found that successful weeks tend to cluster into bursts of high success occurring in sequence in musical careers. Now, we generalize such continuous above-average success into hot streak periods, answering *RQ2*. We do so by applying Piecewise Aggregate Approximation (PAA, Section 4.1.3) to identify such periods in artist and genre time series based on their performances on weekly charts.

We continue to use Rihanna's time series (see Section 4.1.2) to demonstrate how our method works. Figure 4.8 shows her time series after applying PAA. The threshold is set as the 80th percentile, as she has an AR of 73%. We observe two distinct hot streak periods. The first one from April 2008 to April 2009, when she released *Disturbia* and *Take a Bow*, two smash hits that reached the #1 position on Hot 100. The second hot streak lasted from May 2010 to May 2012, when she released the album *Loud*, which contains the also #1 hits *What's My Name* and *Only Girl (In the World)*. This period also includes the release of the aforementioned single *We Found Love* in collaboration with Calvin Harris.

Hot streaks are also possible for genres, as musical tastes change over time, defining which genres are popular or not. Thus, we also run PAA and set a threshold for genres' time series. Figure 4.9 shows the results for *rap*. Considering genres as sets of artists, we may interpret such a time series as the success of all *rap* artists. We identify three hot streaks: the first going from September 2002 to September 2008, the second from September 2009 to September 2010, and the last starting in September 2015 and still ongoing at data collection time (August 2020). In the period between 2010 and 2015 (i.e.,

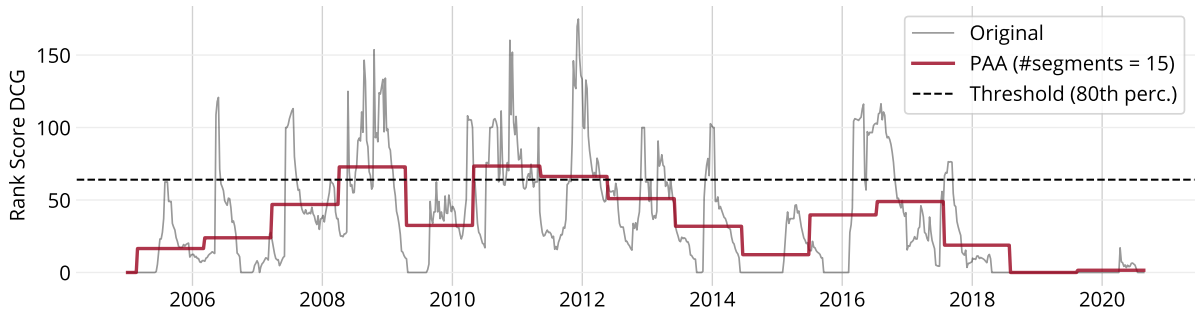


Figure 4.8: Piecewise Aggregate Approximation (PAA) applied to Rihanna's success time series (2005–2020). Periods above the threshold are considered hot streaks.

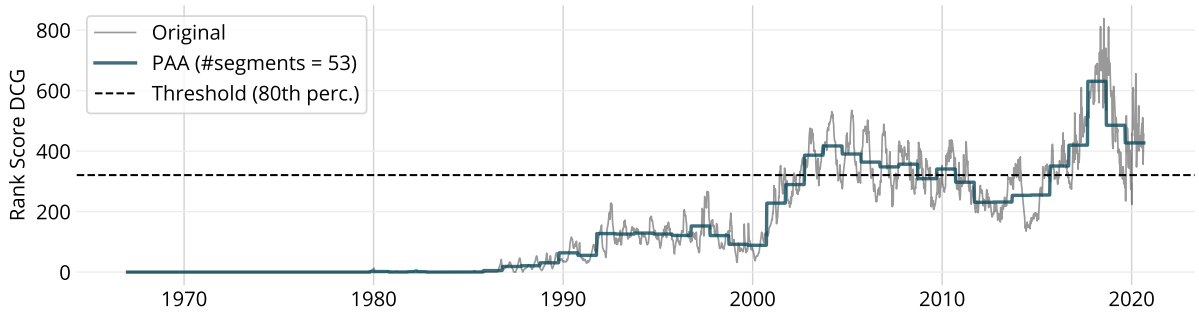


Figure 4.9: Piecewise Aggregate Approximation (PAA) applied to Rap success time series (1967–2020). Periods above the threshold are considered hot streaks.

the last two hot streaks), we highlight the rising of other music genres such as *tropical house* and *electropop*, which may have affected the performance of *rap* artists on charts. Nonetheless, finding such hot streaks reinforces the analyses of Section 4.1.2, as these periods are widely known for the appearance of great names of the *rap* music scene, such as 50 Cent and Kendrick Lamar.

4.2.3 Characterizing Hot Streaks

Here, we deepen the analyses on hot streaks by uncovering several patterns found within musical careers (*RQ3*, Section 4.2.3.1). Furthermore, we perform a comparative analysis on the periods before, during, and after a hot streak to investigate possible trends and effects on musical success (*RQ4*, Section 4.2.3.2).

4.2.3.1 Hot Streak Patterns

In this section, we assess *RQ3* to find if there are specific patterns within artists' careers. We do so by characterizing the hot streak periods considering relevant features such as

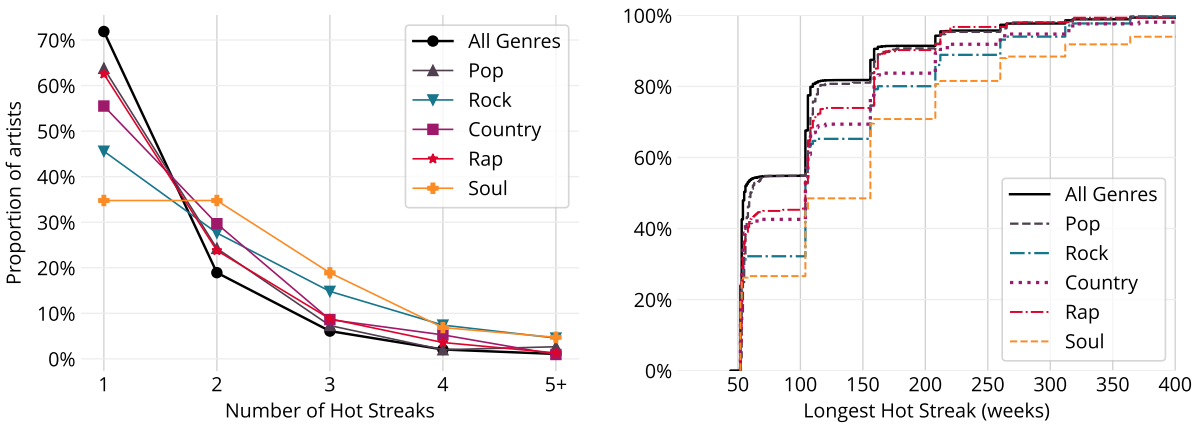


Figure 4.10: Characterization of artists' hot streaks: number of hot streaks (left) and cumulative distribution function of the duration of the longest one in weeks (right).

the number of hot streaks, duration, and position of such a period within the artist's career. Music genres may directly impact artist careers, as different audiences receive and interact with new releases differently. Thus, we perform our analyses by comparing artists from five popular genres in the United States: *pop*, *rock*, *country*, *rap*, and *soul* (see Section 4.1.1). We also compare these music genres with the aggregate of artists from all genres (i.e., the whole artist set) to provide a general overview in the analysis. Artists are grouped considering their Spotify genre list.

We first investigate the number of hot streaks per artist. Figure 4.10 (left) shows that, in general, the majority of artists (about 70% to 90%) have between one and two hot streak periods in their careers. Such a pattern happens in all considered genres, but there are genres with a higher percentage of artists with two or more hot streaks, such as *soul* and *rock*. This phenomenon may happen because these genres have been popular since the early times of Hot 100, dominating the charts between the 1960s and the 1980s (see Figure 1.1). Consequently, artists belonging to these genres have a longer and more established career, increasing the probability of hot streaks. In contrast, as *pop* and *rap* emerged later in the 1980s, artists have shorter timelines to be considered in our model, reducing the occurrence of such periods.

Next, we analyze the duration of hot streaks (HS). As artists may experience more than one HS, we consider only the longest one for each artist. In Figure 4.10 (right), we examine this information through a Cumulative Distribution Function (CDF), which informs, for a given number of weeks (x-axis), the proportion of artists who have the longest HS with duration up to that value. For instance, almost 50% of *soul* artists have their longest hot streak with at most 104 weeks (two years), while for *pop* this proportion rises to nearly 80%. Therefore, we can conclude that in general, *pop* artists have shorter hot streaks than *soul* ones.

Table 4.1 illustrates our findings, showing the top five artists with the most HS periods, as well as the duration of the longest one. All artists are widely known by

Table 4.1: Top 5 artists with more hot streak periods (HS), considering all music genres. Artists are sorted by the number of HS and the duration of the longest one.

Artist	Genres	HS	Longest HS	Period of the Longest HS
Bruce Springsteen	classic rock, heartland rock, mellow gold, permanent wave, rock, singer-songwriter	8	260 weeks	1983-10-22 to 1988-10-15
Michael Jackson	pop, r&b, soul	8	260 weeks	1991-10-12 to 1996-10-05
Mariah Carey	dance pop, pop, r&b, urban contemporary	7	624 weeks	1989-10-14 to 2001-09-29
The Manhattans	classic soul, disco, funk, motown, philly soul, quiet storm, soul, southern soul, urban contemporary	7	312 weeks	1972-11-04 to 1978-10-28
The Rolling Stones	album rock, british invasion, classic rock, rock	7	260 weeks	1964-11-14 to 1969-11-08

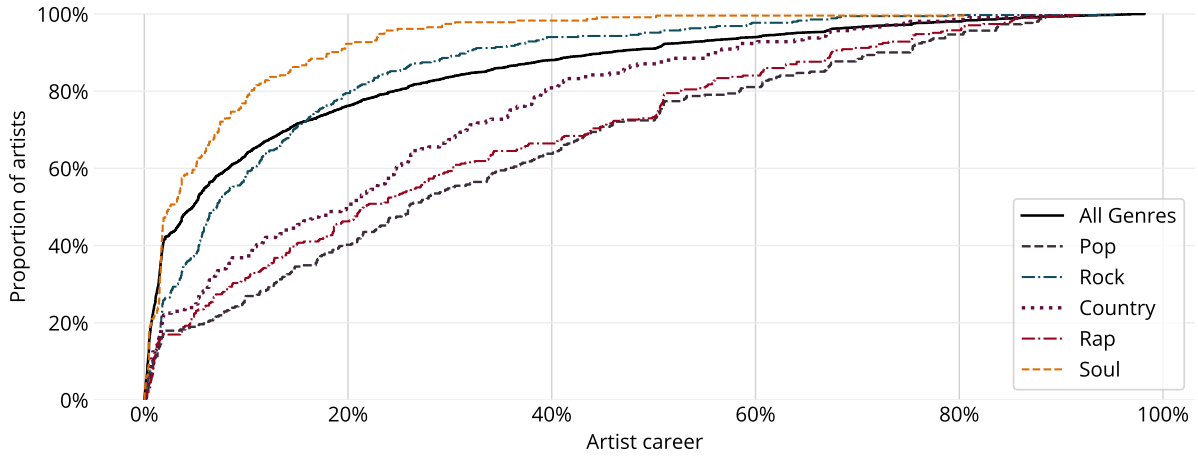


Figure 4.11: Cumulative Distribution Function (CDF) of the location of the first hot streak within artist careers, grouped by genre. Artist timelines are described in percentages, in which 0% represent the debut week and 100% is the last week collected in our dataset.

the audience and recognized by the critic through nominations and wins at the Grammy Awards,[10] the music industry's highest honor. Also, four of them (except for The Manhattans) are in the Billboard's *Greatest Artists of All Time* list.[11] Michael Jackson, known as the king of pop, has eight hot streaks, and the longest one lasts from October 1991 to October 1996. In this period, he released some of their most famous hits, such as *Black or White* and *You Are Not Alone*, which topped the charts. His albums *Dangerous* and *HIStory: Past, Present and Future, Book I* also reached the #1 position of Billboard 200, the magazine's album parade.

The last aspect analyzed is the position of hot streak periods within artist careers. Previous studies on other domains show that such periods are temporally localized and happen at any point of an individual's sequence of works [65, 101]. Moreover, as we detect

[10]Recording Academy Grammy Awards: https://www.grammy.com/
[11]Greatest of All Time: https://www.billboard.com/charts/greatest-of-all-time-artists

more than one hot streak for several artists, we choose to focus only on the first one. Thus, we can investigate at which point of their careers artists experience their first stardom. Figure 4.11 shows the cumulative distribution of the location of the first hot streak for artists from selected genres. In general, almost 80% of the artists have their first burst of success early in their careers (i.e., in the first 20% of their timelines). Nonetheless, this percentage is higher for *soul* and *rock* artists, reaching 90% and 85% of the artists, respectively.

Finally, it is important to note that artists may have careers of different sizes depending on their debut date, as the last date in the time series is always the same (i.e., the collection date). However, as music genres are in constant evolution, new artists emerge in all of them and coexist with the most experienced ones, and thus we believe that our cross-genre comparison is still valid. Overall, the characterization of hot streaks and the genre-aware analysis indicate that music genres are indeed a relevant feature to understand artist careers. Hence, we answer our third research question (*RQ3*) by providing evidence that there are specific patterns for hot streaks when considering different music genres.

4.2.3.2 Impact Around Hot Streaks

After detecting hot streak patterns within artists' careers, we move on to the following research question (*RQ4*), which aims to investigate what happens around a hot streak. That is, we look at the periods before and after the hot streak itself. Once more, as artists may experience more than one hot streak in their careers, we now choose only the longest one. Following the methodology of Garimella and West [38], we consider periods of equal length before and after each hot streak in our analysis. In other words, for a given artist with a hot streak of n weeks, we consider the n weeks before and after such a period. We continue to look at the five music genres from the previous section, but our results can be easily extended to the other genres from our dataset.

We first analyze the number of songs around hot streaks for each genre. As every hot streak is composed of several weeks, we aggregate the number of songs in the charts using the mean value. We do the same for the periods before and after a hot streak. Figure 4.12 presents the average number of songs for artists before, during, and after a hot streak. In all cases, artists from all genres seem to have more songs on the charts during a hot streak, as expected. Such a metric is directly related to success because the more songs an artist has on the Hot 100, the higher the probability of achieving real success. Moreover, *pop* and *rock* artists manage to have more songs on the chart during a hot streak when compared to other genres.

Although the 95% confidence interval (CI) provides a certain level of robustness to our results, we perform statistical tests to ensure them. First, we execute ANOVA (Analysis of Variance) [45] to compare the averages of the three groups (i.e., before,

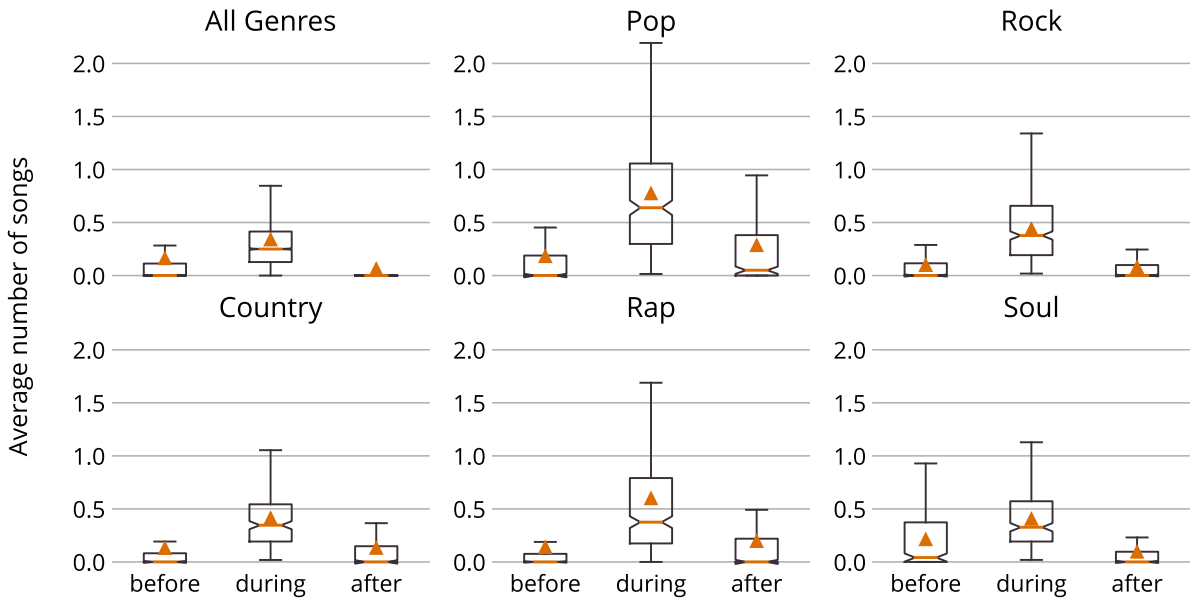


Figure 4.12: Average number of songs before, during and after the longest hot streak of artists from selected genres. Notches represent the 95% confidence interval (CI) around the median (orange line). The orange triangle is the mean value.

Table 4.2: Pairwise comparison of the average number of songs for selected genres around a hot streak (before, during, and after) using Tukey's HSD test. Cross-marked values indicate the difference is not statistically significant ($p \geq 0.05$).

	before vs. after	**before vs. during**	**during vs. after**
All	higher	lower	higher
Pop	lower	lower	higher
Rock	higher	lower	higher
Country	×	lower	higher
Rap	×	lower	higher
Soul	higher	lower	higher

during, and after) by testing a null hypothesis that all means are equal. In our case, it is rejected for all music genres with $p < 0.05$, meaning the differences are indeed significant. Nonetheless, ANOVA does not tell which periods are significantly different from each other, and thus the next step is to perform a pairwise evaluation to verify the statistical significance of each one. We do so by conducting Tukey's honestly significantly differenced (HSD) test [108] as a post-hoc comparison.

Table 4.2 summarizes the results for the Tukey's HSD test, and except for the comparison between *before* and *after* for *country* and *rap*, all pairwise comparisons indicate statistically significant differences ($p < 0.05$). Therefore, we can now affirm that the average number of songs *during* a hot streak is statistically higher when compared with the periods *before* and *after*. Furthermore, considering all artists and also for *rock* and *soul* artists, the average number of songs is higher before a hot streak than after, indicating that there may be a growing success trend in artists' careers before the stardom. However,

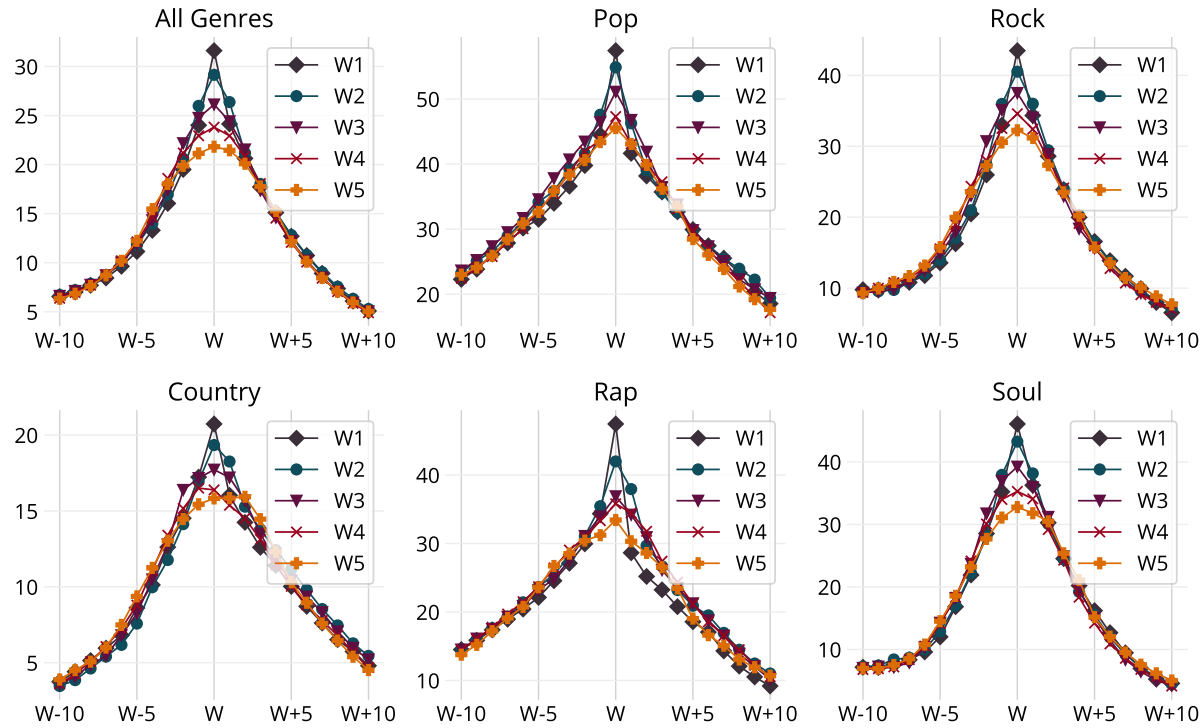


Figure 4.13: Success around the five most impactful weeks (W1-W5) for artists from selected genres, measured by Rank Score DCG. Note: the y-axis varies according to the genre.

we observe the opposite for *pop* artists, as they have more songs after a hot streak than before it. This may be a piece of evidence that artists from such a specific genre behave in a different way and manage to maintain success for longer periods, confirming the findings from Section 4.2.3.1.

Impact around the most successful weeks. We also look at the impact around specific weeks to understand the success dynamics of a specific point of an artist's career. Here, we analyze the ten weeks before and after the five most impactful weeks (i.e., with the highest Rank Score DCG), denoted by $W1, \cdots, W5$. Figure 4.13 shows the curves for the selected genres, and in all of them, we observe a growing success trend before the considered week, as pointed out in the previous analysis. Similarly, the success after such a week does not disappear immediately, reinforcing an effect observed in the charts in which a song continually loses positions in the charts over the weeks. Hence, the set of analyses provided in this section shed light on what happens around a hot streak, answering properly *RQ4*.

Table 4.3: Classification evaluation results. Metric values are presented with a 95% confidence interval (CI) and bold values indicate that a classifier is statistically better for that metric.

	LR	LinearSVC	Perceptron	SGD
Accuracy	**0.90993 ± 0.00001**	0.90791 ± 0.00005	0.88236 ± 0.01954	0.90505 ± 0.00190
Precision	0.73536 ± 0.00004	0.72764 ± 0.00015	0.67466 ± 0.07060	0.73035 ± 0.00508
Recall	0.78935 ± 0.00004	0.78977 ± 0.00030	0.77301 ± 0.09725	0.75886 ± 0.01923
F1 Score	**0.76140 ± 0.00002**	0.75743 ± 0.00015	0.70383 ± 0.03769	0.74405 ± 0.00828
F1 Weighted	**0.91116 ± 0.00001**	0.90935 ± 0.00005	0.88551 ± 0.01623	0.90572 ± 0.00229

4.2.4 Hot Streak Prediction

After identifying and characterizing hot streaks, the next step is to verify whether such periods are predictable for genres or not. We do so by using Music-oriented Hot Streak Binary Classification (MHSBC) from Section 4.1.4. The experimental evaluation of our model is presented in Section 4.2.4.1, whereas Section 4.2.4.2 contains the factor analysis behind predictions.

4.2.4.1 Experimental Evaluation

From the problem definition and the experimental setup defined in Section 4.1.4, we now evaluate the MHSBC results to answer *RQ5*. For a fair comparison, we run the chosen classifiers separately. We also execute each algorithm ten times, varying the *random state* parameter to govern the method's random choices. Thus, we can get a confidence interval for the evaluation metrics. As mentioned in the previous section, the train-test split follows a chronological order, meaning that we test the classifiers using unseen data (i.e., whether a week in 2003 is part of a hot streak period for a given genre based on data up until 2002).

Table 4.3 presents the results for all classifiers considered in this study. All of them outperform the baseline classifier, which predicts the most frequent class ($accuracy = 0.850$). Thus, our model reveals to be better than simply guessing, as all classifiers have higher accuracy values. The baseline does not provide F1-score, as it does not make predictions (i.e., it simply returns the majority class for every instance). Considering the four selected classifiers, Linear Regression (LR) is the one with the best results, with an average accuracy of 0.910 and an average F1-score of 0.761, which are significantly higher (95% CI) than the other algorithms. Hence, we choose such a classifier as the best one for MHSBC.

To strengthen our results, we also evaluate the four classifiers using the area under the Receiver Operating Characteristic (ROC) curve, which is also a widely used perfor-

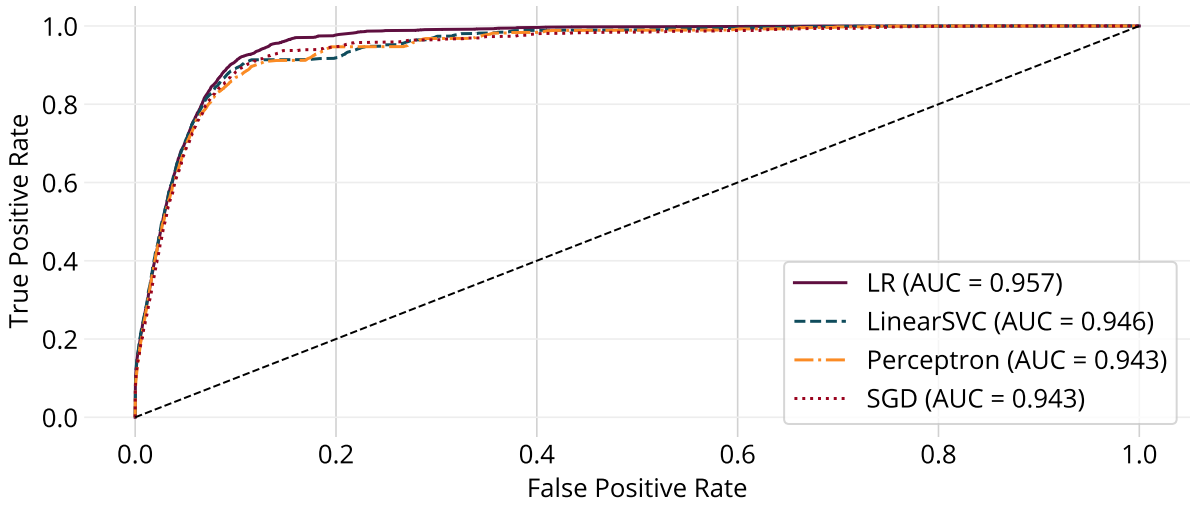


Figure 4.14: ROC curves for selected classifiers with their area under the curve (AUC) values. The black dashed line represents a random classifier.

mance measure for classification problems. ROC is a probability curve based on the False Positive Rate (x-axis) against the True Positive Rate (y-axis), and the area under this curve (AUC) tells how good the classifier is at distinguishing between classes. In our case, the higher the AUC, the better the classifier is at predicting hot streak and non-hot streak weeks correctly. The AUC values for our classifiers are presented in Figure 4.14. In this evaluation, Logistic Regression is also the best classifier, outperforming the other methods with $AUC = 0.957$.

Therefore, we are now able to answer *RQ5* as it is indeed possible to predict, for a given week, if it belongs to a hot streak period for music genres. Such a result is highly relevant for both record labels and artists, as it may guide them in the definition of future partnerships and collaborations. Choosing featured artists from genres within a hot streak period may attract a bunch of new listeners and therefore increase sales and streaming numbers. Next, we investigate which factors are the most important in the hot streak prediction.

4.2.4.2 Feature Importance

As seen in the previous section, using machine learning methods for hot streak prediction produces results with high accuracy. However, understanding why and how a model makes a certain prediction can be as crucial as the outcome itself, shedding light on the "black box" within the learning algorithms. In this section, we use SHAP (SHapley Additive exPlanations) [67] in our classification model to allow its interpretability. In short, SHAP is a game-theoretic approach to explain the output of a machine learning model, assigning for each feature an importance value for a particular prediction. From a global perspective, the importance values can be aggregated to show how much each

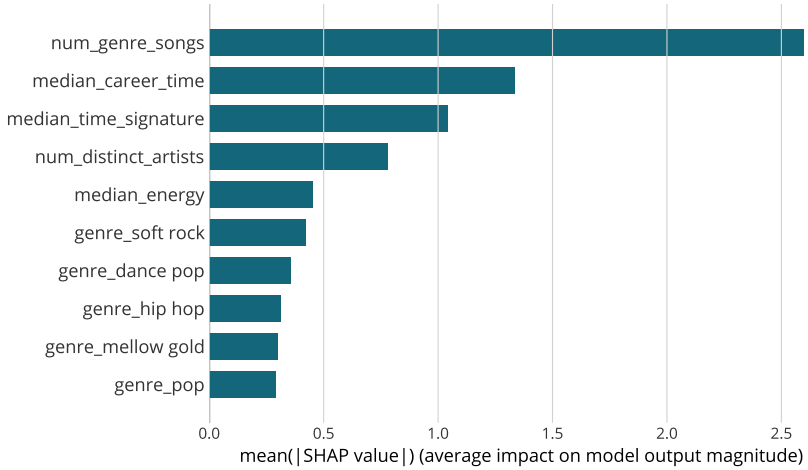


Figure 4.15: Features with the highest absolute mean SHAP values.

predictor contributes to the target variable, either positively or negatively.

First, we analyze the features with the highest absolute mean SHAP values. Figure 4.15 is a summary plot with the top 10 features with the highest impact on hot streak prediction. The result shows features such as the number of songs, median time signature, median career time, and the number of distinct artists are the most significant to our model, thus having high predictive power. The descriptive genre features obtained from the binarization conducted in Section 4.1.4.2 (e.g., genre_pop and genre_rap) also appear in the ranking, but their average SHAP values are close to zero, requiring further investigation.

Next, we examine the positive and negative relationships of the predictors with the target variable. Figure 4.16 goes further in the summary plot, using SHAP values to show the distribution of the impact of each feature in the model output. Features are ranked by their mean absolute SHAP value and each point on the x-axis tells if the effect of that value is associated with a higher or lower prediction. The color scale informs whether the feature value is high (red) or low (teal) for that instance. For example, higher amounts of songs and artists impact positively on the prediction, that is, they are related to the presence of hot streaks. Such a relation is reasonable since the more artists and songs from a genre are present in a weekly chart, the higher the probability of the week belonging to a hot streak.

The median time signature (an acoustic feature obtained from Spotify which specifies how many beats are in each bar) has also a positive impact on hot streak weeks. Thus, high loads for such a metric can raise hot streak prediction. On the other hand, high values for median career time impact negatively the prediction, i.e., the higher the median career for artists, the lower the probability for a week to be within a hot streak. Overall, the aforementioned features have a significant impact on the prediction, and therefore they are the main factors driving hot streak periods. Such results answer our

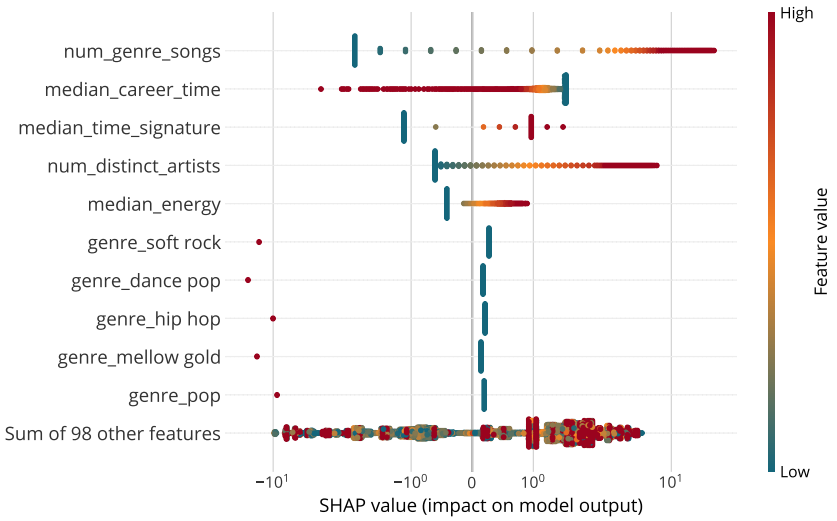


Figure 4.16: SHAP values. Features are sorted by the sum of SHAP value magnitudes over all samples. The color represents the feature values (red high, teal low).

last research question (*RQ6*), as we are able to detect factors that do influence (either positively or negatively) hot streaks.

4.3 Overall Considerations

In this chapter, we used data from music charts to analyze musical careers (i.e., artists and genres). Based on data from Billboard Hot 100, the main song ranking in the United States, we built time series that represent success on weekly charts. We measured success by aggregating the artists' song positions in each chart. Therefore, our goal was to investigate hot streaks in such careers, which correspond to continuous periods with success above the average observed until then. We point out our contributions by answering our six research questions (RQs) as follows.

RQ1. How do the most impactful weeks in musical careers are distributed over time? From the success time series for artists and musical genres, we found that the most successful weeks are clustered in time (Section 4.2.1).

RQ2. Does this behavior generalizes into continuous periods of high impact (i.e., hot streaks)? Yes. We detected hot streaks in musical careers by using the Piecewise Aggregate Approximation (PAA), which is a method for reducing the dimensionality of time series (Section 4.2.2).

RQ3. Are there specific hot streak patterns for distinct musical genres? Yes. Our characterization analysis revealed general and specific patterns of hot streaks for

artists of different genres. We evaluated characteristics such as quantity, duration and appearance of the first hot streaks (Section 4.2.3.1).

RQ4. What happens before, during and after a hot streak period? Overall, we found that artists have more songs on the charts during periods of hot streaks. Besides, the career peaks appear and disappear gradually over time (Section 4.2.3.2).

RQ5. Is it possible to predict whether a week belongs to a hot streak period? Yes. We proposed the Music-oriented Hot Streak Binary Classification (MHSBC) as a model to assess the hot streak prediction task. Our findings revealed that our model was successful with an F1-score of 0.761. (Section 4.2.4.1).

RQ6. What are the factors that influence hot streak periods? We used SHAP values to detect features that increase the predictive power of our model. We identified that factors such as the number of songs present in the charts and the artists' career time are relevant to the classifier, as well as the *time signature* and *energy* acoustic features (Section 4.2.4.2).

Overall, our findings represent a step further in the science behind musical success, as we observe the temporal evolution of artists' careers and their success. Being able to understand the dynamics around hot streak periods and also predict their occurrence is relevant not only to the scientific community but to the music industry as a whole. For the first, it may contribute to the development of more complex models, while for the latter it helps to describe the listeners' behavior and success trends over artists and genres. Therefore, both musicians and record labels may orientate their future releases to achieve or maintain their success levels. In short, the real value of identifying hot streaks is in revealing the fundamental patterns that govern individual careers.

Limitations. Our study has some limitations that can affect the coverage of our results. First, as we built our time series based on Billboard Hot 100, we consider only data from the American market. Therefore, our results may not be generalized to other markets, as they have specific regional factors that shape musical success. Also, the artists' musical genres obtained from Spotify do not follow specific patterns, and thus we deal with overspecialized genres (e.g., Texas hip hop) that may blur our results. Finally, as we split data temporally in the classification task, the algorithm may not perform well for genres with few hot streak instances in the training set.

Chapter 5

Detecting Collaboration Profiles in Success-based Music Genre Networks

In the previous chapter, we assessed musical careers by analyzing hot streaks of individual artists and bands. Such periods correspond to phases in which the individual success is above-normal. Genre appeared as a relevant feature, as there are different patterns of hot streaks according to the genre. In this chapter, we go up one level of abstraction and focus on music genres and their relation to success. Specifically, we add a new dimension to our analyses by considering the collaborations that connect different genres. Therefore, our hypothesis is that success is not only related to the performance on charts, but also to the genre connections that make hit songs.

Indeed, musicians teaming up is nothing new but has risen far beyond the norm. Remaining an industry of creative growth, it is only natural for music (i.e., all musical scene members) adapting to new conditions and redefining its layout. Through cross-genre collaboration, artists are naturally venturing into new domains and working outside of the category to which they had originally been ascribed to. Such a collaboration phenomenon may be reshaping music global environment, by challenging segments of certain genres to come up with something entirely new [100]. Moreover, this gradual evolution is becoming a driving force in creating a more collaborative scenario, making music one of the most innovative art forms.

As this creative market changes, it becomes more unpredictable; and doing both predictive and diagnostic analyses in such a context remains challenging. Still, we believe factors leading to an ideal musical partnership can be understood by exploring collaboration patterns that directly impact its success [100, 21, 12]. Hence, in this chapter, we aim to unveil the dynamics of cross-genre connections and collaboration profiles in success-based networks (i.e., connections formed by genres of artists who cooperate and create hit songs). We include other music markets in our analyses, as only considering data from the United States may not represent the global music scenario (as seen in Chapter 4). Therefore, we organize our study with the following research questions (RQs).

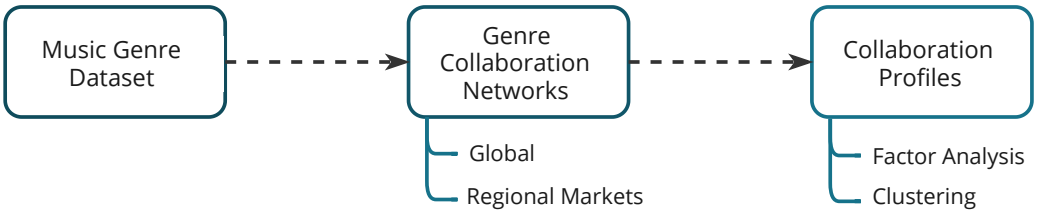


Figure 5.1: Proposed methodology to uncover collaboration profiles in genre networks.

RQ1. Does the regional aspect impact on popular genres and their hit songs?

RQ2. How has genre collaboration evolved over the past few years?

RQ3. Which are the potentially intrinsic factors and indicators that influence the collaboration success?

We answer such research questions through the remainder of this chapter, which is organized as follows. Section 5.1 details our methodology to build the genre networks and to find the collaboration profiles. Then, Section 5.2 presents and discusses the results of experiments over genre collaboration profiles and their relation to musical success. Finally, we address overall considerations in Section 5.3.

5.1 Methodology

This section presents the methodology proposed to answering our three research questions. First, we build a novel dataset with enhanced genre information using data from Spotify (Section 5.1.1). Then, from such data, we model genre collaboration networks for both regional and global music markets (Section 5.1.2). Finally, we use network metrics to uncover collaboration profiles within the genre networks (Section 5.1.3). The methodology steps are summarized in Figure 5.1.

5.1.1 Music Genre Dataset

Over recent years, the world has seen a dramatic change in the way people consume music, moving from physical records to streaming services. Since 2017, such services have become the main source of revenue within the global recorded music market. Thus, as in

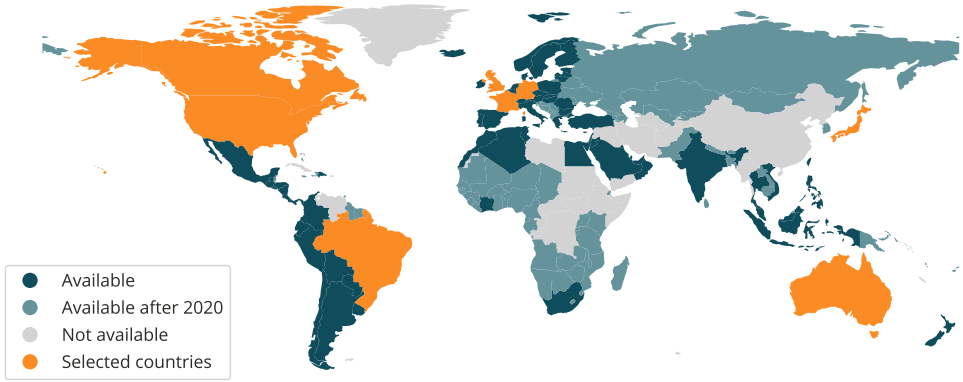


Figure 5.2: Spotify presence worldwide at the time of data collection (May 2020).

Chapter 4, we build our dataset by using data from Spotify. It provides a weekly chart of the 200 most streamed songs for each country and territory it is present, as well as an aggregated global chart.

Considering that countries behave differently when it comes to musical tastes, we collect global and regional charts from January 2017 to December 2019, considering eight of the top 10 music markets according to IFPI: United States (1st), Japan[1] (2nd), United Kingdom (3rd), Germany (4th), France (5th), Canada (8th), Australia (9th), and Brazil (10th). Data from South Korea (6th) and China (7th) were not available in Spotify as of May 2020 (collection date).[2] Figure 5.2 illustrates the presence of Spotify worldwide and the selected markets. We also use Spotify API[3] to gather information about the hit songs and artists present in the charts, such as all collaborating artists within a song (since the charts only provide the main ones) and their respective genres, which is the core of this work. Our final dataset contains 1,370 charts from 156 weeks, comprising 13,880 hit songs and 3,612 artists from 896 different music genres.

Then, we perform a processing phase on the artists' genres, because Spotify assigns a list of very specific genres to each artist. In most cases, artists' genres present a high degree of detail, which may overcomplicate our analyses. For example, Jay-Z (one of the most popular rappers in the United States) is assigned to both *east coast hip hop* and *hip hop* genres, which may be described only by *hip hop*. To simplify our modeling and further analyses, we choose to map all specific genres to more embracing and well-established *super-genres*. Note that the regional aspects are not lost in such a mapping, because our analyses are made separately for each considered market. Hence, the 896 existing genres are now mapped into 162 *super-genres*. The dataset is publicly available

[1]The first Japanese weekly chart is from August 31, 2017.

[2]Due to the COVID-19 pandemic (which started to be globally acknowledged in February 2020), we have decided not to update such a dataset to 2021. We prefer to keep most of it without the potential changes and bias introduced into streaming consumption due to the altered pandemic routine.

[3]Spotify API: https://developer.spotify.com/

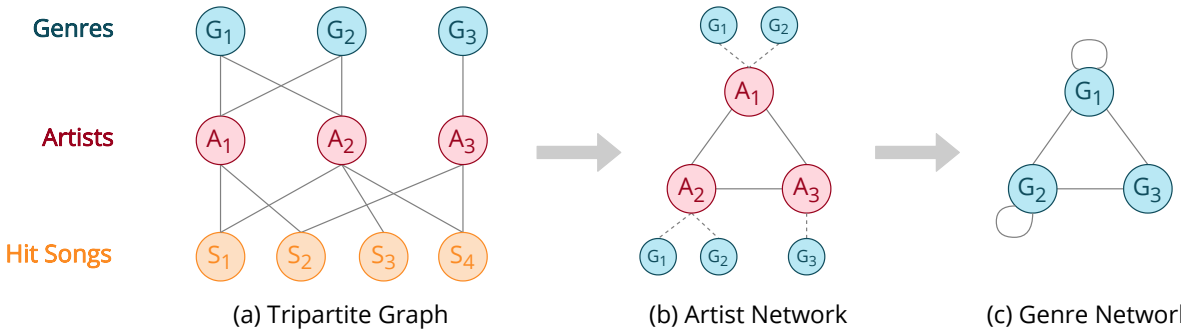


Figure 5.3: Reduction from the tripartite (a) to the one-mode Genre Collaboration Network (c). The intermediate step is an Artist Network with genre information (b). Artists and genres are linked when hit songs involve both nodes.

on our project page[4] and exploratory analyses on it provide relevant information about hit songs in music markets, answering then *RQ1*.

5.1.2 Genre Collaboration Network

To answer *RQ2*, we model genre collaboration using the Network Science framework.[5] A Collaboration Network is usually modeled as a graph formed by nodes (vertices) that may be connected through edges. For example, nodes represent artists and are connected by an edge if the respective pair of artists has collaborated in a song. Now, to analyze the interactions between genres, we model music collaboration as a tripartite graph, in which nodes are divided into three sets: genres, artists, and hit songs; i.e., the minimum elements to evaluate success. The building process of the genre network from the tripartite model is illustrated in Figure 5.3(a). Collaborative hit songs are sung by two or more artists, regardless of their participation (e.g., a typical *feat.* or a duet[6]). We also equally consider all genres linked to an artist because they shape how such an artist is seen by fans and the music industry.

The analyses and characterization of multipartite graphs is often a complex task, as most of connectivity metrics and algorithms are not properly defined for this type of graph. Thus, for directly analyzing the interaction between musical genres, we reduce the tripartite model into a one-mode network in which nodes are exclusively genres. However, such a reduction is only possible by executing an intermediate step: building the artist collaboration network, which is presented in Figure 5.3(b). In such a network, two artists

[4]Project Bàde: https://bit.ly/proj-Bade
[5]For formal definitions and more details on Network Science, see Chapter 3.
[6]The main types of collaboration include *featuring* (two artists collaborating on the same song), *and* (artists who share equal rights), *vs.* (DJ contest), and *with* (duet).

are connected when both collaborate in one or more hit songs. The genre information is not lost, as it is linked directly to the artists.

We may now build the final one-mode genre network by connecting the genres of artists who collaborate in the artist network. The edges are undirected and weighted by the number of hit songs involving artists from both genres, as illustrated by Figure 5.3(c). In addition, self-loop edges are allowed in our modeling, as there are several hit songs from artists of the same genre (intra-genre collaboration). For example, the song *Old Town Road*[7] by Lil Nas X and Billy Ray Cyrus generates an edge between these artists in the intermediate network; and each of Lil Nas X's genres (*pop rap, country pop* and *hip hop*) is linked to Cyrus' only genre (*country*) with weight 1.

5.1.3 Collaboration Profiling

This section presents our approach to uncover significant factors that compose a successful music genre collaboration. Inspired by Silva et al. [100], we first extract information from the success-based networks by evaluating six edge-dependent metrics. We perform an Exploratory Factor Analysis on such metrics to define factors, and then perform a cluster analysis to uncover collaboration profiles to investigate the key driving factors on successful collaborations and then answer *RQ*3.

Exploratory Factor Analysis (EFA). EFA is a statistical method designed to underline patterns of correlations among observed variables and extract latent factors [28]. Generally, EFA identifies the number of common factors and the pattern of factor loadings (correlations). It assumes and asserts that manifest (observed) variables are expressed as a linear combination of factors and measurement errors. Each factor explains a particular variance in the variables and may find hidden data patterns. Besides, EFA is largely used by data scientists to better interpret the results, as it reduces the number of analyzed variables. In our model, we run EFA to evaluate the following network edge metrics: Weight, Common Neighbors, Neighborhood Overlap, Preferential Attachment, Edge Betweenness, and Resource Allocation.[8]

There are two main issues when executing an EFA: *(i)* determining the number of factors to retain for analysis, and *(ii)* selecting the final structure for how the measured variables relate to the factors. For the former, we use the Parallel Analysis criteria [51], which is based on random data simulation. The suggested number of factors to extract is

[7] #1 Song of 2019 according to Billboard Year-End Hot 100 Chart: https://billboard.com/charts/year-end/2019/hot-100-songs

[8] See Chapter 3 for definitions

then provided and based on examining the *scree plot* [23] of factors of the observed data with that of a random data matrix of the same size as the original. Finally, the EFA is performed using the well known Ordinary Least Squares (OLS) factoring method and an oblique rotation, allowing factors to correlate with each other.

Cluster Analysis. The second step of our approach is to perform a cluster analysis to group similar music genre connections based on the aforementioned factors. We use DBSCAN [33] as a clustering algorithm, which assigns data points to the same cluster if they are *density-reachable* from each other. We choose such an algorithm since it supports outlier detection and does not require a predefined number of clusters. Thus, two parameters are required to run DBSCAN: ϵ defines the radius of neighborhood around a point x; and $MinPts$ (minimum points) is the minimum number of neighbors within the ϵ radius.

To choose the optimal ϵ value, we use a method based on k-nearest neighbor distances, which calculates the average of the distances of every data point to its k nearest neighbors. In general, the value of k is specified by the user and corresponds to the $MinPts$ parameter. As a general rule, the $MinPts$ can be derived from the number of dimensions D in the dataset as $MinPts \geq D + 1$. Since there are six topological metrics, we set $MinPts = 7$. Therefore, we set the ϵ value as the k value in which there is a sharp change in the curve of the k-distances.

5.2 Results and Evaluation

Following the steps of our methodology presented in the previous section, we now perform an experimental evaluation to answer the proposed research questions. First, we examine Spotify charts from our dataset for each market and year to detect popular genres (*RQ1*, Section 5.2.1). Then, we characterize the genre collaboration network to understand the evolution of each market (*RQ2*, Section 5.2.2). Finally, to answer *RQ3*, we perform a two-step approach to find intrinsic elements behind collaboration success: factor analysis (Section 5.2.3) and collaboration profile detection (Section 5.2.4).

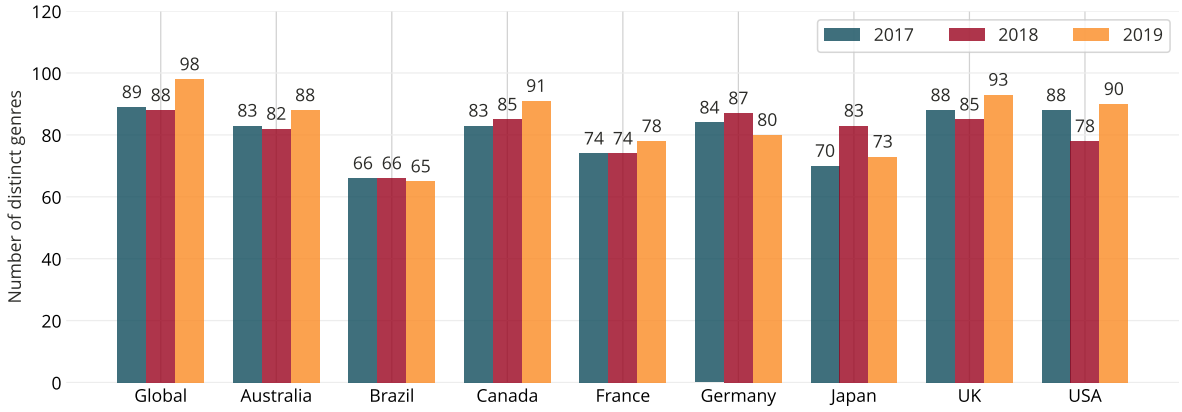


Figure 5.4: Number of distinct genres from Spotify charts for each market (2017-2019).

Table 5.1: Most popular music genres in each considered market from 2017 to 2019.

	Genre	Songs	Arts.		Genre	Songs	Arts.		Genre	Songs	Arts.
	pop	1,715	424		pop	1,790	402		pop	1,772	371
	hip hop	1,192	281		rap	1,762	209		hip hop	1,138	232
Global	rap	1,184	195	*Canada*	hip hop	1,511	232	*UK*	rap	974	167
	pop rap	845	130		pop rap	1,355	149		dance pop	922	178
	dance pop	832	165		trap	1,139	154		pop rap	660	120
	pop	1,646	411		hip hop	2,604	352		j-pop	797	163
	rap	873	165		pop	1,665	479		pop	705	210
Australia	dance pop	822	171	*Germany*	rap	1,223	205	*Japan*	dance pop	438	103
	hip hop	792	191		dance pop	650	159		j-rock	312	72
	pop rap	657	133		pop rap	462	109		r&b	276	82
	pop	1,072	256		pop	3,138	470		rap	2,057	231
	sertanejo	565	82		hip hop	2,660	285		hip hop	1,719	241
Brazil	brazilian funk	559	156	*France*	rap	2,526	245	*USA*	pop	1,704	340
	dance pop	415	101		francoton	1,097	82		pop rap	1,518	139
	electro	307	93		dance pop	391	119		trap	1,370	172

5.2.1 Musical Genres Overview

To answer *RQ1*, we perform an exploratory analysis of the charts from nine markets (global and eight countries) over three years. We first take an overview of the diversity of genres from each market. Figure 5.4 illustrates the number of distinct music genres extracted from the artists who perform the hit songs. Including the global scenario, most of the markets (six out of nine) present an increasing number of genres, suggesting that listeners from such markets are becoming more open to new music styles. Regarding global charts, they present the highest number of distinct genres (98 in 2019), which is expected since they aggregate data from all markets covered by Spotify. However, Brazil, Germany, and Japan present a lower number of distinct genres than their peak in 2018, indicating a behavior different from the rest. They are not English-speaking countries, which may contribute to the development of specific regional music ecosystems.

Next, we focus on music preferences by investigating the most popular genres in

Table 5.2: Network characterization for global and three regional markets, representing the groups of countries with similar network evolution. Underlined values are the highest metric value for a specific market throughout the considered period.

Metric	*Global*			*USA (Group 1)*			*Brazil (Group 2)*			*UK (Group 3)*		
	2017	**2018**	**2019**	**2017**	**2018**	**2019**	**2017**	**2018**	**2019**	**2017**	**2018**	**2019**
Genres (nodes)	72	79	89	76	73	83	58	63	61	74	76	79
Collaborations (edges)	564	583	709	542	522	670	453	524	392	610	605	627
Avg. degree	15.7	14.8	15.9	14.3	14.3	16.1	15.6	16.6	12.9	16.5	15.9	15.9
Avg. weighted degree	256.9	247.4	236.7	324.6	287.9	241.4	136.1	133.0	95.3	216.5	203.6	159.5
Density	0.221	0.189	0.181	0.190	0.199	0.197	0.274	0.268	0.214	0.226	0.212	0.204
Avg. Clustering Coef.	0.743	0.757	0.754	0.762	0.760	0.726	0.770	0.758	0.677	0.724	0.754	0.738
Self-loops	24	21	28	25	22	27	24	29	27	28	25	30

each market. Table 5.1 presents the five most popular genres in terms of hits songs. Indeed, each country has its own musical inclinations, although the predominant genres are mostly *pop*/*pop rap*, *hip-hop*, and *rap*. Such preference may be due to the increasing number of collaboration songs among artists from different musical genres, as revealed in Figure 1.2: growing collaborations of *pop*, *rap*, *hip-hop*, and *r&b* in recent years. Also, except for *r&b*, they are the main genres at the top-5 genre lists on most markets; i.e., such genres are among both the most collaborative ones and the most listened on the globe. Moreover, there are three markets with local genres on their top-5 list: Brazil with *sertanejo* and *brazilian funk*; France with *francoton*; and Japan with *j-pop* and *j-rock*. Although local, such genres are potentially good choices for record companies to encourage musical collaborations. Note local engagement shapes the global environment, ensuring that music culture within such countries can develop and progress. In fact, some local genres are crossing such a frontier and becoming extremely popular worldwide, as *hip hop* in the 1990s and *reggaeton* more recently.

5.2.2 Network Characterization

After analyzing the charts, we build the genre collaboration network for each market and year to find out how genres connect to answer *RQ*2. With nine markets (global and eight countries) during three years, we analyze 27 networks.[9] For each network, we calculate basic statistics on its nodes and edges, as well as structural metrics (Chapter 3). Such statistics are useful to describe the dynamics of the networks (i.e., genre connections) and may produce insights about the state of each music market. Table 5.2 shows the network characterization for selected markets.

First, the global genre networks reveal the world is more open to new successful

[9] All networks can be visualized in https://bit.ly/proj-Bade

genres (number of nodes/genres growth). Also, the number of genre connections (edges) increased considerably, meaning more collaborative hit songs are coming from artists whose genres are not linked in prior networks, opening up new opportunities for those genres to acquire new listeners. The networks average degree remains stable, while its weighted value decreases over the years. This could reveal a growth in the number of collaborations of well-established genres with emerging ones, represented by edges with low weight values (few hit songs). Still, such low-degree emerging genres may become popular shortly, expanding their collaborations to other unexplored genres. For instance, *k-pop* connections double as it spreads worldwide, approaching genres such as *reggaeton* (e.g., the collaboration between J-Hope from BTS and Becky G in the song *Chicken Noodle Soup*, September 2019).

For regional markets, we classify the countries into three groups, according to the similarities in networks' evolution: *(i)* USA and Canada; *(ii)* Brazil, France, Germany and Japan; *(iii)* UK and Australia. As the global network, countries in the first group have an increasing average degree and a decreasing average clustering coefficient, thus indicating a stronger tendency to diversify the inter-genre collaborations. Then, the second group includes non-English speaking countries with decreasing connectivity metrics in 2019, after a peak in 2018. The last group has countries in which more genres are becoming successful, while the connections are not increasing in the same proportion. Such country groups reinforce the insights from Section 5.2.1, in which some countries presented similar behavior regarding popular music genres.

Overall, considering regional markets individually becomes important for producers and record labels, as they are delivering more global hits over time. Their distinct behavior emphasizes the strength of cultural aspects on determining how music is consumed and the success of a given genre or artist. The similarities in genre networks revealed three distinct groups of countries, which share a cultural and/or geographical proximity. Therefore, in each market, genre connections may reveal distinct profiles, which are an important tool for analyzing successful genre collaborations.

5.2.3 Factors Behind Collaborations

In this section, we address the first step of the methodology to uncover collaboration profiles in the genre networks (*RQ3*, Section 5.1.3). Here, we use Exploratory Factor Analysis (EFA) to identify the common factors and the relationships among the edge-dependent metrics of all 27 success-based networks: Weight (W), Neighborhood Overlap (NO), Common Neighbors (CN), Preferential Attachment (PA), Resource Allocation

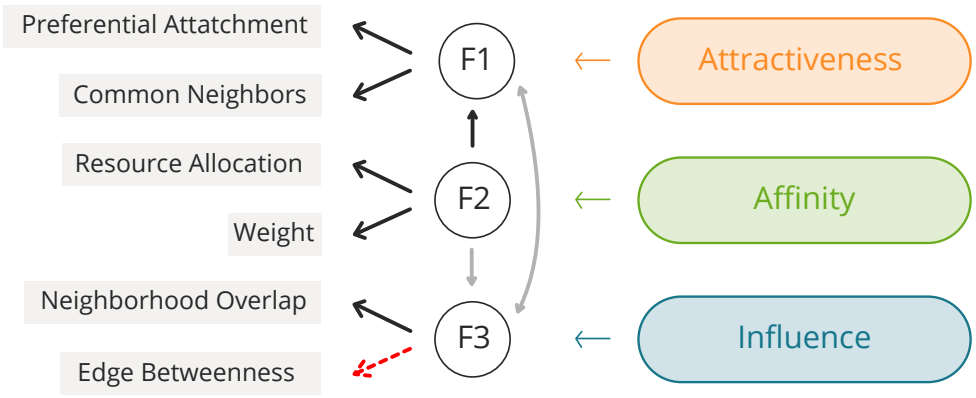


Figure 5.5: Factor Analysis diagram. Solid and red dashed lines represent positive and negative correlations, respectively. Dark and lighter lines represent strong $[0.6 - 1.0]$ and weak $[< 0.6]$ correlations, respectively.

(RA), and Edge Betweenness (EB). Overall, the analysis results suggest a three-factor structure within those six metrics, which helps to analyze the profiles in the next section. A graphical representation of the emerging structure is in Figure 5.5, and the details on the Factor Analysis are in Appendix B. As the three factors are conceptually coherent, we labeled them as follows.

Attractiveness (F1). Factor 1 has high loads for both PA and CN metrics, with a positive correlation between them. Specifically, values close to zero indicate that two nodes are not close and attracted, while higher values indicate closer nodes. Therefore, this factor corresponds to the predisposition of two nodes to connect in the future.

Affinity (F2). Factor 2 has high loads for both RA and W metrics, with a positive correlation between them. High values indicate strong social ties, and lower ones indicate weak ties. Hence, this factor measures both the frequency of collaboration between two nodes and the social strength.

Influence (F3). Factor 3 has high loads for both NO and EB metrics, with a negative correlation between them. Edges with low NO and high EB certainly consist of local bridges in the network. That is, they represent a bridge-like connector between two "social circles". Therefore, this factor corresponds to the importance level of an edge with access to different regions in the network.

5.2.4 Genre Collaboration Profiles

Following the methodology proposed to uncover genre collaboration profiles (*RQ3*), we perform a cluster analysis on the network edges using the DBSCAN algorithm. The

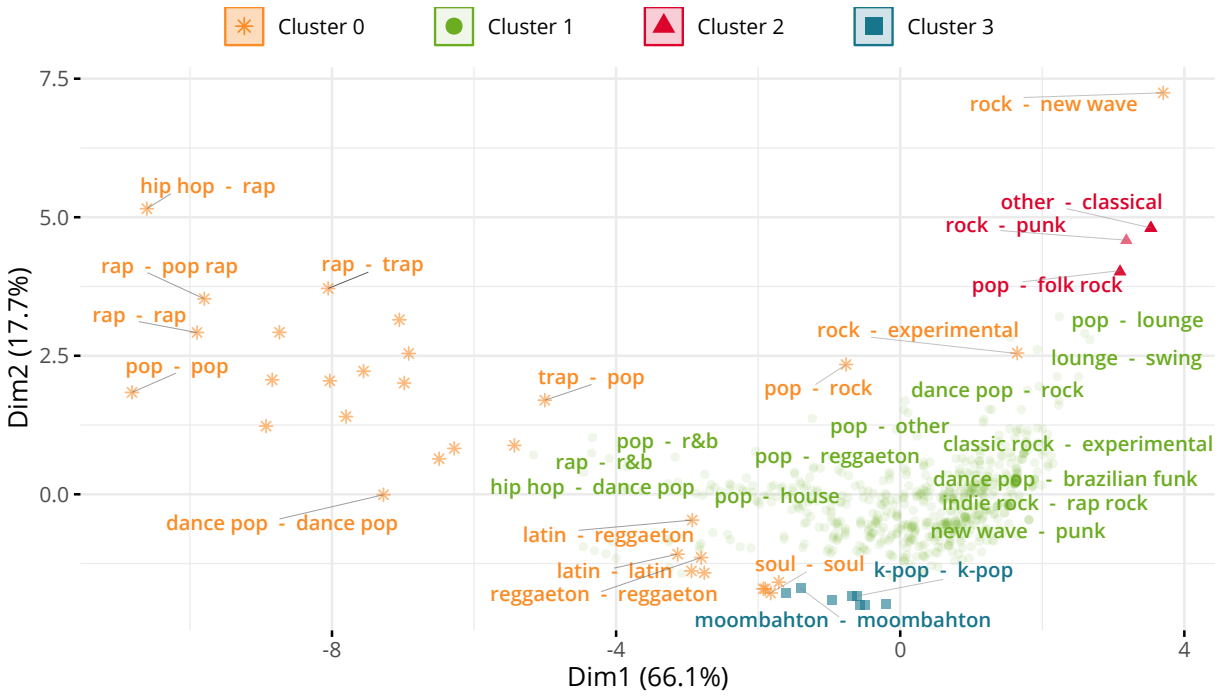


Figure 5.6: Clustering result for the United States network in 2019, with examples of some genre collaborations in each cluster.

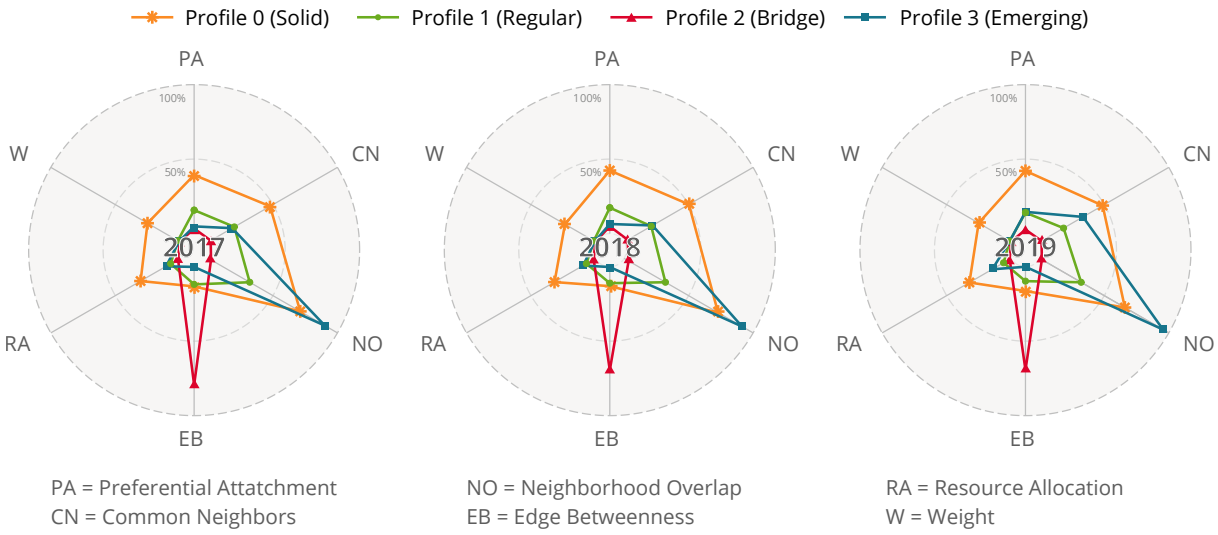


Figure 5.7: Collaboration profiles for all markets (2017-2019).

detailed steps and intermediary results are in Appendix B. Overall, four distinct clusters were detected in at least one of the 27 collaboration networks. As an example including all clusters, Figure 5.6 shows the result of the US network in 2019, where Cluster 0 groups the outliers identified by DBSCAN (data points in low-density regions, i.e., not associated with any proper cluster). Clusters 1 and 2 are slightly overlapping, but each covers groups of high-density data points, which is successful information in this analysis. We can also certainly conclude Cluster 3 is separate from the others. Next, we describe each cluster.

Now that we have detected a set of predominant clusters on all modeled networks,

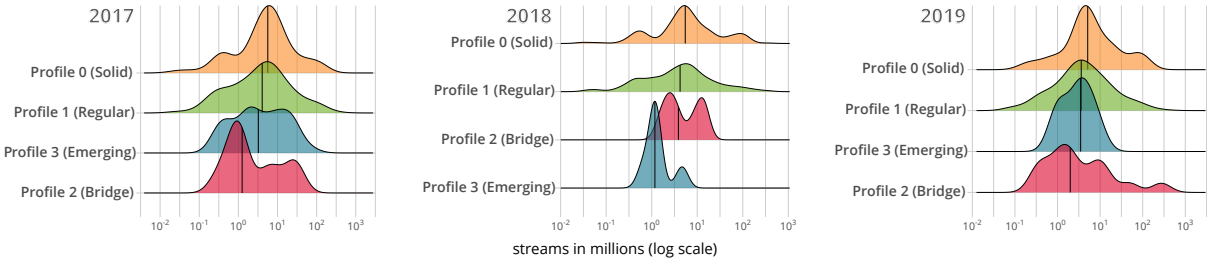


Figure 5.8: Density ridgeline plots of streams in millions for each cluster (log scale). Clusters are sorted by their median stream values (darker vertical lines).

Table 5.3: Total number of *intra*- and *inter*-genre collaborations in each profile.

Collab	*Solid*			*Regular*		
	2017	**2018**	**2019**	**2017**	**2018**	**2019**
Inter-genre	140 (49%)	125 (42%)	103 (51%)	1,828 (99%)	1,916 (98%)	2,165 (94%)
Intra-genre	145 (51%)	174 (58%)	99 (49%)	23 (1%)	34 (2%)	128 (6%)
Collab	*Bridge*			*Emerging*		
	2017	**2018**	**2019**	**2017**	**2018**	**2019**
Inter-genre	10 (100%)	7 (100%)	16 (100%)	3 (7%)	1 (17%)	0 (0%)
Intra-genre	0 (0%)	0 (0%)	0 (0%)	40 (93%)	5 (83%)	7 (100%)

the next step is to look at their characteristics for profiling them and defining proper identities. First, for each network, we calculate the mean of the normalized metrics values grouped by each cluster id. Then, for each year, we plot radar charts for each profile with the mean values of each market present in that profile. Figure 5.7 shows such radar charts, where each cluster is represented by a polygon that exhibits its identity. To compare the metric values' magnitude of each cluster, we adopt the following scale: *low* is the bottom 30^{th} percentile; *medium* is between 30^{th} and 80^{th} percentile; and *high* is the top 20^{th} percentile. Such scale is based on the annual general values, i.e., considering the grouped normalized features of all markets by year.

The differences among the three plots represent minimal changes over the years. However, the distinct shapes show each cluster is *high* or *low* in certain features. Particularly, Cluster 0 presents collaborations with *high* values for *Attractiveness* and *Affinity* factors, but *medium* values for *Influence*. With a similar shape, Cluster 1 presents *medium* values for all four factors. On the other hand, Cluster 2 presents *high* values only for *Influence*, with *low Attractiveness* and *Affinity*. Finally, Cluster 3 is the group with major differences over the years: in general, its collaborations have *medium Attractiveness* and *Affinity*, and *low Influence*. Overall, each curve depicts a distinct collaboration profile, acting as a class descriptor of a cluster.

With the collaboration profiles settled, we can now answer *RQ3*. First, we analyze the distributions of success rate, and then the number of *intra*- and *inter*-genre collaborations for each profile. Here, we define success rate as the average of total streams of songs belonging to the music genres that compose the collaboration (edge) in that year.

Figure 5.8 shows the success density ridgeline plots for each profile, indicating that Profiles 0 and 1 are composed of the most successful music genre collaborations, on average. With results from Table 5.3, in general, the most successful profiles are those composed of more inter-genre collaborations. Such a result may indicate a strong correlation between musical success and inter-genre collaborations. Indeed, by teaming up with one (or more) person of a different musical style in a song, both artists may draw from one another's fan bases; i.e., they may promote themselves to new public who could increase their fan base and audiences. To summarize the characteristics of the collaboration profiles, we name each as follows.

- Profile 0 is **Solid Collaboration (Solid)**, composed of well-established collaborations between most popular genres (super-genres), which have been going on for decades. Examples include: *rap* and *hip-hop*, whose collaborative albums are hugely popular; and *hip-hop* and *pop*, whose separating line (between both genres) has become completely blurred in the last decade, mainly in the USA;

- Profile 1 is **Regular Collaboration (Regular)**, composed of the most common collaborations in all markets, which are very similar to solid collaborations but not as engaged. For instance, collaborations between *hip-hop/rap/pop* and *jazz/blues/soul*, which can be typical in many markets, but not as consolidated when compared to *Solid* ones;

- Profile 2 is **Bridge Collaboration (Bridge)**, composed of collaborations with high influence, representing bridge-like connectors between two areas of a network (mostly between divergent music styles). Such collaborations may be possible sources of investment to increase connectivity and strengthen ties among different audiences. One example is collaborations between *gospel* and other genres, such as *rap* and *MPB* (*Brazilian Popular Music*); and

- Profile 3 is **Emerging Collaboration (Emerging)**, formed mainly of collaborations between regional genres. Such partnerships generally occur within the same genre; possibly between one (or more) unknown artist and one (or more) established artist; or maybe in order to easily reach that genre audience. We propose the term *emerging* because such a profile can be seen as a transition phase for beginners, until they establish their fan bases. Examples of regional genres here include *k-pop* (popular music from South Korea), *moombahton* (fusion genre of *house* music and *reggaeton* (from Washington, D.C.), and *forró* (a popular musical genre from Brazilian Northeastern Region).

5.3 Overall Considerations

In this chapter, we analyzed and identified collaboration profiles in success-based music genre networks. We built such networks based on chart data obtained from Spotify regarding global and eight regional markets from 2017 to 2019. The temporal aspect is also considered, since there is an individual network for each year in order to analyze the temporal evolution in musical markets. We then assessed our goal by answering three research questions (RQs).

RQ1. Does the regional aspect impact on popular genres and their hit songs? Yes. Our results suggested that analyzing regional markets individually is fundamental to properly understand such scenarios, as local genres play a key role in determining hit songs and popular artists (Section 5.2.1).

RQ2. How has genre collaboration evolved over the past few years? Besides the differences in the evolution of regional markets, genre collaborations are also increasing, with emerging local genres achieving global success. Through network characterization, we found a similarity between three groups of countries (USA & Canada, UK & Australia, and non-English speaking countries), which present specific patterns in their evolution (Section 5.2.2).

RQ3. Which are the potentially intrinsic factors and indicators that influence the collaboration success? The networks' structures reveal three main factors that describe a genre collaboration: *Attractiveness*, *Affinity* and *Influence* (Section 5.2.3). Such factors uncover four different collaboration profiles: *Solid*, *Regular*, *Bridge* and *Emerging*, which act as class descriptors of successful partnerships (Section 5.2.4).

Overall, detecting genre collaboration profiles is a powerful way to assess musical success by describing similar behaviors within collaborative songs from multiple angles. Our findings may act as base material for further research tasks, e.g., prediction and recommendation. The former enables predicting the success of a given song/artist/album, while the latter can be used to point out potentially successful genre/artist collaborations. This not only benefits the MIR community, but also the music industry as a whole. In fact, music industry CEOs may maximize expected success by properly investing in potential collaborations. Finally, artists may also profit by identifying the most suitable partnerships to lead the album to early stardom. In conclusion, this work sheds light on the science behind the collaboration phenomenon, providing potential impact to the music industry.

Limitations. One limitation of this work is that we consider a short period in our analyses (2017 to 2019) since Spotify started to provide chart data at the end of 2016. Thus, we cannot make long-term conclusions on the market dynamics. Besides, we do

not analyze data from South Korea and China (sixth and seventh biggest music markets, respectively) due to the unavailability of data from such countries, which are extremely important to the global music scenario. Regarding the profiling approach, EFA and DBSCAN results may be affected by their parameters and the data dimensionality.

Chapter 6

Mining Exceptional Collaboration Patterns on Hit Songs

The genre perspective is fundamental when analyzing the impact of collaborations on musical success since each genre has a distinct audience that behaves in its own way. In the previous chapter, we built success-based genre networks and uncovered collaboration profiles that bring an additional dimension to hit songs. For example, Lady Gaga's Grammy-nominated album *Chromatica*[1] has collaboration as one of its biggest strengths. The collaborations with Ariana Grande (*pop*), Elton John (*rock*), and Blackpink (*k-pop*) contributed to maintain her among the most prominent pop names nowadays, as well as introducing her to new audiences. In fact, the partnership with Grande in the song *Rain on Me* won the Grammy for Best Pop Duo/Group Performance in 2021. Such an intra-genre collaboration (they both are *pop* artists) has a *Solid* profile, which comprises well-established connections between popular genres.

Indeed, the rise of collaborations in the music market highlights its dynamic and unpredictable nature. Given the diversity of collaborations between artists from several genres, it becomes challenging to conduct predictive and descriptive analyses in such a context. For example, it may be relevant to record labels to uncover frequent genre collaborations which achieve a higher level of success to plan future song releases. Thus, in this chapter, we go further in the study of genre collaborations by using the genre networks and collaboration profiles to mine exceptional patterns of musical genres in songs that have been successful in recent years, i.e., to verify if there is a relationship between the combination of different musical genres and success. Specifically, we aim to achieve this goal through the following research questions (RQs):

RQ1. Compared to the global scenario, do regional markets present distinct patterns of frequent genre combinations in hit songs?

RQ2. In collaborative hit songs (i.e., with more than one artist), are there connection patterns between genres that achieve above-normal success?

[1] *Chromatica* was nominated to Best Pop Vocal Album in the 63rd Grammy Awards (2021).

RQ3. Is it possible to identify and recommend combinations of music genres that are promising and relevant to each market?

The remainder of this chapter is organized to answer all such questions. First, we introduce our methodology based on data mining techniques in Section 6.1. Then, Section 6.2 details the experimental evaluation and presents the results. Finally, we make our overall considerations on the findings in Section 6.3.

6.1 Methodology

This section presents the methodology used to find both frequent and exceptional patterns in hit songs. We use two data mining techniques in our experiments: Frequent Itemset Mining (FIM) and Subgroup Discovery (SD), whose definition and fundamental concepts are detailed in Chapter 3. Regarding the research questions (RQs), to answer *RQ1* and *RQ3*, we model hit songs as transactions to find frequent itemsets and association rules (Section 6.1.1). On the other hand, *RQ2* requires to perform an SD algorithm on the genre collaboration network from Chapter 5 (Section 6.1.2).

6.1.1 Modeling Hit Songs as Transactions

In data mining, discovering frequent itemsets and association rules requires a transactional dataset. In such modeling, the dataset instances (i.e., transactions) are composed of a list of items.[2] Here, we use the Music Genre Dataset (MGD) presented in Chapter 5 to obtain Spotify success data of nine music markets worldwide (global and eight countries) from 2017 to 2019. MGD provides the list of songs that entered in weekly Top 200 charts for each market and year, as well as acoustic features that describe such songs. In addition, it provides relevant information on the artists who interpret the songs, such as the genre list of each one.

We model an individual transactional dataset for each market and year to find the most frequent genre combinations and association rules. We define each hit song as a single transaction in which the items are the musical genres of the artists who sing it. If a song has more than one artist and they all share a common genre, this genre appears

[2]See Chapter 3 for formal definitions.

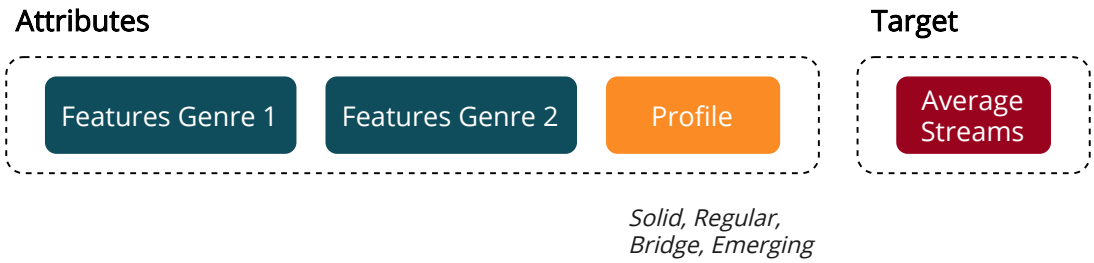


Figure 6.1: Representing the edges of Genre Collaboration Networks as instances of the Subgroup Discovery (SD) problem.

only once in the transaction. Therefore, we are not interested in the number of repeated genres in a song, but in the diversity of different genres that compose it.

For example, consider the remix version of *Despacito* by Luis Fonsi, Daddy Yankee, and Justin Bieber, which spent 16 consecutive weeks in the #1 position on Billboard Hot 100 in 2017. The transaction correspondent to this song would then comprise the genres of all three artists: *latin*, *tropical*, *pop*, *reggaeton*, and *hip hop*. Besides the fact that *latin* and *tropical* are shared by Fonsi and Yankee and *pop* is shared by Fonsi and Bieber, each genre is counted once. Thus, the final transaction for *Despacito* is represented by the tuple `(hip hop, latin, tropical, pop, reggaeton)`.[3]

6.1.2 Network Subgroup Discovery

In order to properly answer *RQ2*, i.e., to find genre connections with above-normal success, we use the genre collaboration networks from the previous chapter as the input of a Subgroup Discovery (SD) task. We consider each market and year separately to preserve the temporal and regional aspects in our analyses. Recalling the definition of the SD problem from Chapter 3, its input is a dataset composed of a set of attributes describing the instances and a target variable, used to distinguish the behavior of the subgroups from the whole dataset. Thus, a subgroup is said to be exceptional if the distribution of the target in its instances deviates from the dataset.

In our approach (summarized in Figure 6.1), for each market and year, we consider the network edges as the instances of the SD model, representing the collaboration between musical genres. Therefore, it is necessary to select attributes that describe the nodes individually and also characteristics already known from the collaboration. Hence, the attribute set for each instance is composed of features from the two genres of the respective edge, as well as the collaboration profile (*Solid*, *Regular*, *Bridge* and *Emerging*) for such

[3]Sorted by alphabetical order.

an edge. To describe each genre, we select the following acoustic features[4] from Spotify: *acousticness*, *danceability*, *duration_ms*, *energy*, *liveness*, *loudness*, *speechiness*, *tempo*, and *valence*.

As such features are provided for individual songs, we assign each genre the median values of all songs from artists belonging to such a genre. These values are then discretized based on the quartiles for each variable. That is, values in the first quartile (below the 25th percentile) are classified as *low*, whereas values in the second and third quartiles (between the 25th and 75th percentile) are *medium*. Values above the 75th percentile are then classified as *high*. Finally, we set the average number of streams of each edge as our target variable, as it is a success metric provided by Spotify.

6.2 Results and Applications

This section presents the results and discussions for each analysis carried out to answer our the research questions: frequent pattern mining (*RQ1*, Section 6.2.1), network subgroup discovery (*RQ2*, Section 6.2.2), and association rules (*RQ3*, Section 6.2.3).

6.2.1 Genre Frequent Patterns

Discovering changes in genre preferences shows the dynamic nature of the music market. As an important cultural artifact, the way music is consumed in different countries may be influenced by language, demographics, and other social aspects. In addition, musicians are naturally venturing into new domains and working outside of the style they had initially emerged from, resulting in a massive variety of new songs and musical tastes. Therefore, in this section, we answer *RQ1* by investigating whether genre combination varies at a country level. That is, we analyze if the association of distinct musical genres in hit songs has specific patterns in global and regional markets.

We focus on finding the most frequent genre associations by applying a Frequent Itemset Mining (FIM) method from the set of hit songs in each musical market from the Music Genre Dataset (MGD). We do so by running the Python implementation of Apriori,[5] one of the most used state-of-the-art FIM algorithms [3]. Here, we perform a

[4]The definitions of the acoustic features are given in Chapter 2.
[5]PyFIM: https://borgelt.net/pyfim.html

Table 6.1: Top 5 most frequent patterns in global and English-speaking markets (2019).

Market	Pattern	Support	Market	Pattern	Support
Global	('dance pop', 'pop')	0.271	*Australia*	('dance pop', 'pop')	0.294
	('latin', 'reggaeton')	0.173		('rap', 'hip hop')	0.162
	('hip hop', 'trap')	0.172		('electropop', 'pop')	0.145
	('rap', 'hip hop')	0.168		('rap', 'pop rap')	0.145
	('rap', 'trap')	0.151		('pop rap', 'hip hop')	0.131
UK	('dance pop', 'pop')	0.285	*USA*	('hip hop', 'rap')	0.305
	('rap', 'hip hop')	0.159		('trap', 'rap')	0.289
	('tropical house', 'pop')	0.133		('pop rap', 'rap')	0.261
	('tropical house', 'dance pop')	0.127		('trap', 'hip hop')	0.246
	('tropical house', 'dance pop', 'pop')	0.125		('pop rap', 'hip hop')	0.230

temporal and regional analysis, running the algorithm separately for each market and year (2017 to 2019). Following the methodology of Section 6.1.1, we define the transactions of our FIM task as hit songs, whose items are the musical genres of each artist who sing them. As our goal is not to evaluate the period in which each song remained on the charts, they are included only once in the dataset. The frequent genre patterns are evaluated using relative support, which is the proportion of transactions that contain such a pattern.

Overall, as language is crucial for listening to music, we divide our eight regional markets into two distinct groups: English and non-English speaking countries. The former includes Australia, Canada, the United Kingdom, and the United States, while the latter comprises Brazil, France, Germany, and Japan. We then perform our analyses comparing the countries with each other and the patterns found in the global charts, which is an aggregation of all territories in which Spotify is available. Here, we present the results for selected markets and years for readability purposes. The complete data with frequent patterns for all markets over time are shown in Appendix C.

Table 6.1 presents the five most frequent genre patterns in hit songs for global and English-speaking markets in 2019. Itemsets are sorted by their relative support value, i.e., their frequency. Regarding the global scenario, there is a strong presence of mainstream genres such as *pop*, *hip hop*, and *rap*. These genres include regional versions of themselves (e.g., *Chicago rap* is included in *rap*), which may contribute for their high support values. Besides, the combination of the regional genres *latin* and *reggaeton* appear in 17.3% of all global hit songs, showing their popularization across the world. The Latin music expansion has its roots in the early 2000s with names such as Shakira and Ricky Martin. In the late 2010s, it has achieved a higher level of popularity led by artists such as Bad Bunny, J Balvin, and Karol G.

Analyzing the English-speaking countries individually, we note a high similarity in the popular genre combinations. For instance, the union of *hip hop* and *rap*, which is present in 30.5% of hit songs in the United States, is also relevant in Australia (16.2%),

Table 6.2: Top 5 frequent patterns in global and non-English speaking markets (2019).

Market	Pattern	Support	Market	Pattern	Support
Global	('dance pop', 'pop')	0.271	*Brazil*	('brazilian funk', 'pop')	0.177
	('latin', 'reggaeton')	0.173		('electro', 'brazilian funk')	0.102
	('hip hop', 'trap')	0.172		('sertanejo', 'brazilian funk')	0.097
	('rap', 'hip hop')	0.168		('electro', 'pop')	0.080
	('rap', 'trap')	0.151		('trap', 'hip hop')	0.064
France	('hip hop', 'pop')	0.584	*Japan*	('j-rock', 'j-pop')	0.283
	('rap', 'hip hop')	0.449		('other', 'j-pop')	0.140
	('rap', 'pop')	0.423		('anime', 'j-pop')	0.138
	('rap', 'hip hop', 'pop')	0.393		('dance pop', 'pop')	0.133
	('francoton', 'pop')	0.174		('r&b', 'j-pop')	0.108

Canada[6] (27.2%) and the United Kingdom (15.9%). All such countries present a strong similarity to the global market, as they share several cultural aspects, including language. In fact, English is the most widely spoken language worldwide in terms of countries where it is official, accounting for 59 countries in all continents.[7]

On the other hand, the analysis of frequent genre patterns for non-English speaking countries reveals a strong regional component in most countries. Table 6.2 presents the five most frequent genre associations in 2019 for three countries:[8] Brazil, France, and Japan. All such countries have patterns with regional rhythms, such as *francoton* in France, and *brazilian funk* and *sertanejo* in Brazil. However, Japan stands out in this regard, as all five patterns have regional styles. The main genres in such a market market include *j-pop*, *j-rock* and *anime*. Besides, our results reveal the absence of genres such as *hip hop* and *rap* in Japan, which are present in all other markets. In all four countries, the presence of local genres increased over time, revealing a tendency of the population to value their own culture and consequently promote it globally.

t

6.2.2 Exceptional Genre Collaborations

In this section, we answer *RQ2* by finding exceptional genre collaboration patterns from hit songs, i.e., collaborations in which the success is above the average in the whole dataset. Therefore, we perform a subgroup discovery (SD) analysis in our genre collaboration networks. We maintain the notion of temporality in our analyses as we build,

[6]We do not show data for Canada in Table 6.1 since its ranking is very similar to the USA.
[7]World Atlas (Access on May 3, 2021): https://bit.ly/3ePt7CK
[8]Germany is not shown in Table 6.2 as its ranking is similar to Global.

for each market, a collaboration network from hit songs of each year (2017, 2018, and 2019). Hence, we analyze 27 distinct collaboration networks (three annual networks for nine music markets). Following the methodology presented in Section 6.1.2, we consider the network edges as the instances of our SD problem, described by acoustic features of the genres and the collaboration profile between them. We also set the average number of streams as the target value.

Here, we use the *Beam Search* algorithm from the *pysubgroup* Python library[9] [62]. This algorithm finds the relevant subgroups according to a predefined target variable (here, the average number of *streams*), evaluated by a quality metric. In this work, we use the function *StandardQFNumeric* from the same library, which handles numeric target variables. For a given subgroup SG and a parameter α, this function is defined by Equation 6.1. N_{SG} is the number of instances in the subgroup, N is the total number of instances in the dataset, μ_{SG} is the average of the target variable within the subgroup, and μ is the average of the target variable in the dataset. In our experiments, we empirically choose $\alpha = 0.5$, as we want to emphasize the difference in the target variable rather than the subgroup size.

$$q(SG, \alpha) = \left(\frac{N_{SG}}{N}\right)^{\alpha} (\mu_{SG} - \mu) \tag{6.1}$$

Table 6.3 shows the exceptional subgroups found by BeamSearch in the networks for nine global and regional markets from 2017 to 2019. Regarding the global network, there are no exceptional subgroups in 2017 and 2018, which may indicate a homogeneous behavior in the collaborations during this period. Following the popularization of streaming services, there is a change in 2019, when a specific subgroup with an average stream count four times higher than expected emerges in the network. Such a subgroup includes edges in which the first genre has low acousticness, medium danceability, high degree (i.e., connectivity), medium duration, and high energy. In contrast, the second one has medium values for energy, loudness, and tempo. In addition, the edges composing the subgroup share the *Regular* collaboration profile. An example is the collaboration between *electro house* and *pop*, which happens in the song *Happier* by Marshmello and Bastille. The song was released in August 2018, but its popularity grew until 2019, as it spent 27 weeks in the top 10 of the Billboard Hot 100.

In a lower-level analysis, all regional markets have subgroups with attributes and target distributions different from each other. However, when analyzing each subgroup individually, we note some connections that repeat in some countries. For instance, in 2018, the edge between *dubstep* and *pop* belongs to exceptional subgroups in five markets: Australia, Canada, Germany, United Kingdom, and the United States. Except for Germany, all countries are English-speaking and share several cultural aspects. Besides, the Internet and social media advance provides a global platform where users can share and

[9]pysubgroup: https://github.com/flemmerich/pysubgroup

Table 6.3: Exceptional subgroups in networks from selected markets (2017-19).

Market	Year	Subgroup	N	E	S_{SG}	S_D	Q
Global	2017	no exceptional subgroups found					
	2018	no exceptional subgroups found					
	2019	acousticness='low' ∧ danceability='medium' ∧ degree='high' ∧ duration_ms='medium' ∧ energy='high' ⇒ energy='medium' ∧ loudness='medium' ∧ tempo='medium' [profile='regular']	15	20	185.95	45.14	70.79
Australia	2017	danceability='high' ∧ liveness='low' ⇒ liveness='low' ∧ valence='high'	5	9	19.55	3.11	3.59
	2018	acousticness='low' ∧ danceability='medium' ∧ valence='low' ⇒ speechiness='medium'	12	11	20.39	3.87	5.37
	2019	acousticness='low' ∧ danceability='medium' ∧ degree='high' ⇒ energy='medium' ∧ loudness='medium' ∧ speechiness='medium' [profile='regular']	14	19	11.05	2.41	3.97
Brazil	2017	danceability='medium' ⇒ acousticness='high' ∧ degree='medium' ∧ duration_ms='low' ∧ energy='high'	14	19	18.69	6.13	5.14
	2018	speechiness='high' ⇒ acousticness='high' ∧ danceability='high' ∧ duration_ms='low'	3	2	96.95	4.82	12.79
	2019	acousticness='medium' ∧ liveness='high' ∧ speechiness='high' ⇒ danceability='high' ∧ valence='high'	4	3	67.08	4.41	8.33
Canada	2017	liveness='low' ⇒ acousticness='medium' ∧ danceability='high' ∧ degree='medium' ∧ speechiness='medium'	6	8	14.26	2.17	3.18
	2018	danceability='medium' ∧ energy='high' ∧ valence='low' ⇒ speechiness='medium'	12	11	20.36	2.36	4.78
	2019	acousticness='low' ∧ danceability='medium' ∧ degree='high' ∧ speechiness='medium' ⇒ danceability='medium' ∧ loudness='medium' ∧ tempo='medium' [profile='regular']	12	16	10.95	1.79	3.33
France	2017	acousticness='medium' ∧ degree='medium' ∧ tempo='low' ⇒ duration_ms='medium'	12	25	8.28	2.74	2.45
	2018	danceability='medium' ∧ duration_ms='medium' ∧ liveness='high' ∧ loudness='high' ⇒ tempo='medium'	14	19	8.50	2.91	2.54
	2019	danceability='high' ∧ liveness='medium' ∧ valence='high' ⇒ acousticness='medium'	3	2	27.93	4.16	3.24
Germany	2017	liveness='low' ⇒ acousticness='medium' ∧ degree='medium' ∧ energy='high' ∧ speechiness='medium'	6	8	35.92	3.65	7.79
	2018	acousticness='low' ∧ loudness='high' ∧ tempo='high' ∧ valence='low' ⇒ tempo='medium'	12	11	23.84	4.45	6.87
	2019	energy='high' ⇒ liveness='low' ∧ tempo='low'	9	22	20.03	3.87	5.36
Japan	2017	danceability='medium' ∧ liveness='medium' ⇒ danceability='medium' ∧ duration_ms='medium' ∧ loudness='medium' ∧ tempo='low' ∧ [profile='regular']	11	16	0.76	0.27	0.19
	2018	loudness='high' ⇒ danceability='low' [profile='solid']	2	2	8.36	0.35	1.11
	2019	acousticness='low' ∧ danceability='low' ∧ speechiness='low' ⇒ valence='low'	3	2	12.30	0.42	1.62
UK	2017	liveness='low' ⇒ acousticness='medium' ∧ degree='medium' ∧ duration_ms='medium' ∧ tempo='low'	11	17	27.23	4.19	7.73
	2018	acousticness='low' ∧ energy='high' ∧ valence='low' ⇒ speechiness='medium' ∧ tempo='medium'	11	10	37.80	6.61	8.72
	2019	acousticness='low' ∧ danceability='medium' ∧ degree='high' ∧ liveness='medium' ⇒ danceability='medium' ∧ loudness='medium' ∧ tempo='medium' [profile='regular']	13	19	15.78	3.44	5.46
USA	2017	loudness='medium' ∧ tempo='low' ⇒ duration_ms='medium' ∧ speechiness='medium' ∧ tempo='low'	6	4	94.58	13.38	17.58
	2018	acousticness='low' ∧ degree='medium' ∧ energy='high' ∧ valence='low' ⇒ speechiness='medium'	13	12	85.66	14.20	22.03
	2019	acousticness='low' ∧ danceability='medium' ∧ degree='high' ⇒ energy='medium' ∧ loudness='medium' ∧ speechiness='medium' [profile='regular']	13	18	58.60	6.65	14.67

N: number of nodes **E:** number of edges S_{SG}: average *streams* in subgroup (10^6). S_D: average *streams* in network (10^6). **Q**: quality metric value (10^7).

promote their musical tastes. The American DJ Marshmello appears again in two of the most popular songs involving *dubstep* (one of his musical genres) and *pop*: *Wolves* with Selena Gomez and *Friends*, in partnership with Anne-Marie.

We also highlight two regional markets with a strong presence of local genres in exceptional subgroups, revealing the popularity of such genres in their countries. Popular subgroups in Japan involves connections of the *anime* genre with itself (i.e., intra-genre collaboration) and *j-pixie* in 2018 and 2019, respectively. Besides, Brazilian regional genres appear in subgroups from all considered years. Specifically, in 2019, the subgroup that comprises the edge between *afrofuturism* and *pagode baiano* has an average number of streams more than ten times bigger than the whole network. Such a huge success is boosted by songs such as *Bola Rebola* by Anitta, Tropkillaz, J Balvin, and MC Zaac, which debuted in #1 in Brazil's daily chart in Spotify[10] with more than 1.2 million streams. Therefore, as the second and tenth biggest music markets in the world,[11] such countries reinforce the importance of considering regional markets individually, as their engagement shapes the global environment.

6.2.3 Recommending Promising Genre Associations

Using the data mining framework offers a wide range of possibilities to perform descriptive analyses in datasets. For instance, the frequent genre patterns mined in Section 6.2.1 can be used to uncover association rules, which inform how items (i.e., music genres) are associated with each other. Recalling the definition from Chapter 3, association rules are represented with expressions from the type $A \rightarrow B$, representing the occurrence of an itemset B (i.e., consequent) given that A (i.e., antecedent) also happens. There are several metrics to evaluate such rules on different perspectives, including likelihood (*confidence*) and surprise (*lift value*). In this section, we answer *RQ3* by using such rules to detect and recommend outstanding genre associations.

We define promising rules according to their lift value, which measures their level of surprise. The intuition behind lift is that it is the ratio between the joint probability of the antecedent and the consequent, and their probability of being statistically independent. Therefore, values above 1 mean that the consequent is much more likely to happen than expected, given the occurrence of the antecedent. In contrast, values below 1 represent the opposite. As we aim to find the most promising genre associations, we look for rules with high lift values. Here, we still perform an individual analysis for each music market

[10]Spotify Chart, 2019/02/22: https://spotifycharts.com/regional/br/daily/2019-02-22
[11]IFPI Global Music Report 2019: https://gmr.ifpi.org

Table 6.4: Association rules in global and regional markets sorted by lift value (2019).

Market	Rule	Lift	Confidence
Global	('latin', 'reggaeton') → tropical	7.922	0.468
	('latin') → tropical	7.821	0.462
	('reggaeton') → tropical	7.722	0.456
Australia	('tropical house') → house	7.655	0.342
	('tropical house', 'pop') → house	7.173	0.321
	('tropical house', 'pop') → electro	7.111	0.670
Brazil	('hip hop') → trap	6.187	0.434
	('brazilian funk', 'pop') → pagode baiano	5.473	0.425
	('hip hop') → pop rap	5.235	0.303
Canada	('r&b') → soul	7.485	0.226
	('dance pop') → tropical house	3.214	0.243
	('dance pop', 'pop') → tropical house	3.160	0.239
France	('rap', 'pop') → hip hop	1.301	0.900
	('rap', 'pop') → francoton	1.263	0.325
	('hip hop', 'pop') → rap	1.259	0.796
Germany	('dance pop') → tropical house	5.909	0.400
	('dance pop') → electro	5.908	0.338
	('dance pop', 'pop') → tropical house	5.824	0.394
Japan	('r&b') → j-rap	8.067	0.228
	('dance pop') → electro	4.348	0.283
	('dance pop', 'pop') → electro	4.284	0.279
UK	('rock') → indie rock	8.370	0.364
	('rock') → indie	6.216	0.231
	('pop rap', 'hip hop') → trap	5.682	0.660
USA	('pop rap', 'pop', 'rap') → r&b	2.990	0.291
	('pop', 'rap') → r&b	2.888	0.281
	('hip hop', 'pop') → r&b	2.878	0.280

and year, using the Apriori algorithm to find the relevant rules.

Table 6.4 present the three most promising rules for each market in 2019[12]. We sort the rules by their lift values, but we also present the confidence value to enrich our analyses. The results for the global market reveal the strong association of regional genres such as *latin*, *reggaeton*, and *tropical*. Analyzing the lift value for the first rule, we can affirm that the occurrence of the genre *tropical* when *latin* and *reggaeton* co-occur in a hit song is almost eight times than the expected. Such a result indicates that adding *tropical* in songs containing *latin* and *reggaeton* may increase in up to 7.922 times the chances of reaching the Top 200 chart. Besides, the rule confidence informs that *tropical* is present in 46.8% of the transactions (i.e., hit songs) that contain *latin* and *reggaeton*. Indeed, Latin musical genres had a boom in 2019, following the continuous growing trend observed in the late 2010s. Despite not achieving the #1 position as Luis Fonsi and Daddy Yankee

[12]For readability purposes, the complete results for 2017 and 2018 are in Appendix C.

did in 2017 with *Despacito*, artists such as Karol G, Bad Bunny, Ozuna, and J Balvin demonstrated the power of such genres, as they managed to put two or more songs in the charts.

We also note the presence of local genres in outstanding association rules, mainly in non-English speaking countries. For instance, *francoton* is associated with *rap* and *pop* in France, whereas the probability of *j-rap* occurring is eight times higher given the presence of *r&b* in Japanese hit songs. In Brazil, the genre *pagode baiano* (consequent) appears on 42.5% of the songs containing *brazilian funk* and *pop* (antecedent). In addition, the occurrence of the antecedent of such a rule increases more than five times the chances of the consequent in Brazilian hit songs. Thus, combining such genres increases considerably the chances of a song to reach Brazilian charts. The singer-songwriter Anitta is an example of such an effect, as her music style list includes all three genres aforementioned. She is one of the most popular artists in the country, and her singles *Onda Diferente* (with Ludmilla) and *Combatchy* (with Lexa, Luísa Sonza, and MC Rebecca) contributed to the high relevance of associating such music genres.

Overall, association rules are a powerful tool to understand musical success, as they reveal the level of combination between musical genres in global and regional markets. Similar to the previous sections, local genres play a fundamental role in regional markets, reinforcing their distinct cultural identities. Besides, using lift values to evaluate rules allows recommending promising genre combinations based on their high association level. Such an approach provides considerable benefits to artists, as they can plan their subsequent releases by choosing artists from genres with a high level of association with their own to collaborate. In addition, record labels may use our findings to diversify their set of artists and promote collaborations with high potential of success between them. Indeed, music is a dynamic and unpredictable industry, but this strategy may help guide artists and record labels to develop approaches to achieve success and increase their numbers.

6.3 Overall Considerations

In this chapter, we investigate the relation between the combination of different musical genres and success under a data mining perspective. Using data from the Music Genre Dataset, which contains Spotify chart information from several markets, we perform descriptive analyses to identify frequent genre combinations and exceptional subgroups in genre collaboration networks. We conducted temporal analyses for both global and regional markets, i.e., we run the data mining algorithms individually for each market and year (2017 to 2019). Such an approach is helpful to reveal the evolution of musical

tastes over time and show how cultural aspects shape local music markets. We address such an objective by answering three research questions (RQs).

RQ1. Compared to the global scenario, do regional markets present distinct patterns of frequent genre combinations in hit songs? Yes. We modeled hit songs as transactions in which the items are their musical genres to run a Frequent Itemset Mining algorithm. We found that there is indeed a difference in popular genre patterns in regional markets, mainly in non-English speaking countries (Section 6.2.1).

RQ2. In collaborative hit songs (i.e., with more than one artist), are there connection patterns between genres that achieve above-normal success? Yes. We use a Subgroup Discovery technique in genre collaboration networks to detect genre connections in which the success metric (i.e., the average number of streams) deviates from the whole network. We identified exceptional subgroups in all markets, and each one presented distinct results that show the importance of considering the local component in success analyses (Section 6.2.2).

RQ3. Is it possible to identify and recommend combinations of music genres that are promising and relevant to each market? Yes. We mined association rules to recommend promising genre combinations based on their level of surprise. Again, we found that local genres play a fundamental role in regional markets as they are included in most of the relevant associations (Section 6.2.3).

In conclusion, performing diagnostic analyses is crucial in music, as it allows the understanding of some relevant aspects behind success. Following the findings from previous chapters, our results reinforce the importance of analyzing regional markets, as they behave differently compared to the global scenario or even to the United States (i.e., the biggest music market in the world). For example, in the past few years, the world has seen local genres such as *reggaeton* and *k-pop* becoming extremely popular worldwide. Therefore, our findings provide benefits to artists and record labels, as they serve as a first step in developing strategies to promote their work across the world.

Limitations. One limitation of this work is that we do not consider how many times a hit song appears in Spotify charts each year when building our transactional database. Such a strategy would emphasize the genre frequency, by counting the song several times; but, on the other hand, the distinctness of hit songs would be lost. In addition, we do not use the position in the charts in our analyses. Thus, songs that reached the Top 10 are treated equally to songs from the bottom of the charts. Further experiments must evaluate the impact of such information in genre popularity compared to the analyses already performed here.

Chapter 7

Concluding Remarks

In the past few years, music has been transformed by digital technologies and analytics. The popularization of streaming services facilitates the analyses on this subject, as they provide a wealth of data about how music is discovered and consumed. One example is the study of music genres, which are fundamental in the musical scenario by aggregating songs with common characteristics. Popular genres are constantly changing, and the notion of genres themselves is blurred as never before. Acting as one of the most prominent high-level music descriptors, analyzing the music ecosystem from a genre perspective is highly relevant to understanding its dynamics. Therefore, in this work, we analyzed artist collaboration from a genre perspective to better understand how genre connections impact musical success. Here, we presented our contributions according to the three specific research goals (RGs) that guide this work.

RG1. Understand the temporal evolution of both artist and genre careers, by identifying and predicting periods of high impact in such careers (i.e., hot streaks). We investigated the temporal evolution of musical success to identify and predict periods of high impact in such careers. We modeled success in the music industry by building time series for artists and genres based on weekly chart positions. Using such series, we proposed a method to detect the continuous periods in which success is above average, i.e., *hot streaks*. The characterization of such hot streaks demonstrated that music genres have specific behavior patterns that must be observed. As genre is a crucial factor in musical success, we presented a genre-based model to predict hot streak periods. Our results showed that our model can successfully predict whether a week belongs to a hot streak period. Moreover, we found that the occurrence of such periods depends not only on the songs present in the charts but also on information obtained from the genres and artists, such as collaboration.

RG2. Analyze the dynamics of cross-genre connections by detecting collaboration profiles in success-based networks (i.e., connections formed by genres of artists who cooperate and create hit songs) Indeed, artist collaboration is a strong force driving music nowadays. It is deeply related to music genres, as such connections usually help artists bridge the gap between styles and genres and approach new audiences. Therefore, our next research goal (*RG2*) is to analyze the dynamics of cross-genre connections based on artists who cooperate and create hit songs. We built genre collaboration networks using data from global and regional markets. Our results showed that analyzing distinct markets worldwide is fundamental, as local genres play a key role in determining popular songs and artists. Furthermore, the networks' structure (i.e., connectivity metrics) revealed three distinct factors to describe genre connections: *Attractiveness*, *Affinity*, and *Influence*. From such factors, we performed a cluster analysis to uncover four different collaboration profiles: *Solid*, *Regular*, *Bridge*, and *Emerging*. Such profiles are an important tool to assess musical success, as they act as class descriptors of successful partnerships.

RG3. Mine frequent genre patterns within hit songs in recent years, i.e., investigating the relationship between combining different music genres and musical success In order to deepen the knowledge regarding the impact of genre collaboration in musical success, we investigate the combination of different music genres in hit songs by mining both frequent and exceptional genre patterns (*RG3*). We modeled hit songs as transactions for each regional market and found frequent genre combinations by running a Frequent Itemset Mining algorithm. Next, we applied a Subgroup Discovery technique in the collaboration networks to reveal unusual genre connections that achieve a high level of success. All such analyses revealed a difference in popular genre patterns in regional markets, mainly in non-English speaking countries. In such markets, local genres play a crucial role in musical success, producing a high impact in the global scenario. Finally, we used association rules to recommend promising genre combinations. Again, regional genres are included in the most relevant associations, reinforcing the fact that local engagement shapes the global environment.

Overall, this work provides an extensive diagnosis and relevant insights on how genres relate to musical success. We contribute to the Hit Song Science field by advancing the knowledge on musical genres and collaborations as features in success-based models and making available a novel dataset with regional and temporal information. Such findings benefit not only scientists but also the music industry. In fact, in the past few years, several music actors have been using data insights to perform diagnostic analyses on the market and support business decisions. For instance, Instrumental is a data-driven British service that aims to use artificial intelligence and machine learning techniques to discover high potential talents and offer the most promising partnerships for independent artists.[1]

[1]About Instrumental: https://www.weareinstrumental.com/about

From such a perspective, our work may contribute to music companies by enhancing such predictive and recommendation models with genre information. Therefore, both artists and record labels benefit from our findings, as they shed light on the science behind musical success and contribute to developing strategies to promote musical content across the world.

7.1 Research Outcomes

In addition to this master thesis, our research generated other products, such as four scientific publications and two publicly available datasets with enhanced musical genre information, as follows.

- **Oliveira et al.** [77]: *Musical Genre Analysis Over Dynamic Success-based Networks*. Publication at the Workshop of Theses and Dissertations in Databases (WTDBD) of the 35th Brazilian Symposium on Databases (SBBD 2020);
- **Oliveira et al.** [78]**:** *Music Genre Dataset (MGD)*, a dataset on musical success over time in global and regional markets with enhanced artist and genre collaboration information;
- **Oliveira et al.** [79]**:** *Detecting Collaboration Profiles in Success-based Music Genre Networks*. Research paper in the Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR 2020). This paper received the Best Poster Presentation Award;
- **Oliveira et al.** [80]**:** *Music-oriented Hot Streak Information Collection (MUHSIC)*, a dataset on musical careers with artist and genre hot streak data;
- **Oliveira et al.** [81]**:** *MUHSIC: An Open Dataset with Temporal Musical Success Information*. Publication at the Dataset Showcase Workshop (DSW) of the 36th Brazilian Symposium on Databases (SBBD 2021);
- **Oliveira et al.** [82]**:** *Musical Success in the United States and Brazil: Novel Datasets and Temporal Analyses*. Research article in the Journal of Information and Data Management (JIDM);

7.2 Future Work

Based on our work on genre collaboration and music success, we identify and discuss new research directions that require further investigation. These topics are not the only open research problems within Hit Song Science (HSS) and Music Information Retrieval (MIR), but they are key factors that may contribute to advancing such fields.

Dealing with multiple sources. Data integration is one of the main issues in many Computer Science research fields. In Hit Song Science, this topic is becoming more relevant and necessary, as there is no unique data source for all necessary features and data. For instance, to the best of our knowledge, there is no data source that provides both acoustic and lyrics-based features. Furthermore, the lack of a unique and universal identifier for each music makes an integration involving several data sources very challenging. Besides, information such as the musical genre(s) of a given song is not standardized in all data sources, mainly due to the blurred line existent between music styles that are close to each other.

Regional markets' diversity. Most studies on Hit Song Science use data from the American market (e.g., Billboard Hot 100 Chart). This may be because the United States is the biggest music market in the world, which may facilitate the acquisition and use of such data. Research studies that consider music markets other than the USA focus mainly on European countries, such as the United Kingdom. However, there are many other relevant markets with distinct characteristics and behavior, which require an individual analysis of success. For example, South Korea, China and Brazil are among the top 10 music markets in the world,[2] with a vibrant music scene and popular regional genres. Such genres have become popular in the global scenario as they connect with other well-established music genres (e.g., collaborations involving *pop*, *k-pop*, and Latin genres such as *reggaeton*). Therefore, as local engagement shapes the global environment, future work must consider the regional aspect, thus ensuring that music culture within such countries are accounted for.

Lack of standardized success metrics. Defining the popularity of a song is still a challenge, and each study in HSS uses specific success metrics. For instance, in this work we used chart-based metrics (Chapter 4) and the number of streams (Chapters 5 and 6). Therefore, as there is no standard, researchers are unable to perform a fair comparison between their work and the existent literature on the subject. Hence, finding a way to properly generalize success would support future work on HSS to more accurately capture popularity definitions. Moreover, it would enable transposing their findings to a commonly

[2]IFPI Global Music Report: http://gmr.ifpi.org/

understood metric, which then allows a complete evaluation by comparing performance with current work (as the one presented here).

Importance of social aspects. The ever-growing popularization of social networks in the last two decades has deeply changed the music industry. The propagation of songs in such platforms is fundamental in their success, as the viral phenomenon of songs in social media may lead a newly released one to stardom or even lead back a great hit from the past to the top of the charts. Since marketing has great impact on the future success of songs, it is increasingly important to consider the latest social platforms and features, which could as well give strong indications of a song's hit potential. Although such features have been used in previous research, novel approaches on HSS need to combine both audio and social data to enhance model efficiency.

Bibliography

[1] Fabian Abel et al. Analyzing the blogosphere for predicting the success of music and movie products. In *ASONAM*, pages 276–280, Odense, Denmark, 2010.

[2] Charu C. Aggarwal. *Recommender Systems - The Textbook*. Springer, 2016. ISBN 978-3-319-29657-9. doi: 10.1007/978-3-319-29659-3.

[3] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *VLDB*, volume 1215, pages 487–499, 1994.

[4] Zayd Al-Beitawi, Mohammad Salehan, and Sonya Zhang. Cluster analysis of musical attributes for top trending songs. In *53rd Hawaii International Conference on System Sciences*, pages 1–7, Honolulu, HI, USA, 2020. ScholarSpace. doi: 10.24251/HICSS.2020.017. URL https://doi.org/10.24251/HICSS.2020.017.

[5] 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*, pages 859–864, New York, NY, USA, 2019. IEEE. doi: 10.1109/ICMLA.2019.00149. URL https://doi.org/10.1109/ICMLA.2019.00149.

[6] Carlos Vicente Soares Araujo, Marco Antônio Pinheiro de Cristo, and Rafael Giusti. A model for predicting music popularity on streaming platforms. *RITA*, 27(4):108–117, out 2020. ISSN 2175-2745. doi: 10.22456/2175-2745.107021. URL https://doi.org/10.22456/2175-2745.107021.

[7] Carlos V.S. Araujo et al. Predicting music success based on users' comments on online social networks. In *WebMedia*, pages 149–156, Brazil, 2017. doi: 10.1145/3126858.3126885.

[8] Tom Arjannikov and John Z. Zhang. An association-based approach to genre classification in music. In *ISMIR*, pages 95–100, Taipei, Taiwan, 2014.

[9] Noah Askin and Michael Mauskapf. What makes popular culture popular? product features and optimal differentiation in music. *American Sociological Review*, 82(5):910–944, sep 2017. doi: 10.1177/0003122417728662. URL https://doi.org/10.1177/0003122417728662.

[10] Simcha Avugos, Jörn Köppen, Uwe Czienskowski, Markus Raab, and Michael Bar-Eli. The "hot hand" reconsidered: A meta-analytic approach. *Psychology of Sport and Exercise*, 14(1):21–27, 2013.

[11] Peter Ayton and Ilan Fischer. The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & cognition*, 32(8):1369–1378, 2004.

[12] Claudio Baccigalupo et al. Uncovering affinity of artists to multiple genres from social behaviour data. In *ISMIR*, pages 275–280, Philadelphia, USA, 2008.

[13] Jotthi Bansal and Matthew Woolhouse. Predictive power of personality on music-genre exclusivity. In *ISMIR*, pages 652–658, Malaga, Spain, 2015.

[14] Michael Bar-Eli, Simcha Avugos, and Markus Raab. Twenty years of "hot hand" research: Review and critique. *Psychology of Sport and Exercise*, 7(6):525–553, 2006.

[15] Albert-László Barabási. *Network science*. Cambridge University Press, 2016.

[16] Natércia A. Batista, Michele A. Brandão, Gabriela B. Alves, Ana Paula Couto da Silva, and Mirella M. Moro. Collaboration strength metrics and analyses on github. In *Proceedings of the International Conference on Web Intelligence, Leipzig, Germany, August 23-26, 2017*, pages 170–178, 2017. doi: 10.1145/3106426.3106480. URL https://doi.org/10.1145/3106426.3106480.

[17] Gregory S Berns and Sara E Moore. A neural predictor of cultural popularity. *Journal of Consumer Psychology*, 22(1):154–160, 2012. ISSN 1057-7408. doi: 10.1016/j.jcps.2011.05.001. URL https://doi.org/10.1016/j.jcps.2011.05.001.

[18] Kerstin Bischoff, Claudiu S. Firan, Mihai Georgescu, Wolfgang Nejdl, and Raluca Paiu. Social knowledge-driven music hit prediction. In *Advanced Data Mining and Applications, 5th International Conference*, pages 43–54, New York, NY, USA, 2009. Springer. doi: 10.1007/978-3-642-03348-3_8. URL https://doi.org/10.1007/978-3-642-03348-3_8.

[19] Dmitry Bogdanov et al. The acousticbrainz genre dataset: Multi-source, multi-level, multi-label, and large-scale. In *ISMIR*, pages 360–367, NYC, USA, 2019.

[20] Brian Brost, Rishabh Mehrotra, and Tristan Jehan. The music streaming sessions dataset. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2594–2600, 2019. doi: 10.1145/3308558.3313641. URL https://doi.org/10.1145/3308558.3313641.

[21] Nicholas J. Bryan and Ge Wang. Musical influence network analysis and rank of sample-based music. In *ISMIR*, pages 329–334, Miami, USA, 2011.

[22] Andrzej Buda and Andrzej Jarynowski. Exploring patterns in european singles charts. In *2015 Second European Network Intelligence Conference*, pages 135–139, New York, NY, USA, 2015. IEEE Computer Society. doi: 10.1109/ENIC.2015.27. URL https://doi.org/10.1109/ENIC.2015.27.

[23] Raymond B Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1(2):245–276, 1966.

[24] Costin Chiru and Oana-Georgiana Popescu. Automatically determining the popularity of a song. In *Rough Sets - International Joint Conference*, pages 392–406, New York, NY, USA, 2017. Springer. doi: 10.1007/978-3-319-60837-2_33. URL https://doi.org/10.1007/978-3-319-60837-2_33.

[25] 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, AMCMM*, pages 83–88, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595935010. doi: 10.1145/1178723.1178736. URL https://doi.org/10.1145/1178723.1178736.

[26] Débora C. Corrêa, Alexandre L. M. Levada, and Luciano da F. Costa. Finding community structure in music genres networks. In *ISMIR*, pages 447–452, Miami, USA, 2011.

[27] Alberto Cosimato et al. The conundrum of success in music: Playing it or talking about it? *IEEE Access*, 7:123289–123298, 2019. doi: 10.1109/ACCESS.2019.2937743.

[28] Anna B Costello and Jason Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation*, 10(1):7, 2005.

[29] Baptiste de La Robertie, Yoann Pitarch, and Olivier Teste. Measuring article quality in wikipedia using the collaboration network. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015, Paris, France, August 25 - 28, 2015*, pages 464–471. ACM, 2015. doi: 10.1145/2808797.2808895. URL https://doi.org/10.1145/2808797.2808895.

[30] Sanjeev Dewan and Jui Ramaprasad. Social media, traditional media, and music sales. *MIS Quarterly*, 38(1):101–121, mar 2014. ISSN 0276-7783. doi: 10.25300/MISQ/2014/38.1.05. URL https://doi.org/10.25300/MISQ/2014/38.1.05.

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

[32] Elena V. Epure, Anis Khlif, and Romain Hennequin. Leveraging knowledge bases and parallel annotations for music genre translation. In *ISMIR*, pages 839–846, Delft, the Netherlands, 2019.

[33] Martin Ester et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Procs. of KDD*, pages 226–231, Portland, USA, 1996.

[34] Jianyu Fan and Michael Casey. Study of chinese and uk hit songs prediction. In *Proceedings of International Symposium on Computer Music Multidisciplinary Research*, pages 640–652, Marseille, France, 2013. The Laboratory of Mechanics and Acoustics. URL https://cmmr2013.prism.cnrs.fr/Docs/CMMR2013Proceedings.pdf.

[35] Limisgy Ramadhina Febirautami, Isti Surjandari, and Enrico Laoh. Determining characteristics of popular local songs in indonesia's music market. In *5th International Conference on Information Science and Control Engineering*, pages 197–201, New York, NY, USA, 2018. IEEE. doi: 10.1109/ICISCE.2018.00050. URL https://doi.org/10.1109/ICISCE.2018.00050.

[36] Bruce Ferwerda, Emily Yang, Markus Schedl, and Marko Tkalcic. Personality and taxonomy preferences, and the influence of category choice on the user experience for music streaming services. *Multimedia Tools Appl.*, 78(14):20157–20190, 2019. doi: 10.1007/s11042-019-7336-7. URL https://doi.org/10.1007/s11042-019-7336-7.

[37] Klaus Frieler, Kelly Jakubowski, and Daniel Müllensiefen. Is it the song and not the singer? hit song prediction using structural features of melodies. *Jahrbuch Musikpsychologie*, 25:41–54, dec 2015.

[38] Kiran Garimella and Robert West. Hot streaks on social media. In *International Conference on Web and Social Media*, pages 170–180. AAAI Press, 2019.

[39] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019. ISBN 978-1-491-96229-9.

[40] Soumya Suvra Ghosal and Indranil Sarkar. Novel approach to music genre classification using clustering augmented learning method (CALM). In *AAAI MAKE*, volume 2600 of *CEUR Workshop Proceedings*, 2020.

[41] Thomas Gilovich, Robert Vallone, and Amos Tversky. The hot hand in basketball: On the misperception of random sequences. *Cognitive psychology*, 17(3):295–314, 1985.

[42] Mark S Granovetter. The strength of weak ties. *American journal of sociology*, 78 (6):1360–1380, 1973.

[43] Michael Hahsler, Matthew Piekenbrock, and Derek Doran. dbscan: Fast density-based clustering with r. *Journal of Statistical Software*, 25:409–416, 2019.

[44] Jonathas G. D. Harb and Karin Becker. Emotion analysis of reaction to terrorism on twitter. In *SBBD*, pages 97–108. SBC, 2018.

[45] Gary W Heiman. *Understanding research methods and statistics: An integrated introduction for psychology*. Houghton, Mifflin and Company, 2001.

[46] Sumyea Helal. Subgroup discovery algorithms: A survey and empirical evaluation. *J. Comput. Sci. Technol.*, 31(3):561–576, 2016. doi: 10.1007/s11390-016-1647-1. URL https://doi.org/10.1007/s11390-016-1647-1.

[47] Darryll Hendricks, Jayendu Patel, and Richard Zeckhauser. Hot hands in mutual funds: Short-run persistence of relative performance, 1974–1988. *The Journal of finance*, 48(1):93–130, 1993.

[48] Romain Hennequin, Jimena Royo-Letelier, and Manuel Moussallam. Audio based disambiguation of music genre tags. In *ISMIR*, pages 645–652, Paris, France, 2018.

[49] Dorien Herremans and Tom Bergmans. Hit song prediction based on early adopter data and audio features. In *International Society for Music Information Retrieval Conference, ISMIR - Late Breaking Demo*, Suzhou, China, 2017. ISMIR.

[50] Dorien Herremans, David Martens, and Kenneth Sörensen. Dance hit song prediction. *Journal of New Music Research*, 43(3):291–302, 2014. doi: 10.1080/09298215.2014.881888.

[51] Lloyd G Humphreys and Richard G Montanelli Jr. An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, 10(2):193–205, 1975.

[52] Laura Igual and Santi Seguí. *Introduction to Data Science - A Python Approach to Concepts, Techniques and Applications*. Undergraduate Topics in Computer Science. Springer, 2017.

[53] 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, may 2018. ISSN 2054-5703. doi: 10.1098/rsos.171274. URL https://doi.org/10.1098/rsos.171274.

[54] Milan Janosov, Federico Battiston, and Roberta Sinatra. Success and luck in creative careers. *EPJ Data Sci.*, 9(1):9, 2020. doi: 10.1140/epjds/s13688-020-00227-w.

[55] Eamonn J. Keogh and Michael J. Pazzani. Scaling up dynamic time warping for datamining applications. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 285–289. ACM, 2000. doi: 10.1145/347090.347153.

[56] Yekyung Kim, Bongwon Suh, and Kyogu Lee. # nowplaying the future billboard: mining music listening behaviors of twitter users for hit song prediction. In *Proceedings of the First International Workshop on Social Media Retrieval and Analysis*, pages 51–56, New York, NY, USA, 2014. ACM. doi: 10.1145/2632188.2632206. URL https://doi.org/10.1145/2632188.2632206.

[57] Willi Klösgen and Jan M Zytkow. *Handbook of data mining and knowledge discovery*. Oxford University Press, Inc., 2002.

[58] Noam Koenigstein, Yuval Shavitt, and Noa Zilberman. Predicting billboard success using data-mining in p2p networks. In *2009 11th IEEE International Symposium on Multimedia*, pages 465–470, San Diego, USA, 2009. IEEE. doi: 10.1109/ISM.2009.73.

[59] Clement Laroche et al. Genre specific dictionaries for harmonic/percussive source separation. In *ISMIR*, pages 407–413, NYC, USA, 2016.

[60] 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*, pages 3–6, New York, NY, USA, 2015. ACM. doi: 10.1145/2802558.2814645. URL https://doi.org/10.1145/2802558.2814645.

[61] Junghyuk Lee and Jong-Seok Lee. Music popularity: Metrics, characteristics, and audio-based prediction. *IEEE Transactions on Multimedia*, 20(11):3173–3182, mar 2018. ISSN 1941-0077. doi: 10.1109/TMM.2018.2820903. URL https://doi.org/10.1109/TMM.2018.2820903.

[62] Florian Lemmerich and Martin Becker. pysubgroup: Easy-to-use subgroup discovery in python. In *ECML/PKDD (3)*, volume 11053 of *Lecture Notes in Computer Science*, pages 658–662. Springer, 2018.

[63] Hanchao Li, Zhouhemu Tang, Xiang Fei, Kuo-Ming Chao, Ming Yang, and Chaobo He. A survey of audio MIR systems, symbolic MIR systems and a music definition language demo-system. In *14th IEEE International Conference on e-Business Engineering*, pages 275–281, New York, NY, USA, 2017. IEEE Computer Society. doi: 10.1109/ICEBE.2017.51. URL https://doi.org/10.1109/ICEBE.2017.51.

[64] David Liben-Nowell and Jon M. Kleinberg. The link-prediction problem for social networks. *J. Assoc. Inf. Sci. Technol.*, 58(7):1019–1031, 2007.

[65] Lu Liu, Yang Wang, Roberta Sinatra, C Lee Giles, Chaoming Song, and Dashun Wang. Hot streaks in artistic, cultural, and scientific careers. *Nature*, 559(7714): 396–399, 2018. doi: 10.1038/s41586-018-0315-8.

[66] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6):1150–1170, 2011.

[67] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774, 2017.

[68] 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, feb 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.2976033. URL https://doi.org/10.1109/ACCESS.2020.2976033.

[69] Yui Matsumoto, Ryosuke Harakawa, Takahiro Ogawa, and Miki Haseyama. Context-aware network analysis of music streaming services for popularity estimation of artists. *IEEE Access*, 8:48673–48685, mar 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.2978281. URL https://doi.org/10.1109/ACCESS.2020.2978281.

[70] Kai Middlebrook and Kian Sheik. Song hit prediction: Predicting billboard hits using spotify data. *CoRR*, abs/1908.08609, 2019. URL http://arxiv.org/abs/1908.08609.

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

[72] 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 Comput. Surv.*, 51(3):45:1–45:46, June 2018. ISSN 0360-0300. doi: 10.1145/3177849. URL https://doi.org/10.1145/3177849.

[73] Mark E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010. ISBN 978-0-19920665-0. doi: 10.1093/ACPROF:OSO/9780199206650.001.0001.

[74] Minh-Tri Nguyen, Duong H. Le, Takuma Nakajima, Masato Yoshimi, and Nam Thoai. Attention-based neural network: A novel approach for predicting the popularity of online content. In *21st IEEE International Conference on High Performance Computing and Communications; 17th IEEE International Conference on Smart City; 5th IEEE International Conference on Data Science and Systems, HPCC/SmartCity/DSS 2019, Zhangjiajie, China, August 10-12, 2019*, pages 329–336, 2019. doi: 10.1109/HPCC/SmartCity/DSS.2019.00058. URL https://doi.org/10.1109/HPCC/SmartCity/DSS.2019.00058.

[75] 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. ACM.

[76] Joseph C Nunes and Andrea Ordanini. I like the way it sounds: The influence of instrumentation on a pop song's place in the charts. *Musicae Scientiae*, 18(4): 392–409, sep 2014. doi: 10.1177/1029864914548528. URL https://doi.org/10.1177/1029864914548528.

[77] Gabriel P. Oliveira, Anisio Lacerda, and Mirella M. Moro. Musical genre analysis over dynamic success-based networks. In *35th Brazilian Symposium on Databases – Workshop on Thesis and Dissertations in Databases*, Porto Alegre, Brazil, 2020. SBC.

[78] Gabriel P. Oliveira, Mariana O. Silva, Danilo B. Seufitelli, Anisio Lacerda, and Mirella M. Moro. MGD: Music Genre Dataset, October 2020. URL https://doi.org/10.5281/zenodo.4778563. https://doi.org/10.5281/zenodo.4778563.

[79] Gabriel P. Oliveira, Mariana O. Silva, Danilo B. Seufitelli, Anisio Lacerda, and Mirella M. Moro. Detecting collaboration profiles in success-based music genre networks. In *Proceedings of the 21st International Society for Music Information Retrieval Conference*, pages 726–732, Montreal, Canada, 2020. ISMIR.

[80] Gabriel P. Oliveira, Gabriel R. G. Barbosa, Bruna C. Melo, Mariana O. Silva, Danilo B. Seufitelli, Anisio Lacerda, and Mirella M. Moro. MUHSIC: Music-oriented

Hot Streak Information Collection, May 2021. URL https://doi.org/10.5281/zenodo.4779003. https://doi.org/10.5281/zenodo.4779003.

[81] Gabriel P. Oliveira, Gabriel R. G. Barbosa, Bruna C. Melo, Mariana O. Silva, Danilo B. Seufitelli, and Mirella M. Moro. MUHSIC: An open dataset with temporal musical success information. In *36th Brazilian Symposium on Databases - Dataset Showcase Workshop*, pages 65–76. SBC, 2021. https://doi.org/10.5753/dsw.2021.17415.

[82] Gabriel P. Oliveira, Gabriel R. G. Barbosa, Bruna C. Melo, Juliana E. Botelho, Mariana O. Silva, Danilo B. Seufitelli, and Mirella M. Moro. Musical success in the united states and brazil: Novel datasets and temporal analyses. *Journal of Information and Data Management*, 13(1):111–126, 2022. URL https://doi.org/10.5753/jidm.2022.2350.

[83] Sergio Oramas et al. Multi-label music genre classification from audio, text and images using deep features. In *ISMIR*, pages 23–30, Suzhou, China, 2017.

[84] François Pachet and Pierre Roy. Hit song science is not yet a science. In *ISMIR*, pages 355–360, 2008.

[85] Matthew Prockup et al. Modeling genre with the music genome project: Comparing human-labeled attributes and audio features. In *ISMIR*, pages 31–37, Malaga, Spain, 2015.

[86] Markus Raab, Bartosz Gula, and Gerd Gigerenzer. The hot hand exists in volleyball and is used for allocation decisions. *Journal of Experimental Psychology: Applied*, 18(1):81, 2012.

[87] Matthew Rabin and Dimitri Vayanos. The gambler's and hot-hand fallacies: Theory and applications. *The Review of Economic Studies*, 77(2):730–778, 2010.

[88] R Rajyashree, Anmol Anand, Yash Soni, and Harshitaa Mahajan. Predicting hit music using midi features and machine learning. In *International Conference on Communication and Electronics Systems*, pages 94–98, New York, NY, USA, 2018. IEEE. doi: 10.1109/CESYS.2018.8724001. URL https://doi.org/10.1109/CESYS.2018.8724001.

[89] Agha Haider Raza and Krishnadas Nanath. Predicting a hit song with machine learning: Is there an apriori secret formula? In *2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*, pages 111–116, New York, NY, USA, 2020. IEEE. doi: 10.1109/DATABIA50434.2020.9190613. URL https://doi.org/10.1109/DATABIA50434.2020.9190613.

[90] Cláudio Rebelo de Sá, Wouter Duivesteijn, Paulo J. Azevedo, Alípio Mário Jorge, Carlos Soares, and Arno J. Knobbe. Discovering a taste for the unusual: exceptional models for preference mining. *Mach. Learn.*, 107(11):1775–1807, 2018. doi: 10.1007/s10994-018-5743-z. URL https://doi.org/10.1007/s10994-018-5743-z.

[91] Jing Ren and Robert J. Kauffman. Understanding music track popularity in a social network. In *25th European Conference on Information Systems*, pages 374–388, Atlanta, GA, USA, 2017. AIS. URL http://aisel.aisnet.org/ecis2017_rp/25.

[92] 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*, pages 95–96, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. doi: 10.1145/2872518.2889402. URL https://doi.org/10.1145/2872518.2889402.

[93] William R Revelle. psych: Procedures for personality and psychological research, 2017.

[94] 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, feb 2006. ISSN 1095-9203. doi: 10.1126/science.1121066. URL https://doi.org/10.1126/science.1121066.

[95] Hendrik Schreiber. Genre ontology learning: Comparing curated with crowd-sourced ontologies. In *ISMIR*, pages 400–406, NYC, USA, 2016.

[96] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014.

[97] Seungkyu Shin and Juyong Park. On-chart success dynamics of popular songs. *Advances in Complex Systems*, 21(3-4):1850008, 2018. doi: 10.1142/S021952591850008X. URL https://doi.org/10.1142/S021952591850008X.

[98] 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, Palo Alto, CA, USA, 2016. AAAI Press. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13129.

[99] Mariana O. Silva and Mirella M. Moro. Causality analysis between collaboration profiles and musical success. In *WebMedia*, pages 369–376. ACM, 2019.

[100] Mariana O. Silva, Laís M. Rocha, and Mirella M. Moro. Collaboration Profiles and Their Impact on Musical Success. In *Procs. of ACM/SIGAPP SAC*, pages 2070–2077, Limassol, Cyprus, 2019. doi: 10.1145/3297280.3297483.

[101] Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. Quantifying the evolution of individual scientific impact. *Science*, 354 (6312), 2016.

[102] Abhishek Singhi and Daniel G Brown. Hit song detection using lyric features alone. In *15th International Society for Music Information Retrieval Conference, ISMIR - Late-Breaking Demo*, Taipei, Taiwan, 2014. ISMIR.

[103] Abhishek Singhi and Daniel G Brown. Can song lyrics predict hits. In *Proceedings of the 11th International Symposium on Computer Music Multidisciplinary Research*, pages 457–471, Marseille, France, 2015. The Laboratory of Mechanics and Acoustics. URL https://cmmr2019.prism.cnrs.fr/Docs/proceedingsCMMR2015.pdf.

[104] James Sundali and Rachel Croson. Biases in casino betting: The hot hand and the gambler's fallacy. *Judgement and Decision Making*, 1(1):1, 2006.

[105] Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, and Eli Woods. Tslearn, A machine learning toolkit for time series data. *J. Mach. Learn. Res.*, 21:118:1–118:6, 2020.

[106] Alexandros Tsaptsinos. Lyrics-based music genre classification using a hierarchical attention network. In *ISMIR*, pages 694–701, Suzhou, China, 2017.

[107] Eleana Tsiara and Christos Tjortjis. Using twitter to predict chart position for songs. In *Artificial Intelligence Applications and Innovations - 16th IFIP*, pages 62–72, New York, NY, USA, 2020. Springer. doi: 10.1007/978-3-030-49161-1_6. URL https://doi.org/10.1007/978-3-030-49161-1_6.

[108] John W Tukey. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114, 1949.

[109] Igor Vatolkin, Günter Rudolph, and Claus Weihs. Evaluation of album effect for feature selection in music genre recognition. In *ISMIR*, pages 169–175, Malaga, Spain, 2015.

[110] Wladston Viana, Ana Paula Couto da Silva, and Mirella M. Moro. Pick the right team and make a blockbuster: a social analysis through movie history. In *SAC*, pages 1108–1114. ACM, 2016.

[111] Xiaomeng Wang, Binxing Fang, Hongli Zhang, and Xing Wang. Analyzing and predicting the popularity of online content using the weak ties theory. In *21st IEEE International Conference on High Performance Computing and Communications; 17th IEEE International Conference on Smart City; 5th IEEE International Conference*

on Data Science and Systems, HPCC/SmartCity/DSS 2019, Zhangjiajie, China, August 10-12, 2019, pages 1743–1748, 2019. doi: 10.1109/HPCC/SmartCity/DSS.2019.00239. URL https://doi.org/10.1109/HPCC/SmartCity/DSS.2019.00239.

[112] 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 *International Conference on Acoustics, Speech and Signal Processing*, pages 621–625, New York, NY, USA, 2017. IEEE. doi: 10.1109/ICASSP.2017.7952230. URL https://doi.org/10.1109/ICASSP.2017.7952230.

[113] Byungjoon Yoo and Kwansoo Kim. Online music ranking service: Ranking mechanism based on popularity and slot effect. In *Pacific Asia Conference on Information Systems*, pages 615–626, Atlanta, GA, USA, 2010. AISeL. URL http://aisel.aisnet.org/pacis2010/31.

[114] Haiqing Yu, Yanling Li, Shujun Zhang, and Chunyan Liang. Popularity prediction for artists based on user songs dataset. In *Proceedings of International Conference on Computing and Artificial Intelligence*, pages 17–24, New York, NY, USA, 2019. ACM. doi: 10.1145/3330482.3330493. URL https://doi.org/10.1145/3330482.3330493.

[115] Mohammed J. Zaki and Wagner Meira Jr. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014. ISBN 9780521766333.

[116] Eva Zangerle et al. Hit song prediction: Leveraging low- and high-level audio features. In *ISMIR*, pages 319–326, Delft, the Netherlands, 2019.

Appendix A

Hot Streak Prediction

In this appendix, we present the details of the Hot Streak Prediction model from Chapter 4. We first describe the features considered in our model (Section A.1) and then detail our experimental setup (Section A.2).

A.1 Feature Description

In this section, we provide a brief description on the features used in our prediction model. In our model, each instance represent a week in Hot 100 for a given music genre (i.e., the set of songs from artists belonging to that genres). The instances are described by a feature set, which is divided into three main categories: genre-related, artist-related, and song-related. All features are obtained from Spotify through its API[1]. Here, we made an effort in keeping the definitions at simple as possible, as formally presenting them is beyond the scope of this work.

1. **Genre-related features**
   - **genre:** the genre name, as provided by Spotify;
   - **num_genre_songs:** number of songs from artists belonging to the given genre;
   - **num_distinct_artists**: number of distinct artists belonging to the genre;
2. **Artist-related features**
   - **median_artists_per_song:** median number of artists per song;
   - **median_career_time:** median career time of artists, from their debut to the week in subject (in days);
   - **median_genres_per_artist:** the median number of distinct genres per artist, as an artist may belong to more than one genre;

[1]Spotify for Developers: https://developer.spotify.com/

3. **Song-related features**

- **num_collab_songs:** number of songs that are collaborations (i.e., interpreted by two or more artists);
- **num_explicit_songs:** number of songs marked with the *Explicit* tag from Spotify;
- **Acoustic features:** related to the audio content, median values for all songs belonging to a genre.
  - **danceability:** informs whether a song is suitable for dancing or not in terms of probability;
  - **energy:** the intensity and activity of a song considering information such as dynamic range, perceived loudness, timbre, onset rate, and entropy;
  - **key:** the estimated overall key of a song, mapped as an integer number (e.g. $C = 0$, $C\# = 1$);
  - **loudness:** the general loudness measured in decibels (dB);
  - **mode:** the general modality of a song (i.e., major= 1 or minor= 0);
  - **speechiness:** the probability of a given song to have spoken words in it;
  - **acousticness:** the probability of a song to be acoustic or not;
  - **instrumentalness:** the probability of a song to be instrumental, i.e., not contain vocals;
  - **liveness:** the probability of a song being performed live, i.e., the presence of an audience in a song;
  - **valence:** the positiveness within a song, in which high valence values represent happier songs, whereas low values means the opposite;
  - **tempo:** the speed of the song, measured in beats per minute (BPM);
  - **time_signature:** the amount of beats in each bar (measure);
  - **duration_ms:** the duration of songs in milliseconds.

A.2 Experimental Setup Details

Here, we give further information on our experimental evaluation by focusing on the parameters of the classifiers used in our model. We developed all experiments using the Python package Scikit-Learn[2]. In our modeling, 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. Table A.1 presents the parameters tuned for each classifier as well as the considered search space.

Table A.1: Parameter grid for tuning the hyperparameters for each considered classifier, with the best values underlined (evaluated by F1-Score).

Classifier	**Hyperparameter**	**Search space**
LR	`penalty`	['l1', 'l2']
	`tol`	[1e-6, 1e-4, 1e-2]
	`C`	[0.01, 0.1, 1.0]
	`solver`	['lbfgs', 'liblinear', 'saga']
LinearSVC	`penalty`	['l1', 'l2']
	`tol`	[1e-06, 1e-04, 1e-02]
	`C`	[0.01, 0.1, 1.0]
	`loss`	['hinge', 'squared_hinge']
Perceptron	`penalty`	['l1', 'l2']
	`tol`	[1e-5, 1e-3, 1e-1]
	`alpha`	[1e-6, 1e-4, 1e-2]
SGD	`penalty`	['l1', 'l2']
	`tol`	[1e-5, 1e-3, 1e-1]
	`alpha`	[1e-6, 1e-4, 1e-2]
	`loss`	['hinge', 'squared_hinge', 'log', 'modified_huber']

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

Appendix B

Genre Profiling Process

Here, we detail the genre profiling process from Chapter 5. We first describe data and network characterization (Section B.1) and then explain the Exploratory Factor Analysis (Section B.2). Finally, we detail the cluster analysis step (Section B.3).

B.1 Data Processing and Network Characterization

When processing Spotify chart data, we make some decisions regarding the definition of success of a genre collaboration and the genre reduction through our mapping:

- In our chart analyses, we consider the number of streams as the success measure for a hit song. Therefore, in our temporal and regional analyses, we define the success of a genre collaboration (i.e. the edge weight) as the average value of total streams of all songs involving those genres within the considered market and period.
- In the mapping proccess from the Spotify-assigned genres to our *super-genres*, we detect 76 out of 896 genres which do not fit into any category. For example, *talent show*, in which artists may belong to other well-established genres (e.g. *pop*, *country* or *hip hop*). Thus, these genres are categorized as *other*. As Spotify artists can be assigned to more than one genre, this categorization do not prejudice our further analyses.

The complete global and regional chart overview and the full characterization of the networks are presented by Tables B.1 and B.2, respectively. To compute the network metrics, we use *NetworkX*[1], a network analysis Python package.

[1]NetworkX: https://networkx.github.io/

Table B.1: Most popular music genres in each considered market in the years 2017, 2018 and 2019.

	2017			2018			2019		
	Genre	**Songs**	**Arts.**	**Genre**	**Songs**	**Arts.**	**Genre**	**Songs**	**Arts.**
Global	pop	635	252	pop	701	257	pop	678	256
	hip hop	362	101	rap	587	134	hip hop	432	180
	dance pop	346	121	hip hop	546	181	rap	412	123
	rap	344	86	pop rap	398	93	trap	319	103
	pop rap	286	81	trap	354	96	dance pop	288	94
Australia	pop	635	266	pop	718	262	pop	684	262
	dance pop	360	138	rap	427	106	rap	325	111
	hip hop	300	101	dance pop	358	116	dance pop	295	102
	rap	280	85	hip hop	358	113	hip hop	275	121
	pop rap	236	84	pop rap	300	86	pop rap	233	87
Brazil	pop	447	158	pop	468	166	pop	358	129
	dance pop	212	76	sertanejo	283	63	brazilian funk	280	114
	sertanejo	178	40	brazilian funk	276	111	sertanejo	265	66
	brazilian funk	170	67	dance pop	149	66	dance pop	103	39
	electro	142	54	electro	140	66	electro	101	41
Canada	pop	703	254	rap	850	151	pop	667	239
	rap	559	113	hip hop	732	152	rap	584	141
	hip hop	510	115	pop	715	245	hip hop	455	139
	pop rap	468	95	pop rap	628	111	pop rap	413	107
	trap	372	83	trap	534	107	trap	370	101
France	pop	1,000	271	pop	1,299	287	pop	1,176	276
	hip hop	770	135	hip hop	1,180	191	hip hop	984	177
	rap	728	125	rap	1,120	166	rap	899	153
	francoton	414	52	francoton	434	59	francoton	366	56
	dance pop	174	86	dance pop	162	73	dance pop	119	61
Germany	hip hop	796	171	hip hop	971	216	hip hop	1,048	238
	pop	621	290	pop	687	293	pop	635	281
	rap	372	105	rap	536	129	rap	429	134
	dance pop	299	126	dance pop	282	114	dance pop	210	86
	pop rap	177	72	pop rap	214	74	trap	166	71
Japan	pop	197	119	pop	417	152	j-pop	489	108
	j-pop	138	65	j-pop	387	125	pop	287	125
	dance pop	131	72	dance pop	259	80	j-rock	183	42
	r&b	89	43	r&b	165	63	dance pop	173	57
	rap	63	36	j-rock	164	55	other	125	43
UK	pop	682	246	pop	763	243	pop	665	234
	dance pop	383	131	hip hop	490	161	hip hop	441	152
	hip hop	355	109	rap	480	122	rap	360	105
	rap	276	83	dance pop	424	137	dance pop	296	107
	pop rap	224	78	pop rap	319	90	pop rap	209	72
USA	rap	673	122	rap	939	159	rap	715	165
	pop	650	210	hip hop	783	159	pop	635	203
	hip hop	594	124	pop rap	653	107	hip hop	548	156
	pop rap	540	93	pop	642	201	pop rap	492	110
	trap	444	91	trap	611	116	trap	488	121

Table B.2: Network characterization for all global and regional markets, grouped according to their similar network evolution. Underlined values are the highest metric value for a specific market throughout the considered period.

Metric	***Global***			***Group 1: USA & Canada***						***Group 3: Other English-speaking markets***					
				USA			*Canada*			*UK*			*Australia*		
	2017	**2018**	**2019**	**2017**	**2018**	**2019**	**2017**	**2018**	**2019**	**2017**	**2018**	**2019**	**2017**	**2018**	**2019**
G	72	79	89	76	73	83	70	71	82	74	76	79	65	71	79
C	564	583	709	542	522	670	540	558	680	610	605	627	512	514	577
AD	15.7	14.8	15.9	14.3	14.3	16.1	15.4	15.7	16.6	16.5	15.9	15.9	15.8	14.5	14.6
AWD	256.9	247.4	236.7	324.6	287.9	241.4	366.3	307.6	212.4	216.5	203.6	159.5	220.6	170.8	140.0
D	0.221	0.189	0.181	0.190	0.199	0.197	0.224	0.225	0.205	0.226	0.212	0.204	0.246	0.200	0.200
ACC	0.743	0.757	0.754	0.762	0.760	0.726	0.739	0.749	0.762	0.724	0.754	0.738	0.718	0.700	0.700
SL	24	21	28	25	22	27	22	23	31	28	25	30	22	23	25
IntraG	4.26%	3.60%	3.95%	4.61%	4.21%	4.03%	4.07%	4.12%	4.56%	4.59%	4.13%	4.78%	4.30%	4.47%	4.33%
InterG	95.74%	96.40%	96.05%	95.39%	95.79%	95.97%	95.93%	95.88%	95.44%	95.41%	95.87%	95.22%	95.70%	95.53%	95.67%

Metric	***Global***			***Group 2: Non-English speaking markets***											
				Brazil			*France*			*Germany*			*Japan*		
	2017	**2018**	**2019**	**2017**	**2018**	**2019**	**2017**	**2018**	**2019**	**2017**	**2018**	**2019**	**2017***	**2018**	**2019**
G	72	79	89	58	63	61	63	63	66	69	75	73	56	71	63
C	564	583	709	453	524	392	465	464	434	555	590	523	350	491	418
AD	15.7	14.8	15.9	15.6	16.6	12.9	14.8	14.7	13.2	16.1	15.7	14.3	12.5	13.8	13.3
AWD	256.9	247.4	236.7	136.1	133.0	95.3	185.1	213.2	153.2	213.8	196.6	152.2	84.3	121.7	68.3
D	0.221	0.189	0.181	0.274	0.268	0.214	0.238	0.238	0.202	0.237	0.200	0.200	0.227	0.198	0.214
ACC	0.743	0.757	0.754	0.770	0.758	0.677	0.778	0.772	0.773	0.759	0.800	0.700	0.748	0.765	0.697
SL	24	21	28	24	29	27	20	22	24	23	24	23	20	24	19
IntraG	4.26%	3.60%	3.95%	5.30%	5.53%	6.89%	4.30%	4.74%	5.53%	4.14%	4.07%	4.40%	5.71%	4.89%	4.55%
InterG	95.74%	96.40%	96.05%	94.70%	94.47%	93.11%	95.70%	95.26%	94.47%	95.86%	95.93%	95.60%	94.29%	95.11%	95.45%

G: number of genres (nodes). **C**: number of genre collaborations (edges). **AD**: average node degree. **AWD**: average node degree (weighted). **D**: network density. **ACC**: average clustering coefficient. **SL**: number of self-loops. **IntraG**: fraction of intra-genre collaborations. **InterG**: fraction of inter-genre collaborations.

***** As Spotify provides Japanese weekly charts only after 08/31/2017, we build Japan's 2017 genre network with data from then.

B.2 Exploratory Factor Analysis in R

To perform an exploratory factor analysis (EFA), we use the *psych* R package [93] with the `fa()` function. Our data consist of 27 music genre networks (i.e., nine music markets), containing six different topological metrics, described in Chapter 3: Weight (W), Common Neighbors (CN), Neighborhood Overlap (NO), Preferential Attachment (PA), Edge Betweenness (EB), and Resource Allocation (RA).

B.2.1 Choosing the Number of Factors

Before conducting the EFA, we must determine an acceptable number of factors. The *psych* package offers a few ways in which the number of factors can be decided. Here, we use the `fa.parallel()` function to obtain the suggested number of factors via the Parallel Analysis [51] criteria. The output gives us the textual output of the suggested number of factors and a *scree* plot [23] of the successive eigenvalues. Figures B.1, B.2 and B.3 show the resulting *scree* plot for each network.

In the *scree* plots generated, blue and red lines show eigenvalues of actual and simulated/resampled data (placed on top of each other), respectively. The number of factors is determined by looking at the large drops in the actual data and spot the point where it levels off to the right. Moreover, we must identify the inflection point where the gap between simulated and actual data tends to be minimum. Analyzing all 27 *scree* plots, we see that the vast majority suggest a number of factors equal to 3.

B.2.2 Factor Analysis

Once the number of suggested factors is determined using `fa.parallel()`, EFA can be performed using the `fa()` function. In order to enable reproducibility, we provide the model parameters settings, as summarized in Table B.3. Specifically, we use the well known Ordinary Least Squares (OLS) factoring method and an oblique rotation, allowing factors to correlate with each other. We can visualize the EFA results using the `fa.diagram()` function, where it takes a `fa()` result object and creates a path diagram showing actor loadings, ordered from strongest to weakest. Factor loadings represent the

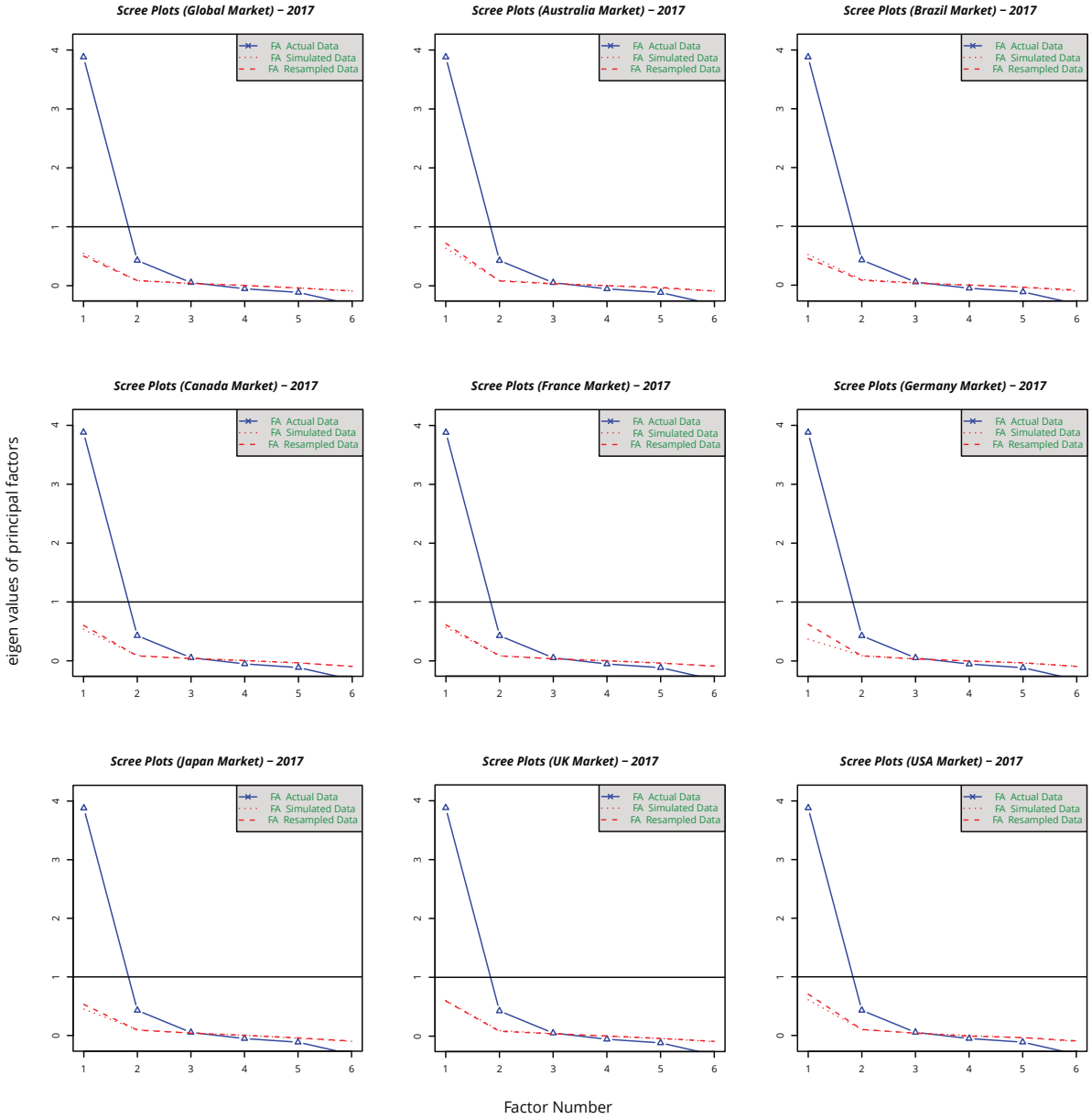


Figure B.1: Scree plots resulting from the Parallel Analysis for each genre network in 2017. Blue and red lines show eigenvalues of actual and simulated/resampled data, respectively. The suggested number of factors can be found in the X-axis position right before the "elbow" in the actual data curve.

correlation between each metric and the underlying factor, and they can range from -1 to 1. Figures B.4, B.5 and B.6 show the resulting factor loadings graph for each network.

Table B.3: Parameter Settings for Exploratory Factor Analysis

Parameters	**Description**	**Value**
`nfactors`	Number of factors to extract	3
`rotate`	Type of rotation	*oblimin*
`fm`	Factoring method	*ols*

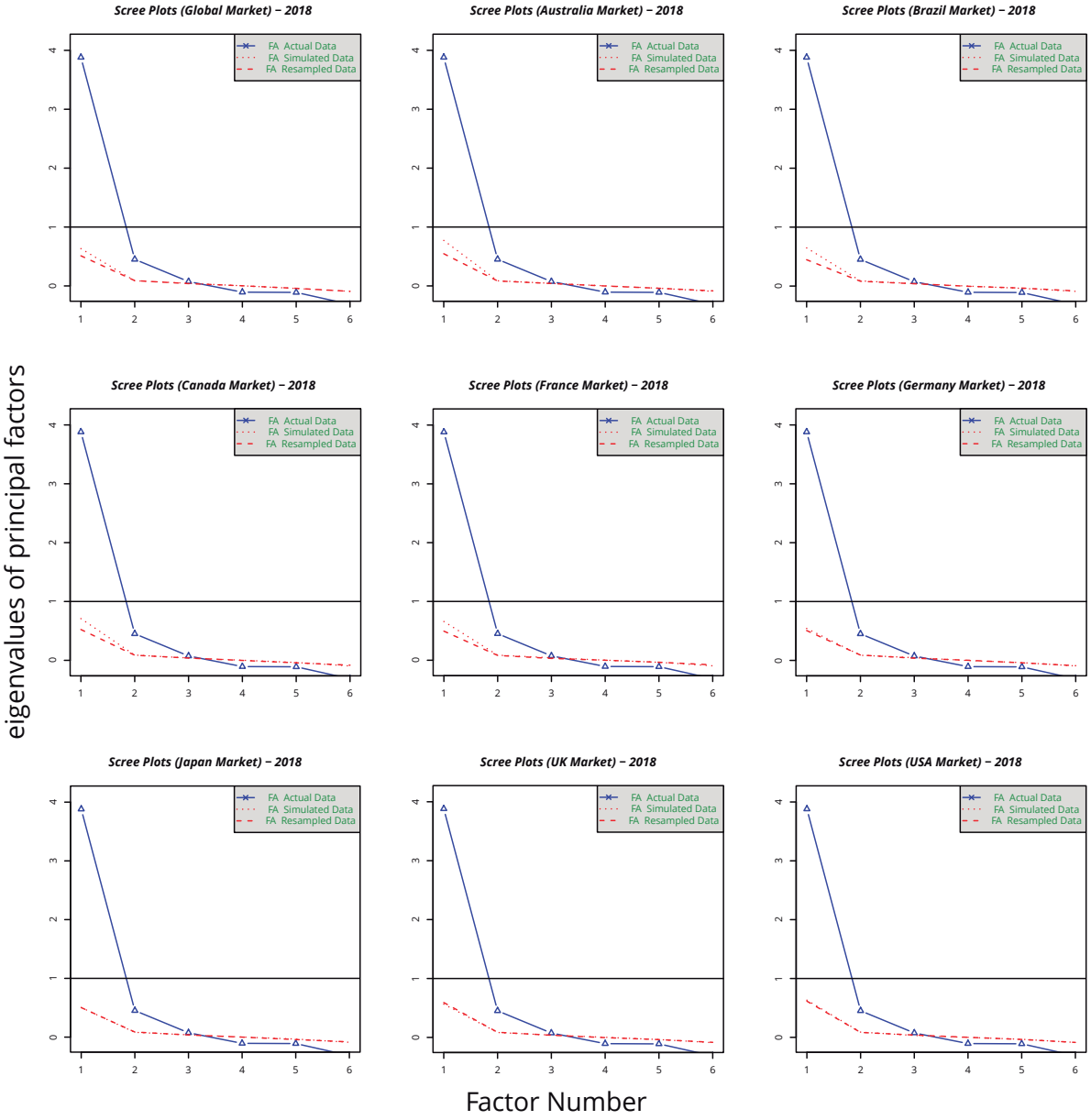


Figure B.2: Scree plots resulting from the Parallel Analysis for each genre network in 2018. Blue and red lines show eigenvalues of actual and simulated/resampled data, respectively. The suggested number of factors can be found in the X-axis position right before the "elbow" in the actual data curve.

B.3 Cluster Analysis - DBSCAN

We use the DBSCAN algorithm [33], which is a classical density-based clustering procedure. DBSCAN clusters the data points by separating areas of high density from areas of low density. It can be used not only to identify clusters of any shape, but also detect noise and outliers in the dataset. Two important parameters are required for DBSCAN:

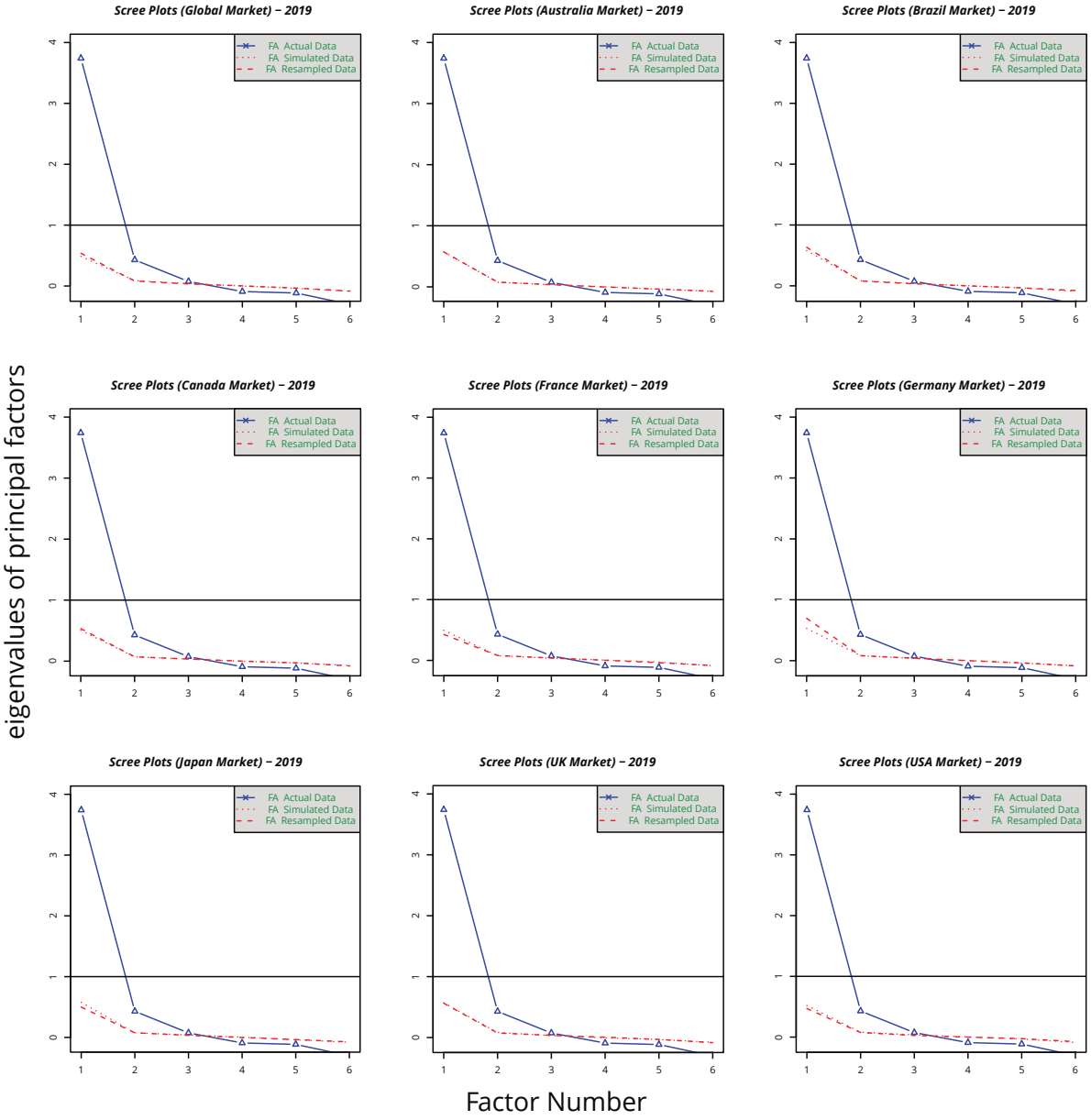


Figure B.3: Scree plots resulting from the Parallel Analysis for each genre network in 2019. Blue and red lines show eigenvalues of actual and simulated/resampled data, respectively. The suggested number of factors can be found in the X-axis position right before the "elbow" in the actual data curve.

1. ϵ: It defines the radius of neighborhood around a data point x. It is called as ϵ-neighborhood of x. Such parameter is crucial to choose appropriately. If the ϵ value is chosen too small then large part of the data will be considered as outliers. Otherwise, the clusters will merge and majority of the data points will be in the same clusters.

2. $MinPts$: Minimum number of neighbors (data points) required to form a dense cluster, within ϵ radius. As a general rule, the $MinPts$ can be derived from the number of dimensions D in the dataset as $MinPts \geq D + 1$. Also, the minimum

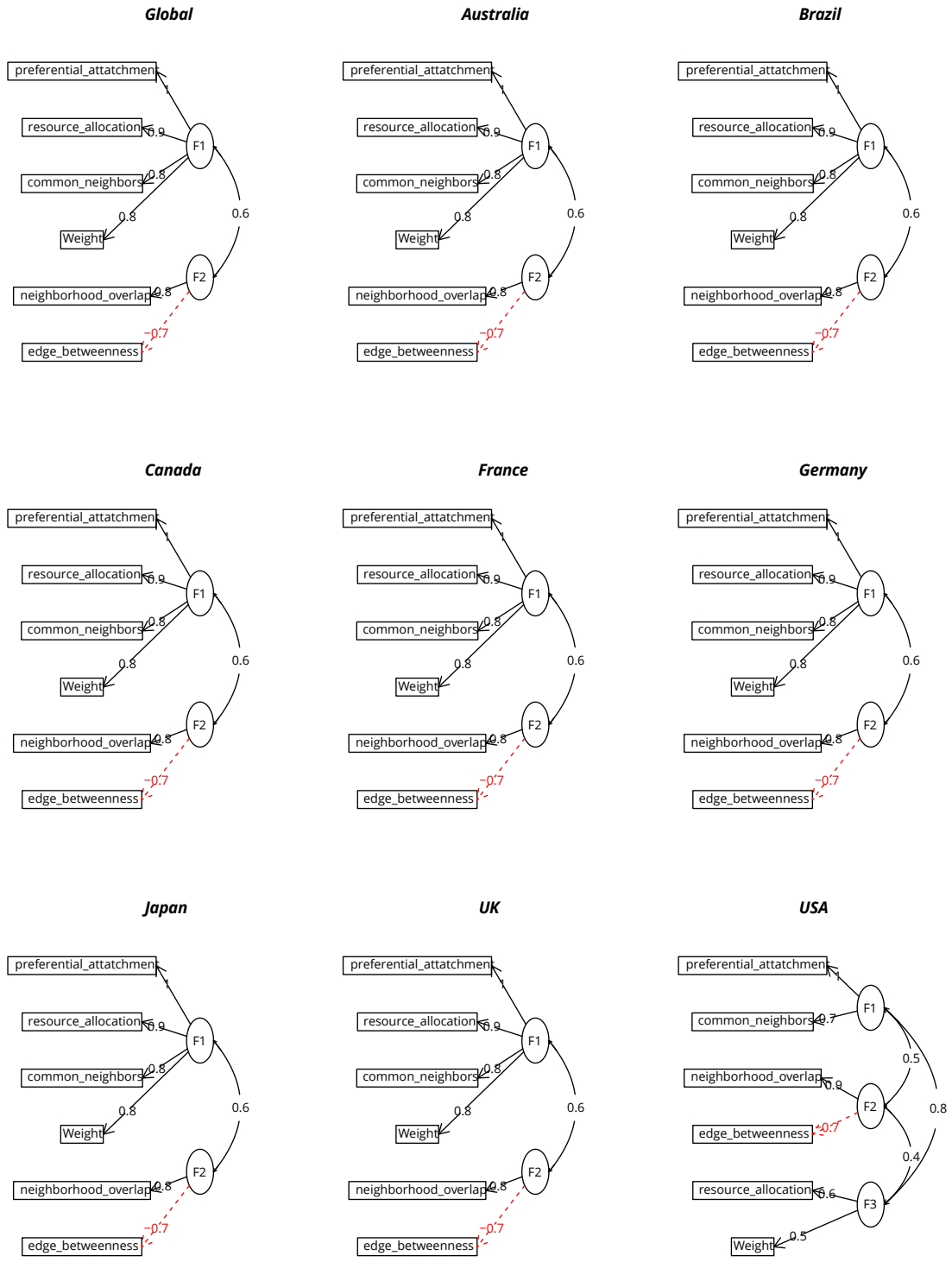


Figure B.4: Exploratory Factor Analysis diagram for each genre network in 2017. Solid and dashed lines represent positive and negative correlations, respectively.

value of $MinPts$ is at least 3.

Here, we use the *dbscan* R package [43] with the `dbscan()` function. As our dataset has six distinct dimensions (i.e. metrics), we set $MinPts = 7$. Then, to choose the optimal ϵ value, *dbscan* relies on a space-partitioning data structure called a *k-d trees*. This data structure allows us to identify the kNN or all neighbors within a fixed radius ϵ. We now

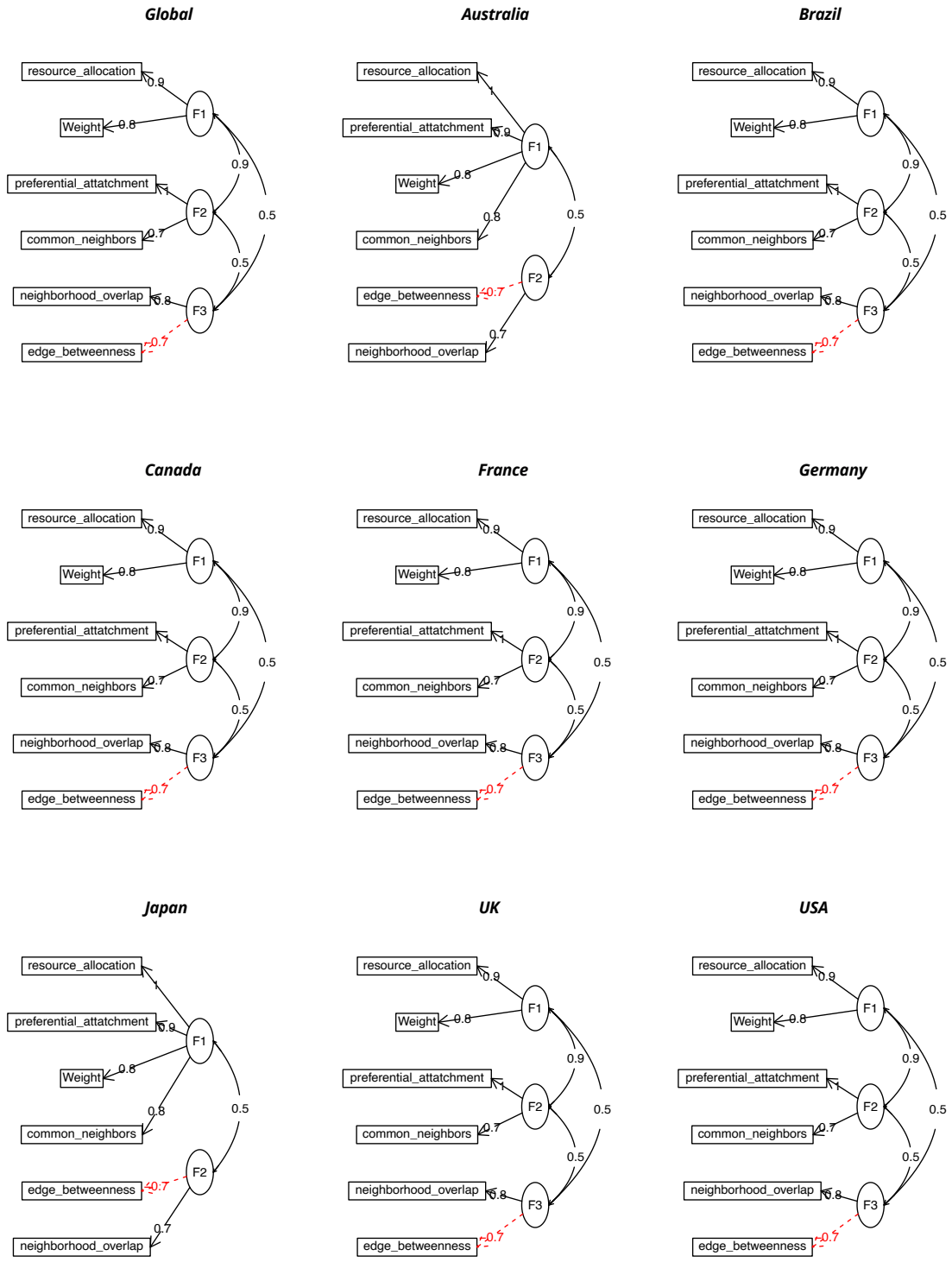


Figure B.5: Exploratory Factor Analysis diagram for each genre network in 2018. Solid and dashed lines represent positive and negative correlations, respectively.

execute the function `kNNdistplot()` with $k = 7$ (must be equal to $MinPts$) to plot the k-distances, which are the average distance of a point to its k-nearest neighbors. Finally, we set ϵ as the k value where where a sharp change take place in the curve. Figures B.7, B.8 and B.9 present such plots for all markets in 2017, 2018 and 2019, respectively. In all markets and years, we observe that this threshold occurs close to $k = 1$, thus being chosen

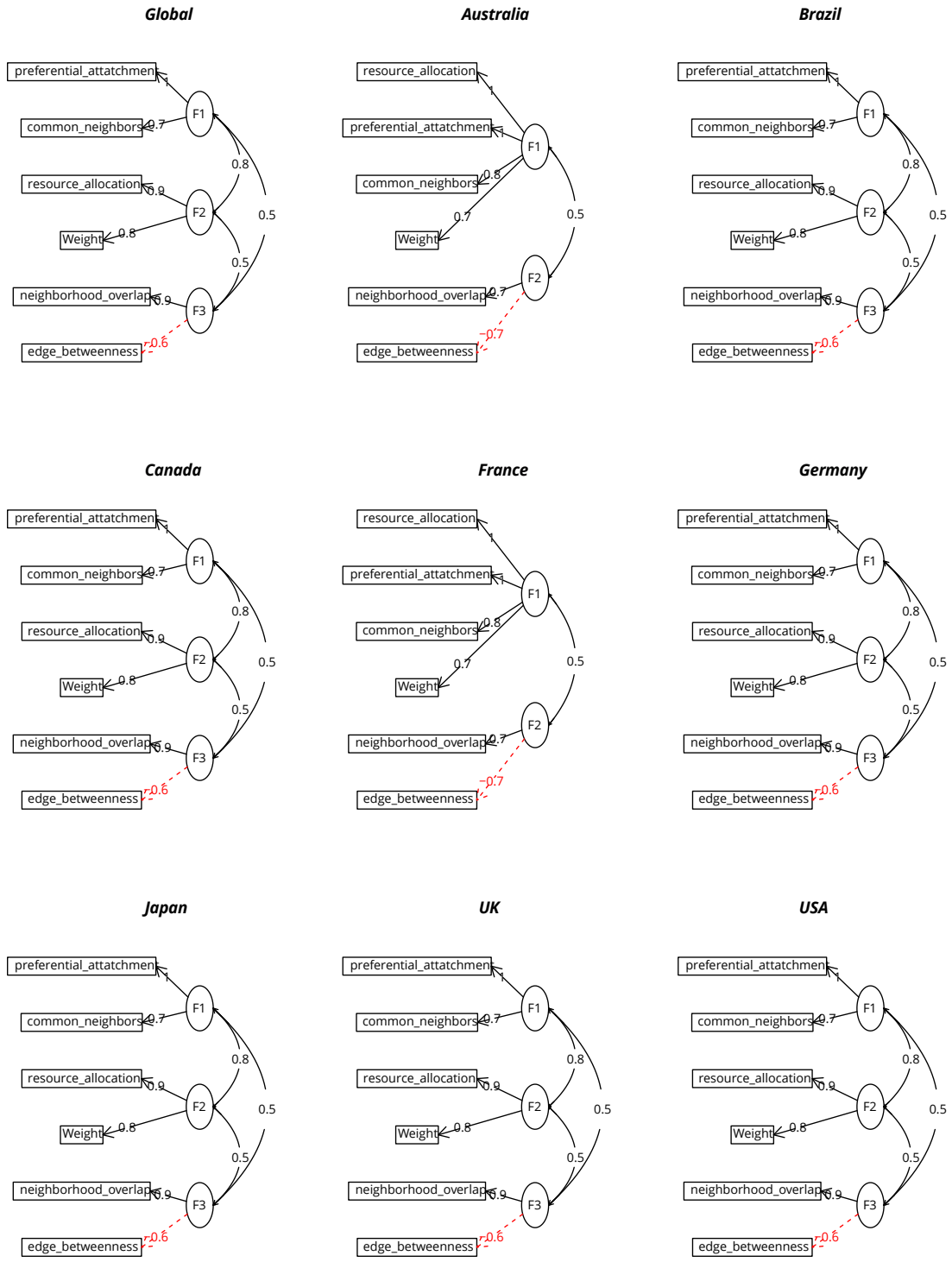


Figure B.6: Exploratory Factor Analysis diagram for each genre network in 2019. Solid and dashed lines represent positive and negative correlations, respectively.

as our ϵ value. The resulting clusters for each market throughout the years are shown by Figures B.10, B.11 and B.12, while the resulting collaboration profiles are presented by Figure B.13.

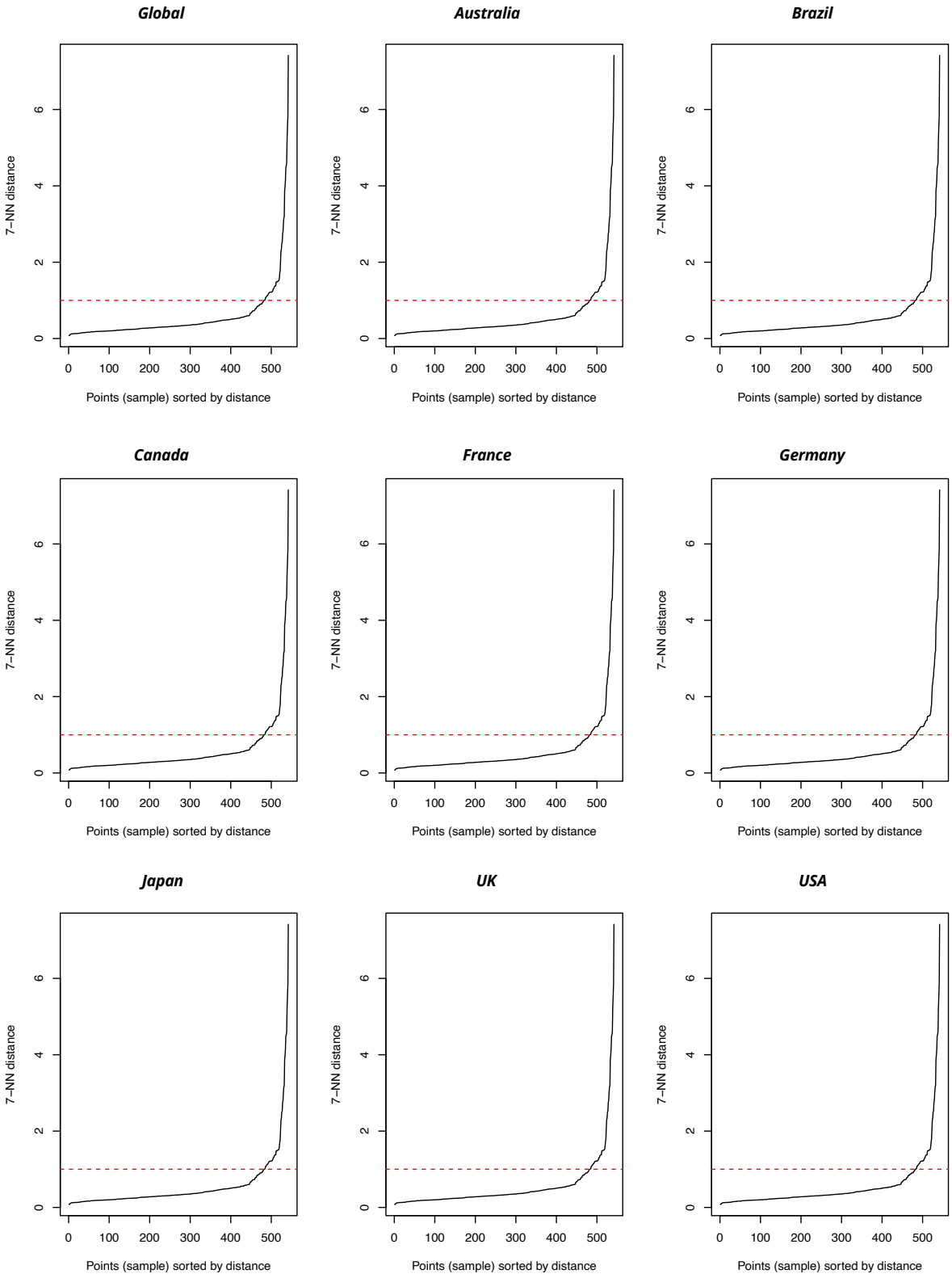


Figure B.7: 7-NN distance plot for each genre network in 2017. Dashed lines represent the threshold where a major change occurs in the curve, chosen as the ϵ parameter of DBSCAN.

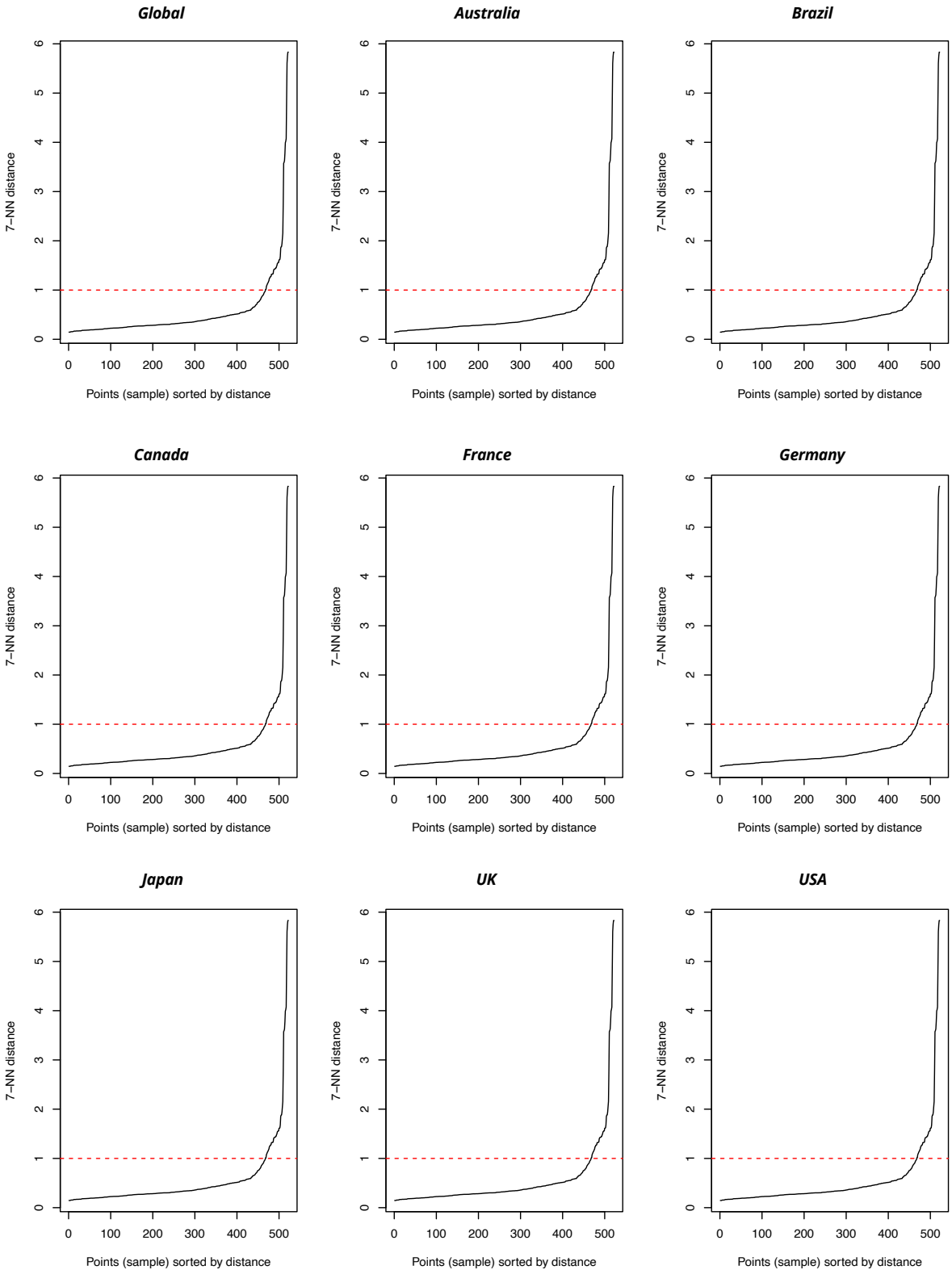


Figure B.8: 7-NN distance plot for each genre network in 2018. Dashed lines represent the threshold where a major change occurs in the curve, chosen as the ϵ parameter of DBSCAN.

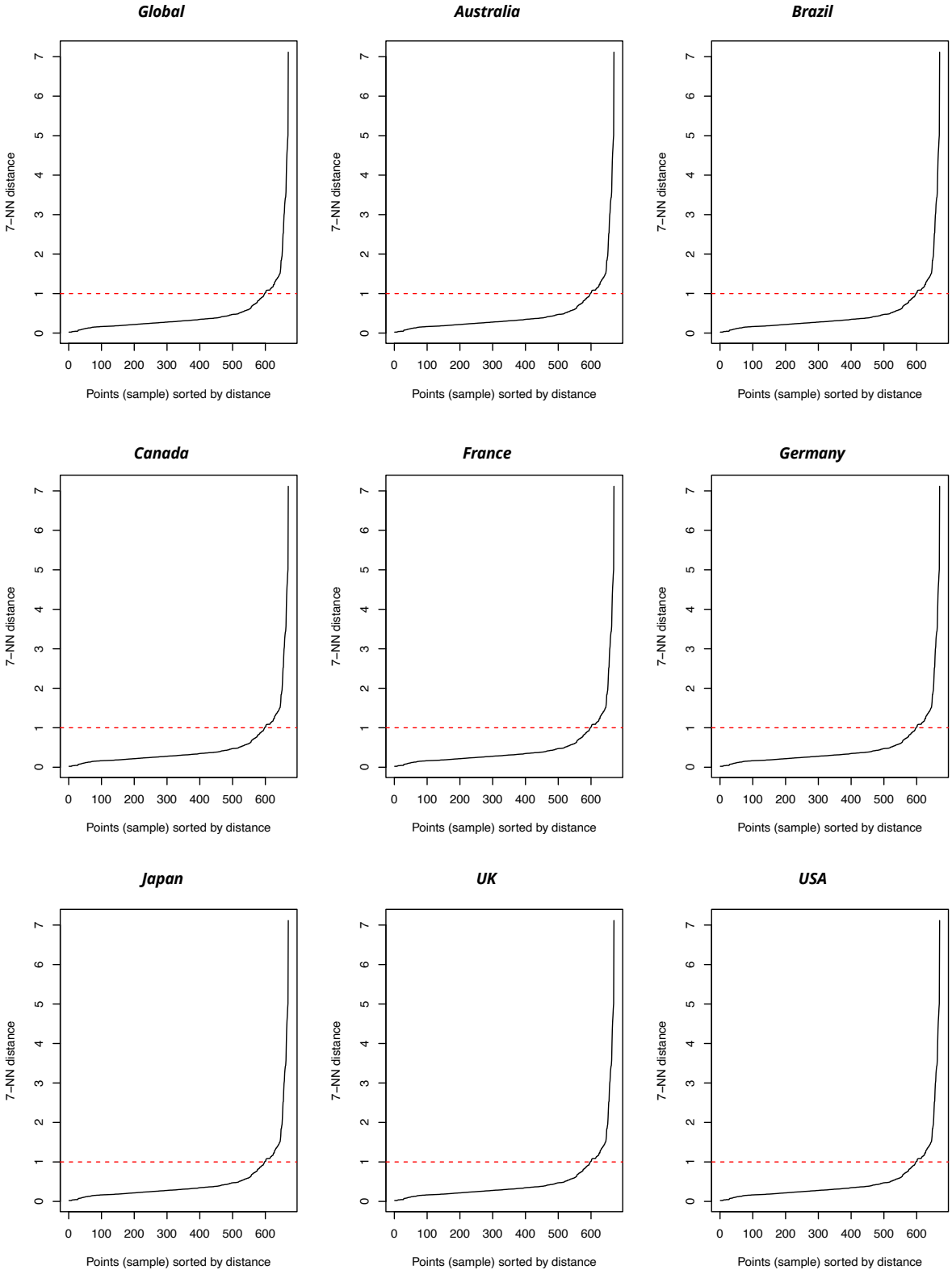


Figure B.9: 7-NN distance plot for each genre network in 2019. Dashed lines represent the threshold where a major change occurs in the curve, chosen as the ϵ parameter of DBSCAN.

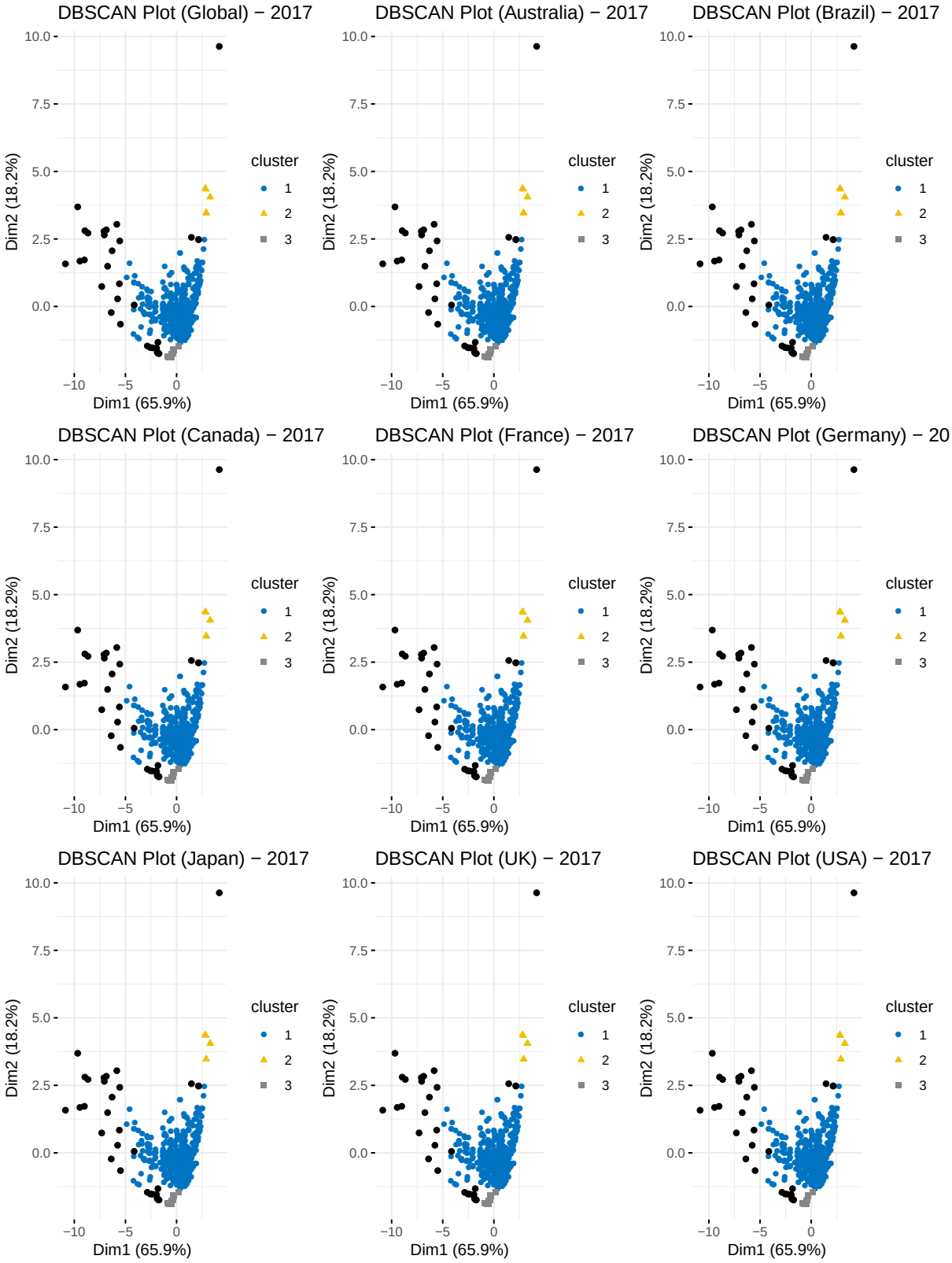


Figure B.10: Clustering of genre collaboration profiles in 2017. The results are generated with DBSCAN algorithm with $MinPts = 7$ and $\epsilon = 1.0$. The clustering is based on the topological metrics. Black points correspond to outliers.

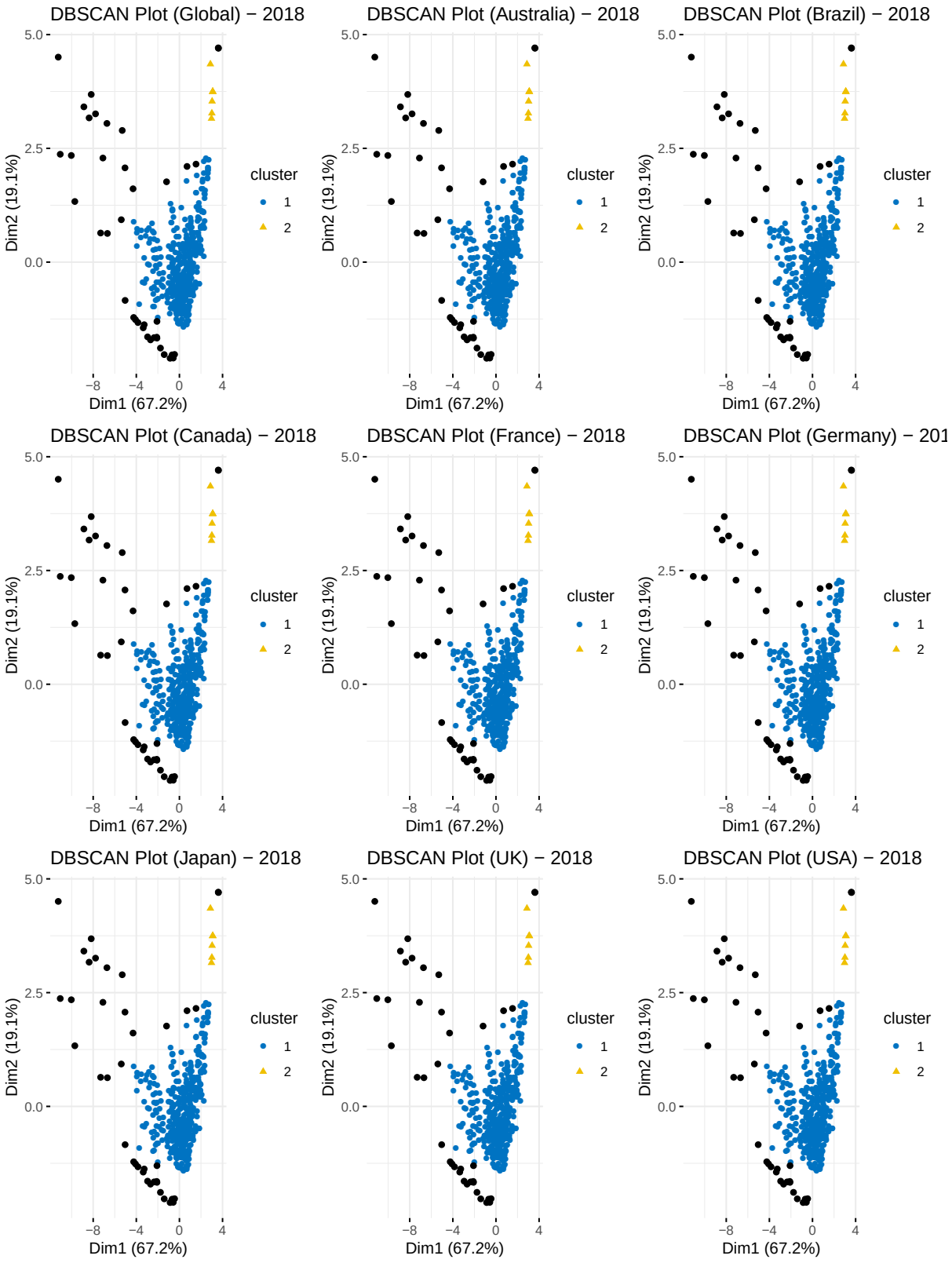


Figure B.11: Clustering of genre collaboration profiles in 2018. The results are generated with DBSCAN algorithm with $MinPts = 7$ and $\epsilon = 1.0$. The clustering is based on the topological metrics. Black points correspond to outliers.

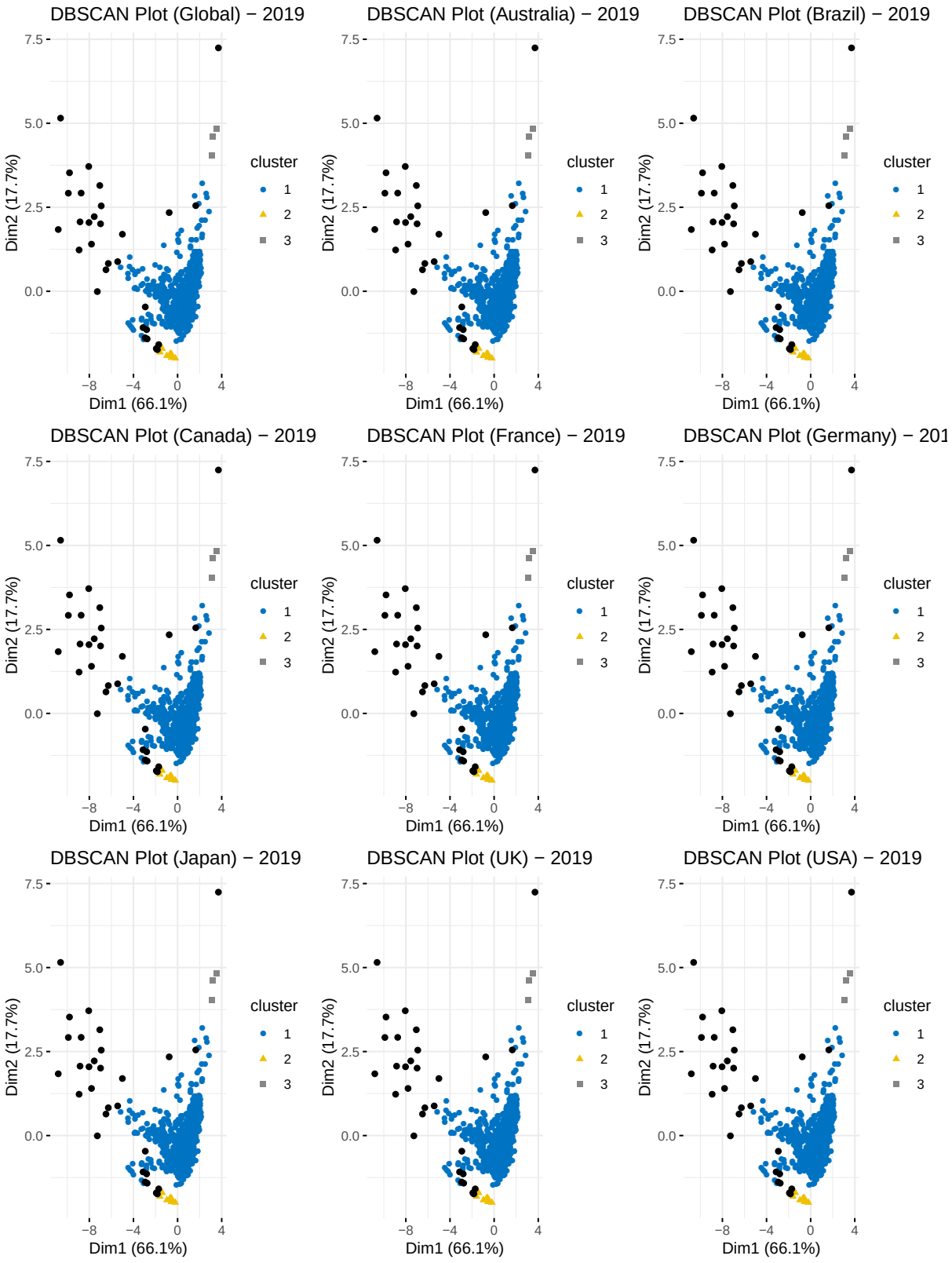


Figure B.12: Clustering of genre collaboration profiles in 2019. The results are generated with DBSCAN algorithm with $MinPts = 7$ and $\epsilon = 1.0$. The clustering is based on the topological metrics. Black points correspond to outliers.

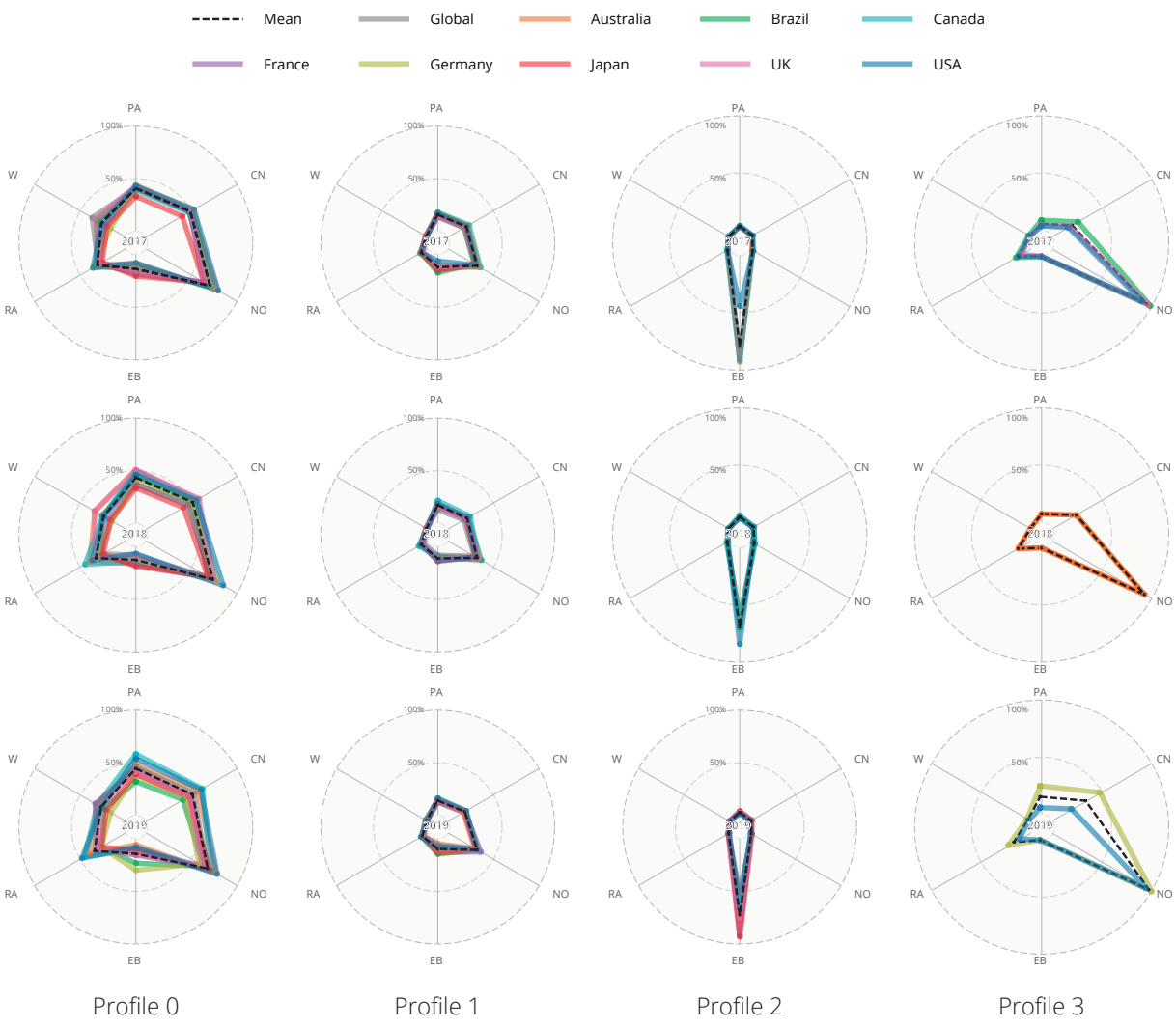


Figure B.13: Radar Plots of each genre collaboration profile, divided by year.

Appendix C

Genre Pattern and Association Rule Mining

Here, we present the details of the algorithm used to detect frequent genre patterns and the association rules in Chapter 6. As mentioned in such chapter, we use the Apriori algorithm [3], which is a classical method for Frequent Itemset Mining (FIM). We run the implementation of the *pyFIM* Python library by Christian Borgelt.[1] As frequent patterns and association rules are defined under the same data mining framework, we use the `apriori()` function for both tasks. Table C.1 presents the parameters used to get our results and their defined values. In the `target` parameter, we use *'c'* for closed itemsets and *'r'* for association rules. In addition, we define the output metrics in the `report` parameter, setting *'s'* for relative support and *'cl'* for a tuple containing the rule confidence and lift value.

Table C.1: Parameters of Apriori to get frequent genre itemsets and association rules.

Frequent Itemsets			Association Rules		
Parameter	**Definition**	**Value**	**Parameter**	**Definition**	**Value**
`target`	type of frequent itemsets	'c'	`target`	type of frequent itemsets	'r'
`supp`	minimum support	6	`conf`	minimum confidence	20
`zmin`	minimum number of items	2	`zmin`	minimum number of items	2
`report`	values to report	's'	`report`	values to report	'cl'

We also present the complete results obtained from such an algorithm per market and year. Therefore, Tables C.2, C.3, and C.4 contain the frequent genre patterns for each market in 2017, 2018, and 2019, respectively. Besides, the association rules for all markets in the same years are presented in Tables C.5, C.6, and C.7.

[1] PyFIM - Frequent Item Set Mining for Python: https://borgelt.net/pyfim.html

Table C.2: Top 5 most frequent patterns in global and regional markets (2017).

Market	Pattern	Support	Market	Pattern	Support	Market	Pattern	Support
	('dance pop', 'pop')	0.393		('dance pop', 'pop')	0.418		('dance pop', 'pop')	0.243
	('rap', 'hip hop')	0.250		('rap', 'hip hop')	0.232		('electro', 'pop')	0.180
Global	('pop rap', 'hip hop')	0.217	*Australia*	('tropical house', 'pop')	0.210	*Brazil*	('brazilian funk', 'pop')	0.138
	('rap', 'pop rap')	0.213		('electro', 'pop')	0.210		('tropical house', 'pop')	0.102
	('rap', 'pop rap', 'hip hop')	0.194		('electropop', 'pop')	0.202		('sertanejo', 'pop')	0.100
	('hip hop', 'rap')	0.397		('hip hop', 'pop')	0.538		('dance pop', 'pop')	0.311
	('rap', 'pop rap')	0.356		('rap', 'hip hop')	0.450		('rap', 'hip hop')	0.196
Canada	('hip hop', 'pop rap')	0.343	*France*	('rap', 'pop')	0.425	*Germany*	('tropical house', 'pop')	0.196
	('dance pop', 'pop')	0.332		('rap', 'hip hop', 'pop')	0.393		('electro', 'pop')	0.182
	('hip hop', 'rap', 'pop rap')	0.328		('francoton', 'pop')	0.251		('electro', 'tropical house')	0.174
	('dance pop', 'pop')	0.336		('dance pop', 'pop')	0.392		('hip hop', 'rap')	0.456
	('r&b', 'j-pop')	0.173		('rap', 'hip hop')	0.209		('pop rap', 'rap')	0.417
Japan	('tropical house', 'pop')	0.137	*UK*	('tropical house', 'pop')	0.201	*USA*	('hip hop', 'pop rap')	0.386
	('pop rap', 'pop')	0.134		('pop rap', 'pop')	0.193		('hip hop', 'pop rap', 'rap')	0.376
	('electro', 'pop')	0.130		('tropical house', 'dance pop')	0.182		('trap', 'rap')	0.341

Table C.3: Top 5 most frequent patterns in global and regional markets (2018).

Market	Pattern	Support	Market	Pattern	Support	Market	Pattern	Support
	('dance pop', 'pop')	0.295		('dance pop', 'pop')	0.331		('brazilian funk', 'pop')	0.188
	('rap', 'hip hop')	0.286		('hip hop', 'rap')	0.263		('sertanejo', 'brazilian funk')	0.110
Global	('pop rap', 'rap')	0.226	*Australia*	('pop rap', 'rap')	0.213	*Brazil*	('electro', 'pop')	0.106
	('trap', 'hip hop')	0.203		('pop rap', 'pop')	0.186		('dance pop', 'pop')	0.104
	('pop rap', 'hip hop')	0.199		('hip hop', 'pop rap')	0.183		('sertanejo', 'pop')	0.076
	('hip hop', 'rap')	0.418		('hip hop', 'pop')	0.578		('rap', 'hip hop')	0.244
	('pop rap', 'rap')	0.351		('rap', 'hip hop')	0.522		('dance pop', 'pop')	0.231
Canada	('pop rap', 'hip hop')	0.317	*France*	('rap', 'pop')	0.461	*Germany*	('tropical house', 'pop')	0.157
	('pop rap', 'hip hop', 'rap')	0.308		('rap', 'hip hop', 'pop')	0.428		('hip hop', 'pop')	0.148
	('trap', 'rap')	0.297		('francoton', 'pop')	0.240		('electro', 'pop')	0.143
	('dance pop', 'pop')	0.245		('dance pop', 'pop')	0.343		('hip hop', 'rap')	0.450
	('j-rock', 'j-pop')	0.191		('rap', 'hip hop')	0.262		('pop rap', 'rap')	0.375
Japan	('r&b', 'j-pop')	0.146	*UK*	('tropical house', 'pop')	0.183	*USA*	('trap', 'rap')	0.344
	('tropical house', 'pop')	0.128		('pop rap', 'rap')	0.175		('pop rap', 'hip hop')	0.328
	('tropical house', 'dance pop')	0.116		('pop rap', 'pop')	0.169		('pop rap', 'hip hop', 'rap')	0.321

Table C.4: Top 5 most frequent patterns in global and regional markets (2019).

Market	Pattern	Support	Market	Pattern	Support	Market	Pattern	Support
	('dance pop', 'pop')	0.271		('dance pop', 'pop')	0.294		('brazilian funk', 'pop')	0.177
	('latin', 'reggaeton')	0.173		('rap', 'hip hop')	0.162		('electro', 'brazilian funk')	0.102
Global	('hip hop', 'trap')	0.172	*Australia*	('electropop', 'pop')	0.145	*Brazil*	('sertanejo', 'brazilian funk')	0.097
	('rap', 'hip hop')	0.168		('rap', 'pop rap')	0.145		('electro', 'pop')	0.080
	('rap', 'trap')	0.151		('pop rap', 'hip hop')	0.131		('trap', 'hip hop')	0.064
	('hip hop', 'rap')	0.273		('hip hop', 'pop')	0.584		('dance pop', 'pop')	0.162
	('trap', 'rap')	0.255		('rap', 'hip hop')	0.449		('rap', 'hip hop')	0.158
Canada	('dance pop', 'pop')	0.253	*France*	('rap', 'pop')	0.423	*Germany*	('hip hop', 'pop')	0.130
	('pop rap', 'rap')	0.252		('rap', 'hip hop', 'pop')	0.393		('tropical house', 'pop')	0.105
	('hip hop', 'pop rap')	0.225		('francoton', 'pop')	0.174		('tropical house', 'dance pop')	0.087
	('j-rock', 'j-pop')	0.283		('dance pop', 'pop')	0.285		('hip hop', 'rap')	0.305
	('other', 'j-pop')	0.140		('rap', 'hip hop')	0.159		('trap', 'rap')	0.289
Japan	('anime', 'j-pop')	0.138	*UK*	('tropical house', 'pop')	0.133	*USA*	('pop rap', 'rap')	0.261
	('dance pop', 'pop')	0.133		('tropical house', 'dance pop')	0.127		('trap', 'hip hop')	0.246
	('r&b', 'j-pop')	0.108		('tropical house', 'dance pop', 'pop')	0.125		('pop rap', 'hip hop')	0.230

Table C.5: Association rules in global and regional markets sorted by lift value (2017).

Market	Rule	Lift	Confidence
Global	('tropical house', 'electro', 'dance pop') → house	7.720	0.532
	('tropical house', 'electro', 'dance pop', 'pop') → house	7.527	0.519
	('tropical house', 'electro') → house	7.504	0.517
	('tropical house', 'electro', 'pop') → house	7.319	0.505
	('tropical house', 'electro', 'dance pop') → electro house	7.253	0.560
Australia	('tropical house', 'electro', 'dance pop') → electro house	6.870	0.548
	('tropical house', 'electro', 'dance pop', 'pop') → electro house	6.759	0.539
	('tropical house', 'electro') → electro house	6.597	0.526
	('tropical house', 'electro', 'pop') → electro house	6.555	0.523
	('hip hop', 'dance pop') → urban contemporary	6.380	0.370
Brazil	('tropical house') → house	7.667	0.367
	('brazilian funk', 'pop') → pagode baiano	7.615	0.271
	('tropical house', 'pop') → house	7.519	0.360
	('rap', 'hip hop') → trap	7.278	0.597
	('rap') → trap	6.880	0.565
Canada	('electro', 'pop') → house	7.825	0.460
	('electro') → house	7.503	0.441
	('tropical house') → house	7.480	0.440
	('electro') → electro house	7.425	0.524
	('electro', 'pop') → electro house	7.421	0.524
France	('dance pop', 'pop') → electropop	5.754	0.283
	('dance pop') → electropop	5.721	0.282
	('dance pop', 'pop') → electro house	5.526	0.231
	('dance pop') → electro house	5.495	0.230
	('dance pop', 'pop') → tropical house	5.005	0.439
Germany	('electro', 'tropical house') → house	8.210	0.536
	('electro', 'tropical house') → electro house	8.137	0.543
	('electro', 'pop') → electro house	7.757	0.517
	('electro', 'pop') → house	7.717	0.503
	('electro') → electro house	7.267	0.485
Japan	('tropical house', 'dance pop') → house	9.907	0.465
	('electro', 'dance pop') → house	9.682	0.455
	('electro', 'dance pop', 'pop') → house	9.412	0.442
	('electro', 'tropical house') → house	9.261	0.435
	('electro', 'tropical house', 'pop') → house	9.198	0.432
UK	('electro', 'tropical house', 'dance pop') → electro house	8.321	0.517
	('electro', 'tropical house', 'dance pop', 'pop') → electro house	8.253	0.513
	('electro', 'tropical house') → electro house	7.979	0.496
	('electro', 'tropical house', 'pop') → electro house	7.976	0.496
	('electro', 'dance pop') → electro house	7.784	0.484
USA	('electropop', 'pop') → indie pop	8.232	0.291
	('electropop') → indie pop	7.981	0.282
	('r&b') → soul	6.696	0.254
	('r&b', 'pop') → soul	6.603	0.250
	('r&b', 'pop') → urban contemporary	6.282	0.331

Table C.6: Association rules in global and regional markets sorted by lift value (2018).

Market	Rule	Lift	Confidence
Global	('electro') → house	9.518	0.431
	('latin') → tropical	9.208	0.542
	('latin') → reggaeton	9.142	0.951
	('reggaeton') → latin	9.142	0.993
	('reggaeton', 'latin') → tropical	9.058	0.533
Australia	('electro', 'pop') → house	7.528	0.395
	('electro') → house	7.513	0.394
	('electro', 'pop') → dance	6.700	0.274
	('electro', 'pop') → tropical house	6.493	0.815
	('electro') → dance	6.367	0.261
Brazil	('electro', 'pop') → house	7.266	0.267
	('electro', 'pop') → tropical house	6.774	0.446
	('electro', 'pop') → electro house	6.757	0.347
	('electro') → house	6.213	0.229
	('hip hop') → trap	6.183	0.496
Canada	('dance pop') → electro	3.369	0.272
	('dance pop', 'pop') → electro	3.244	0.262
	('dance pop') → tropical house	3.094	0.292
	('dance pop', 'pop') → tropical house	3.022	0.285
	('pop', 'rap') → r&b	2.793	0.256
France	('rap', 'pop') → hip hop	1.305	0.912
	('hip hop') → rap	1.261	0.836
	('rap') → hip hop	1.261	0.881
	('hip hop', 'pop') → rap	1.241	0.824
	('hip hop', 'pop') → francoton	1.236	0.318
Germany	('electro') → house	9.286	0.450
	('electro') → dance	8.556	0.350
	('tropical house') → house	8.208	0.398
	('tropical house', 'pop') → house	8.107	0.393
	('tropical house', 'pop') → dance	8.003	0.327
Japan	('electro', 'tropical house', 'dance pop') → house	8.567	0.435
	('electro', 'tropical house', 'dance pop', 'pop') → dance	8.566	0.282
	('electro', 'tropical house') → house	8.437	0.429
	('electro', 'dance pop') → house	8.336	0.423
	('electro', 'dance pop', 'pop') → dance	8.324	0.274
UK	('electro', 'tropical house') → dance	8.456	0.338
	('electro') → dance	7.931	0.317
	('electro', 'tropical house') → house	7.864	0.449
	('electro', 'tropical house') → electro house	7.603	0.265
	('electro') → house	7.497	0.428
USA	('dance pop') → electro	3.995	0.226
	('dance pop') → tropical house	3.770	0.246
	('dance pop', 'pop') → electro	3.754	0.212
	('dance pop', 'pop') → tropical house	3.624	0.236
	('pop', 'hip hop') → r&b	2.960	0.276

Table C.7: Association rules in global and regional markets sorted by lift value (2019).

Market	Rule	Lift	Confidence
Global	('latin', 'reggaeton') → tropical	7.922	0.468
	('latin') → tropical	7.821	0.462
	('reggaeton') → tropical	7.722	0.456
	('reggaeton') → latin	7.623	0.975
	('latin') → reggaeton	7.623	0.987
Australia	('tropical house') → house	7.655	0.342
	('tropical house', 'pop') → house	7.173	0.321
	('tropical house', 'pop') → electro	7.111	0.670
	('tropical house') → electro	7.077	0.667
	('tropical house') → electro house	7.041	0.261
Brazil	('hip hop') → trap	6.187	0.434
	('brazilian funk', 'pop') → pagode baiano	5.473	0.425
	('hip hop') → pop rap	5.235	0.303
	('hip hop') → r&b	4.443	0.263
	('hip hop') → rap	4.034	0.303
Canada	('r&b') → soul	7.485	0.226
	('dance pop') → tropical house	3.214	0.243
	('dance pop', 'pop') → tropical house	3.160	0.239
	('rap', 'pop') → pop rap	2.677	0.880
	('pop rap', 'hip hop', 'rap') → trap	2.638	0.777
France	('rap', 'pop') → hip hop	1.301	0.900
	('rap', 'pop') → francoton	1.263	0.325
	('hip hop', 'pop') → rap	1.259	0.796
	('hip hop') → rap	1.234	0.779
	('rap') → hip hop	1.234	0.853
Germany	('dance pop') → tropical house	5.909	0.400
	('dance pop') → electro	5.908	0.338
	('dance pop', 'pop') → tropical house	5.824	0.394
	('dance pop', 'pop') → electro	5.796	0.332
	('trap') → pop rap	5.002	0.422
Japan	('r&b') → j-rap	8.067	0.228
	('dance pop') → electro	4.348	0.283
	('dance pop', 'pop') → electro	4.284	0.279
	('dance pop') → tropical house	4.273	0.353
	('dance pop', 'pop') → tropical house	4.227	0.349
UK	('rock') → indie rock	8.370	0.364
	('rock') → indie	6.216	0.231
	('pop rap', 'hip hop') → trap	5.682	0.660
	('grime', 'hip hop') → dancehall	5.642	0.209
	('pop rap', 'rap') → trap	5.595	0.650
USA	('pop rap', 'pop', 'rap') → r&b	2.990	0.291
	('pop', 'rap') → r&b	2.888	0.281
	('hip hop', 'pop') → r&b	2.878	0.280
	('pop rap', 'pop') → r&b	2.618	0.255
	('pop', 'rap') → pop rap	2.527	0.904