DOCTORAL DISSERTATION

Successfully defended on the 19th of December 2024

Modeling and Influencing Music Preferences on Streaming Platforms

KRISTINA MATROSOVA

Supervisors:

Manuel Moussallam
Director of Research, Deezer

DEEZER

Thomas Louail
Tenured researcher, CNRS Géographie-Cités

cnrs · GÉOGRAPHIE-CITÉS

Olivier Bodini
Professor, Université Sorbonne Paris Nord

UNIVERSITÉ SORBONNE PARIS NORD

Jury:

Geneviève Vidal (president)
Professor, Université Sorbonne Paris Nord

Christine Bauer (manuscript reviewer)
Professor, University of Salzburg

Florence Levé (manuscript reviewer)
Professor, Université de Picardie Jules Verne

Cynthia C. S. Liem
Associate Professor, Delft University of Technology

"The internet's completely over."

PRINCE (2010)

Preface

This thesis was submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science. It was conducted jointly between CNRS UMR Géographie-cités under the supervision of Thomas Louail, Université Sorbonne Paris Nord under the supervision of Olivier Bodini, and Deezer under the supervision of Manuel Moussallam. This work is part of the 'RECORDS' project, and was funded by the French National Agency of Research (ANR-2019-CE38-0013).

Abstract

Today, music streaming services have become the primary means for people to access and enjoy recorded music. These platforms rely on algorithmic recommendation systems to help users navigate vast music catalogs and provide a personalized listening experience.

For these recommendations to be accurate, patterns must be identified, and user preferences modeled accordingly. There are numerous ways to approach this modeling process, and each method impacts the recommendation outcomes differently. In turn, by consuming the recommended items, users' preferences and listening behaviors may be influenced in various ways. For instance, recommendations may lead to a diversification of their music choices, exposing them to new content, or, conversely, confine them to a niche. These influences may also manifest in more subtle ways, making them harder to measure.

This thesis explores the feedback loop between musical taste and recommendation systems through two main questions: how can we computationally model individual musical preferences using behavioral data from streaming platforms? how can we measure the influence that recommender systems have on shaping those preferences?

We begin by providing an overview of the data that can be collected from streaming platforms, exploring the types of information available about users, music items, and their interactions. We also examine the structure and distribution of these data, which form the foundation for subsequent analysis.

Next, we review the literature on musical taste, a topic that has been widely studied across various fields such as sociology, psychology, and cultural geography, long before the advent of music recommender systems. By reviewing these diverse approaches to measuring musical preferences, we identify insights that can inform and improve recommendation algorithms.

We then address the challenges of representing the musical space. With millions of tracks available on streaming platforms, it is crucial to categorize, label, and identify common patterns among music items in order to accurately model user preferences and generate relevant recommendations.

Finally, we present an overview of music recommendation systems, focusing on the current methods, specific challenges in the music domain, and the

4

role of fairness and bias in influencing user preferences and listening habits.

With this multidisciplinary and multi-modal background in place, we address our research questions. First, we explore how to model musical taste using streaming data, constructing a 'fingerprint' of user preferences based on two different definitions of musical taste: one that highlights individual uniqueness and another that reflects collective preferences. Then, we tackle the challenges in measuring the influence of music recommender systems, using local music consumption as a case study. Ultimately, this thesis aims to bridge the gap between understanding musical taste and developing recommendation systems that are not only personalized but also capable of fostering diverse and satisfying listening experiences.

Résumé

Aujourd'hui, les services de streaming musical sont devenus le principal moyen d'accéder à la musique enregistrée. Ces plateformes reposent sur des systèmes de recommandation algorithmique pour aider les utilisateurs à naviguer dans des catalogues immenses et leur offrir une expérience d'écoute personnalisée.

Pour que ces recommandations soient pertinentes, il faut d'abord identifier les schémas de comportement et modéliser les préférences des utilisateurs. Il existe plusieurs manières de procéder à cette modélisation, chaque méthode peut influencer les recommandations de manière différente. Réciproquement, les recommandations peuvent elles-mêmes affecter les goûts et les pratiques d'écoute des utilisateurs, que ce soit en élargissant leurs horizons musicaux ou, à l'inverse, en les enfermant dans une bulle. Parfois, ces effets sont plus subtils, ce qui les rend plus difficiles à mesurer.

Cette thèse examine la relation entre le goût musical et les systèmes de recommandation à travers deux questions: comment modéliser les préférences musicales individuelles à partir des données comportementales issues des plateformes de streaming? comment mesurer l'impact des systèmes de recommandation sur la formation de ces préférences?

Nous commençons par une présentation des données disponibles sur les plateformes de streaming : les informations relatives aux utilisateurs aux morceaux de musique et à leurs interactions. Nous examinons également la structure et la distribution de ces données sur lesquelles reposent nos analyses.

Ensuite, nous passons en revue la littérature sur le goût musical, un sujet largement exploré dans des domaines comme la sociologie, la psychologie et la géographie culturelle, bien avant l'arrivée des systèmes de recommandation. Que peut-on tirer des approches utilisées dans ces domaines pour mieux comprendre et modéliser les préférences musicales, notamment dans le cadre de la recommandation algorithmique?

Nous abordons ensuite les défis liés à la représentation de l'espace musical. Avec des millions de morceaux disponibles, il est essentiel de catégoriser, étiqueter et identifier les similarités entre les titres pour modéliser efficacement les préférences et fournir des recommandations pertinentes.

Enfin, nous proposons une vue d'ensemble des systèmes de recomman-

dation musicale, en présentant les méthodes actuelles et les défis propres à la recommandation de musique, notamment les notions d'équité et de biais algorithmiques qui peuvent influencer les préférences et les pratiques d'écoute des utilisateurs.

Avec ce cadre pluridisciplinaire et multimodal en place, nous tentons de répondre à nos questions de recherche. Nous commençons par explorer comment modéliser le goût musical à partir des données de streaming, en construisant une 'empreinte' des préférences basée sur deux conceptions du goût musical : l'une axée sur l'individualité l'autre reflétant des tendances collectives. Ensuite, nous nous attaquons aux défis liés à la mesure de l'influence des systèmes de recommandation, en prenant pour exemple la consommation de musique locale. Au final, cette thèse vise à lier la compréhension du goût musical et la création de systèmes de recommandation capables non seulement de personnaliser les suggestions, mais aussi de permettre une écoute diversifiée et enrichissante.

Acknowledgements

During these years, I was surrounded by many people who, in some way or another, have helped me get through this journey, and I would like to express my gratitude to all of them.

First, I would like to thank my supervisor Thomas Louail for creating the RECORDS project, of which this PhD project was part. I really enjoyed this unique experience of working with people from different scientific fields, both academics and people from the industry, and I have learned a lot from each of the interactions we had together. Thanks for introducing me to "social science" literature and way of thinking. Thanks for teaching me how to make nice plots. Also, thanks for all those long speeches you gave us once in a while at the end of a meeting. Often I lost track of the initial thought, and probably so did you, but it was still always fun and insightful. This PhD journey was a first experience for both of us: for me as a student, and for you as a supervisor — something I fully realized only after my defense. You put in a lot of effort, and it showed. I sincerely hope I was just the first of many PhD students you will guide in the future.

I also thank Manuel Moussallam, my other supervisor. You guided and supported me from the beginning to the very end of this journey, both in terms of scientific work and simply by being there on a human level. Thanks for always having great ideas, and for helping with structuring my own. Thanks for your kindness — in moments of panic and despair, your calming words made all the difference (probably a skill you picked up from raising your kids, but it works surprisingly well on anxious PhD students too). Thank you for always making time to help me, even if it means spending the night before a deadline restructuring a paper (that we did not even submit in the end). Also thank you for being such a rock star during Deezer parties!

Last but not least, I would like to thank my director Olivier Bodini. Thanks for jumping on board with me on this project without really knowing what to expect. You did great, especially considering that it is not the kind of research topic and workflow you are used to. Even though Ramshaw and Tarjan (2012)'s 95 pages long paper you made me read will probably haunt me forever, it was a small price to pay for this great adventure together. Thank you for always being honest with me, and saying out loud things that needed to be said, yet always in

8

a kind and caring way. Also, thank you for your help with all the administrative stuff, I can only imagine how much time I saved being able to get the 3-in-1 signature of doctoral school director, team director and PhD supervisor at once from one person.

To the three of you - thank you for your trust, and for being great supervisors, both on the technical side and simply as human beings. It took us some time to figure out how to make this quartet work, but in the end I think we did pretty well.

A big thank you to the jury members for taking the time to engage with my research. Christine Bauer and Florence Levé, thank you for reviewing my manuscript. Cynthia Liem and Geneviève Vidal, thank you for being part of this once-in-a-lifetime day. Writing this after my defense, I am grateful for your questions, which lead to insightful interactions and new research ideas, and of course for granting me the title of PhD. A special thanks to Christine Bauer for representing the jury until the very end of the celebration night!

As I spent most of my PhD journey with Deezer, particularly the Research team, I would like to thank all of them for their warm welcome. Thanks to Marion and Bruno for supporting me during the recruitment process for this thesis. Thanks to Felix for recruiting me for the internship prior to the PhD, otherwise I would probably never have ran into the doctoral job offer. Also, thank you for teaching me how to make my code shiny clean. And for simply being such a great friend. Thanks to Darius - we spent our PhD journeys side by side, and it was a rollercoaster of emotions, but there was no problem that could not be solved after a good tea-and-netflix-gossip break. Thanks to Pierre for being supportive even though you probably never really knew what I was doing exactly. Thanks to Karim for all the nice chats. Thanks to Gabi for the randomness. Anti-thanks for the birthday cake. Overall thanks to Felix & the PhDs for the groove — you made this whole experience a lot more fun.

Thanks to everyone I met and spent time with during seminars and conferences, and a special shoutout to the SysMus22 and RecSys24 teams for all the fun memories we created along the way.

I would also like to thank all of my lifelong friends who have been by my side through many different phases of life, and to the wonderful friends I have met more recently outside of work - Sasha, Masha, Nicolas, the ULTRA team, Georgy, Polina, Nastia, Serguey, Ségolène and many others. Thank you for the laughter, the conversations, and the much-needed distractions. Whether it was sharing drinks, holidays, or spontaneous climbing sessions, you reminded me of life beyond my research. Your presence, even in the simplest moments, gave me energy and perspective throughout this journey. I am lucky to have you all in my life.

Speaking of climbing, thanks to my climbing gym staff for fueling me with coffee and cheering me up while I was writing my manuscript in their

co-working space. Also thanks to the route-setters for creating cool problems on which I could ~~procrastinate~~ unwind when the thesis was not writing itself.

Of course, I would like to express infinite gratitude to my parents, who always put my education first, often at the expense of their own comfort. I would never have been able to come this far without your help, thank you for that. Special thanks to my mom who not only provided me with financial support, but also made me fluent in two foreign languages and was my personal teacher in pretty much all of the school subjects - opening countless doors for me and giving me opportunities I would not have had otherwise.

Finally, I am grateful to my life partner, Amine. You are the one who fully shared this experience with me, every step of the way. Thank you for standing by my side in both moments of success and despair. Thank you for enduring my countless "who needs a PhD anyway?" moments with patience and understanding. Thank you for believing in me even when I couldn't see it myself. Thank you for supporting me from the moment I applied my resume to the day of the defense. Thank you for taking care of me in all possible ways, so that I could fully focus on my work, without worrying about anything else, and without even having to ask. Beyond all that, thank you for simply being the wonderful and beautiful human that you are. It is a true blessing to have you in my life.

To conclude, I would like to cite the one and only Snoop Doggy Dogg: *"Last but not least, I wanna thank me. I wanna thank me for believing in me. I wanna thank me for doing all this hard work. [...] I wanna thank me for never quitting"*.

Contents

Acronyms

ALS Alternating Least Squares. 106, 117, 118

AUC Area Under the Curve. 80, 82

BFI Big Five Inventory. 48, 51, 52, 73

CBF Content-Based Filtering. 102–104, 106, 107, 115

CF Collaborative Filtering. 35, 76, 98, 102, 104–107, 109, 115–120, 125, 142

CNN Convolutional Neural Network. 82, 84, 110

DCBM Degree-Corrected Block Model. 93, 94, 142

GMM Gaussian Mixture Models. 81, 83

k-NN k-nearest neighbors. 81, 82

MF Matrix Factorization. 35, 91, 93, 104–107, 109, 115–117, 119, 120, 142

MIDI Musical Instrument Digital Interface. 81, 82

MIR Music Information Retrieval. 73, 76, 81, 98

MRMR Minimum Redundancy Maximum Relevance. 82

MRS Music Recommender Systems. 18, 19, 21, 41, 70, 71, 74, 96, 100–102, 109, 112–117, 121, 154

MUSIC Mellow, Unpretentious, Sophisticated, Intense, and Contemporary (Rentfrow et al.'s five-factor model). 50, 51, 87, 88

NLP Natural Language Processing. 53, 85

PCA Principal Component Analysis. 67, 74, 82, 93

RS Recommender Systems. 18, 19, 21, 33, 36, 69–71, 76, 91, 99–102, 105–111, 114, 115, 118, 123, 125, 143, 155, 158, 159

SBM Stochastic Block Model. 93, 142, 143

SVD Singular Value Decomposition. 50, 91–95, 105, 106

SVM Support Vector Machine. 82

Chapter 1

Introduction

1.1 General overview

1.1.1 Context

My parents have never been huge music fans. Sure, my mom had a couple of songs that she would hum while doing chores, and my dad always repeated Opus' 'Life is Life' chorus whenever he was in a good mood, but it was pretty much it. As I could not really get any music influence from my family, I had to get it from somewhere else.

Around the age of 5, I was introduced to classical music by my piano teacher, and even through I liked to play the piano, I do not remember being particularly involved with the music itself. Around 10, I started watching MTV charts. I would write down the songs' titles in a notebook in order to download them from eMule[1] later — this was a full on research task as I would do this with any song, no matter if I liked it or not. I don't remember why I started doing this exactly, but I guess that I noticed that music became a subject among kids of my age, and I probably felt the need to create my own music identity. During this process, I bonded with several artists, but the main one was a Russian 'rock' girls band[2], which was very popular at the time.

Starting from there, a clear path was set: I will be a rock fan. I asked for a guitar for my birthday, made friends with other kids who liked rock in my class, and we even created a band of our own. The more I grew up, the more reckless of a teenager I was becoming, the more 'heavy' was the music I listened to. By the age of 17, I was a fully set 'metalhead'.

After high-school, I moved to France to study in university, and slowly started to shift from this label, digging more and more into different kinds of music. For example, I started listening to russian pop, which I never really en-

---

[1]eMule

[2]the best music band that has ever existed (according to my 12 years old self)

joyed before, out of nostalgia. Also, I met a lot of new friends, who definitely influenced my preferences and introduced me to new genres and artists. Finally, I gave a second chance to some music that I did not understand as a teenager, like psychedelic rock or jazz, and some of it suddenly became my favorite.

We all have our own history with music, shaped by different influences and experiences. But can we make generalizations about why we like the music we do? Are there patterns common to all or most people? These are classic questions, and various fields of knowledge have attempted to address them. When I began this thesis, I was driven by my personal experience with music and a desire to understand other people's musical journeys. I was curious about whether there are shared paths in how we develop our musical tastes and how these preferences evolve over time.

The ANR RECORDS project, of which this thesis is a part, was specifically launched to explore these questions in the streaming era. Beyond personal feelings and experiences, streaming data allows us to analyze billions of digital traces from listeners across the globe. This, in theory, provides an almost exhaustive view of music consumption on a massive scale.

With this unprecedented data, new opportunities arise to explore whether computational methods can help us answer questions such as: Are there identifiable pathways in how individuals develop their musical tastes? Can we detect trends that transcend personal histories and social contexts? What role do algorithms and platforms play in shaping these journeys, and can we measure this influence? These are the questions that this thesis seeks to address.

1.1.2 At the origins of musical taste

Where does our musical taste come from?

A common assumption from psychology is that our music preferences mostly form around our teenage years, and do not change that much after early 20s (North and Hargreaves, 1995). The evolution of taste with aging has generally been a big subject among psychologists, and for example it seems that our tolerance for novelty increases as we get older(Berlyne, 1973). Also, some correlations were found between personality and preferred music styles, for example extroverted people seem to usually enjoy energetic music, like hip-hop or funk (Rentfrow and Gosling, 2003). Does all of this mean that I was condemned to like rock and metal as a dreamy and dissident teenager? And was it only natural that I approached more and more 'sophisticated' music while growing up?

Beyond psychological factors, the outside world also has an influence on how our music preferences are shaped. Many different theories have been explored by social scientists over the years. Starting from the 60s, sociologists established correlations between individuals' social background — education, economic, cultural capitals — and their listening and appreciation of specific

genres, artists or music pieces (Bourdieu, 1984). Later on, the omnivore/univore theory emerged, linking a broader and more diverse music preferences to higher social status, and vice versa (Peterson, 1992). More recent work suggests that as we get to be around people from different backgrounds during our lives, this unique set of connections helps us to build our own individual taste (Lahire, 2008). Conducting a proper review of this abundant scientific literature is obviously beyond the scope of this work, but later in this manuscript we will discuss some of these past works.

Our music taste can fluctuate a lot all over our life course, influenced by many things other than our social life. As we use music to regulate our emotions (Saarikallio, 2008), our preferences can evolve according to our mood, activity, minor and major life events. For example, listening to music from our homeland out of nostalgia seems to be quite a typical human behavior (Barrett et al., 2010).

1.1.3 The evolution of available data

'Traditional' data collection methods

As we will see in the following of this manuscript, many scientific studies that aim to better understand human preferences, including when it comes to music, have relied on declarative data collected through surveys and interviews. In some cases, observational data has been used, but if was often recorded in somewhat unrealistic conditions of practice — arbitrary experiments conducted in a lab, with people very involved in the activity being studied. This raises concerns about the representativity of the results obtained from such data.

The first issue is scalability and sample representativeness: it is costly and requires substantial resources, notably to find enough people willing to give their time to participate in a study, and it is even more challenging to ensure that those respondents are representative of a population in general. For example, many psychology studies have been conducted on students, because they are readily available, relatively free, and can be easily motivated to give some of their time in exchange for credits or a small fee. However, such studies may be hard to generalize because they often rely on a small and peculiar demographic group.

Second, the data collection process involves decisions that cannot be revised later. Typically, researchers will decide on a set of questions in advance and chose specific metrics for each, for example represent the preferred music as a Likert scale for a predefined list of music genres. Once data collection is completed, it is usually not possible to collect more data or alter the representation of the data, which can be quite inconvenient, especially considering that these decisions can directly impact a study's results.

Finally, the problem with declarative, self-reported data is that humans are not the most reliable reporters. We all have our biases, we may forget, under or overestimate things, and have a distorted perception of reality influenced by

social norms and personal subjective experiences. In addition, for such a frequent activity like recorded music listening, very few people would practice some kind of 'quantified self-tracking' of when, what, how much they listen to, etc., thereby producing data that might prove useful for the scientific study of this social practice. All of these factors can compromise the accuracy of study results.

The emergence of streaming data

Today, with the widespread use of the Internet and streaming services, the amount of available data has vastly increased. We constantly share detailed information with different companies — our locations, who we are with, what we post, like, comment, and even how long we look at specific ads — all of our actions are carefully stocked in databases counting billions of lines. Music streaming is no exception, and services like Deezer[3], Spotify[4] or Apple Music[5] do their best to keep track of their users' histories of interactions on their platforms. These records, called logs, can typically include the type of action (e.g., stream, like, click, skip, search query), its timestamp, location, duration, context and other information. Considering that these platforms count millions of users, who interact with millions of songs everyday, a tremendous volume of data has been collected over the years.

This data, which accurately reflects real-life music listening behaviors, is more reliable than data gathered from surveys where people report their own behaviors, which can be biased. In an ideal world, this wealth of data could answer many longstanding questions researchers have about music listening habits.

While big data offers significant advantages, it also introduces new challenges. One major issue for research, particularly public-sector research, is accessing this data, as it is primarily collected by private companies that offer these digital services, and they are the ones who decide what data to share and with whom. Fortunately, collaborations efforts like the ANR 'RECORDS' project are more and more frequent, showing promise in bridging the gap between private data holders and public researchers. Still, sharing data with third parties and making open-source datasets must be done with extra caution to ensure users' privacy. A whole branch of research have developed to address this issue, exploring techniques like hashing, partially deleting or adding noise to data, to protect individual privacy while preserving the integrity and usability of the data for research purposes.

Furthermore, while a large volume of data may initially seem advantageous, it also introduces significant computational challenges. Methods and algorithms

---

[3]deezer.com
[4]spotify.com
[5]music.apple.com

that work with smaller datasets can become impossible to use on much larger datasets generated by streaming services.

Finally, as opposed to declarative data, where a respondent answers directly to specific questions, behavioural data needs to be interpreted. Typically, a single user action can have multiple explanations. For example, a user can skip a song because they dislike it, but also because it does not suit their current mood, even though they generally enjoy it. Similarly, the 'like' button, typically seen as an indicator of preference, might also be used to bookmark tracks that a user has not actually listened to yet. In more advanced contexts, researchers might want to get some additional information about the users through indirect indices: for example, analyzing IP addresses to detect user relocation, or using device type as an indicator of social class. This is a problem that usually does not occur with declarative data, where the participants give explicit responses to specific questions.

1.1.4   Recommender systems and musical taste

Streaming platforms have introduced more than just new tools for researchers studying music listening behavior. Their emergence has intertwined these theoretical questions with practical challenges for the platforms themselves.

With millions of tracks available, streaming services rely on recommender systems (Recommender Systems (RS)) (Lü et al., 2012) to help users navigate these vast catalogs of music. These systems analyze past user behavior and preferences to suggest music that listeners might enjoy, typically by finding patterns based on similarities between users or music items. To build effective RS, platforms need to model users' preferences and detect patterns in their listening behaviors. The more accurately these systems understand and cater to user tastes, the more likely users are to engage with the recommendations.

For a long time, the development of music recommender systems (Music Recommender Systems (MRS)) evolved separately from research on musical taste in other disciplines, and only occasional attempts have been made to bridge the two fields. For example, Laplante (2014) conducted an extensive review of studies on musical taste from the social sciences, focusing on how these insights could enhance the design of RS. Soleymani et al. (2015) proposed and demonstrated the success of combining content-based approaches with emotional and psychological aspects of music perception to improve the recommendation of less popular songs. Zangerle et al. (2018, 2020) explored how incorporating users' country-specific, culture-related, and socio-economic features could improve recommendation quality. Separating 'theoretical' research on musical taste from work on MRS is problematic because understanding and modeling users' music preferences is essential for effectively feeding these systems. We hope this thesis contributes to the growing interdisciplinary literature in this

area.

In addition to understanding how to model music preferences for building high-quality MRS, it is important to consider the impact these systems may have on how people consume music. Just as friends or favorite radio stations shape our music tastes, the growing reliance on algorithmic recommendations for discovering new music means that these algorithms can influence our preferences as well (Datta et al., 2018; Anderson et al., 2020). Algorithmic fairness has become a significant concern within the RS community (Wang et al., 2023).

In the music context, algorithms seem to favor more popular artists, which can disproportionately include male artists and those from more developed countries (Kowald et al., 2020; Shakespeare et al., 2020; Lesota et al., 2022). Additionally, these systems tend to perform better for users with more mainstream tastes, potentially marginalizing those with niche preferences Kowald et al. (2020). However, since these systems frequently lack explainability (the ability to provide clear post-hoc explanations for their decisions) and interpretability (the ease with which humans can directly understand how a model works) it is not always easy to identify the source of certain biases Afchar (2023). Moreover, some biases may be difficult to measure due to incomplete, incorrect, or initially biased datasets. We will explore these challenges throughout this thesis.

Music preferences and the way we choose to shape them in a recommendation setup can impact the outcome of recommendations, which might influence our preferences in return. In this thesis, we will explore how streaming MRS both respond to and influence user behavior.

1.2   Research objectives

In this thesis, we will focus specifically on two central questions :

How can we model individual musical taste, both to better understand it and also to use it for recommendation purposes? Streaming platforms provide an overwhelming amount of data related to user preferences, and whether for understanding these preferences or using them in RS, it is crucial to compress this information to extract its essence. Drawing from different definitions in the humanities, we will focus on two in particular: what makes us unique and what is representative of our general preferences. We will then attempt to apply these definitions to streaming data in the form of a concise fingerprint for each user.

How can we measure the influence that RS may have on musical taste? We will begin by discussing the fairness of MRS in general and then address this question through the specific example of local music bias. We will explore population biases that may exist in different datasets, as well as the issues related to the labeling of musical items, both of which can distort our understanding

of algorithmic biases. Finally, we will discuss ways to mitigate these issues in order to create a more accurate picture of existing biases and address them accordingly.

Since these questions are embedded in a broad societal and technological context, we will draw on literature from multiple disciplines to explore them, which is what makes this thesis unique.

1.3 Thesis structure

To begin, we need to become familiar with the data generated by streaming platforms. In Chapter 2, we will explore what this data looks like using Deezer as an example. We will examine the available information about users, the structure of the music catalog and its metadata, and focus on the interactions between users and the items within the catalog. We will discuss how these interactions can be represented, the statistical distributions behind/sustaining these interactions, and what additional insights, such as geographical location or context, can be derived. We will then look at the specific characteristics of streaming data and how it differs from other types of interaction data collected online. Next, we will discuss how behavioral data from streaming can be supplemented with declarative information collected in surveys or interviews. Finally, we will address the importance of making this type of datasets publicly available, outlining the challenges related to data anonymization and how they can be addressed.

Following this, we will move on to three literature review chapters. As mentioned earlier, the topics covered in this thesis are multifaceted and have been studied across different disciplines, so it is important to approach them from multiple angles.

In Chapter 3, we will examine how musical taste has been studied and measured across various fields, including sociology, psychology, ethnomusicology, and data science. We will explore the primary origins of our musical preferences, such as social background, personality traits, and sociological or cultural origins, as well as how our tastes evolve over the course of our lives, both in the short and long term. The goal of this chapter is to understand the different ways we can quantify human, or user, behavior.

We will note throughout this chapter that there are many ways to represent music, and the way we do so can influence how we perceive people's tastes. In Chapter 4, we will shift focus to the music catalog itself, exploring how we can extract information from different musical objects, label them, categorize them, or represent them in a continuous space. We will approach this from two perspectives: first, by using the characteristics of the music itself, either through human experts or automated audio analysis, and second, by relying on aggregated human behaviors and preferences.

Once we have established how to represent both users and music items and how they are interconnected, we will proceed to Chapter 5, where we will discuss how these interactions can be initiated and influenced by RS. We will first provide an overview of the different types of recommendation algorithms. Then, we will discuss the main challenges in RS, both generally and in the specific context of music. Key issues such as the cold start problem, contextual recommendation, balancing novelty with familiarity, and biases and fairness will be covered.

With these foundations in place, we will attempt to answer our two main research questions through two studies that I have conducted during my PhD.

In the first study, we will explore how to model users' musical taste and capture it in a fingerprint, from their behavior on a streaming platform. To construct this fingerprint, we will rely on two definitions of musical taste from the literature — one defining taste as what makes us unique, and the other as what is representative of our overall preferences — and we will show how these two definitions lead to contradictory solutions. Additionally, we will address the users' identifiability through their interactions with music.

In the second study, we will tackle the issue of fairness in MRS, using the specific example of local music consumption. Building on a previous study, we will demonstrate how biases and missing labels in data can lead to a misinterpretation of algorithmic bias, and discuss potential solutions to address these issues.

To conclude, we will summarize the key findings of the thesis and discuss possible future research directions, some directly related to the questions explored in this thesis and some more broadly related to the topic of music consumption and streaming.

1.4   Publications

To conclude this introductory chapter, in this section we list the author's publications that occurred during the PhD thesis :

- Matrosova, K., Marey, L., Salha-Galvan, G., Louail, T., Bodini, O., & Moussallam, M. (2024). Do Recommender Systems Promote Local Music? A Reproducibility Study Using Music Streaming Data. In *Proceedings of the 18th ACM Conference on Recommender Systems* (pp. -).
- Matrosova, K., Moussallam, M., Louail, T., & Bodini, O. (2024). Depict or discern? Fingerprinting musical taste from explicit preferences. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 7(1), 15-29.

Part I

Setting the Stage

Chapter 2

Getting familiar with data from streaming platforms

A music streaming platform is a service, most often accessible through a website and/or a mobile app, that allows users to listen to audio content online in real time, without having to download music files. Platforms, like Deezer, Spotify or Apple Music, operate by hosting large libraries of tens of millions audio files, which users can access via a subscription or 'freemium'[1] model.

The central actors involved include the users (that are supposed to be human listeners), artists (content creators), record labels (rights holders), and the platform itself (as an Internet service provider) [2]. Each actor contributes to, or benefits from, the streaming ecosystem in different ways. Artists provide the content, record labels manage the licensing and royalties, and platforms curate, deliver, and recommend the content to listeners. Users can interact with the platform in multiple ways, such as searching for music, streaming songs, liking different music items (songs, albums, artists) or creating playlists. They can also explore curated playlists and receive recommendations based on their listening behavior. Each of these actions generates data that platforms use to understand preferences and provide personalized experiences.

Data collection on music streaming platforms occurs through a variety of mechanisms. During user registration, platforms typically gather basic information such as the user's age, location, gender, and sometimes even musical preferences, which helps create an initial personalized experience. On the content side, music providers (such as artists or record labels) submit audio files, along with some metadata, including details like the song title, artist name, genre, release date, and album information. Users can interact with the catalog in many different ways. First, they can use the search bar or simply navigate through

---

[1]A business model where basic services are provided for free, typically supported by ads, while advanced features or content require payment.

[2]The Music Streaming Economy – Part 1: The International Music Streaming Boom

the platform in order to find the music they are interested in. They can listen to music by playing (or streaming) and skipping (going to the next track). Finally, they can save the items (tracks, albums, artists) of their choice by liking them (saving it to a list of favourite items) or by adding songs to a specific playlist. All of these interactions are recorded in timestamped logs, that link a unique user ID to a specific item ID. These logs also capture the type of action and the context of each interaction, such as the duration, device used, location, app context (how has the item been accessed by the user) and other relevant details.

In this chapter, our goal is to provide an overview of what data from a music streaming platform can look like, in order to better understand how we can use it in research and for recommendation. In the following sections, we will describe, and also highlight the challenges and limitations of the available data about users, music items, and user-items interactions. To do so we will take the example of different datasets extracted from Deezer, as well as some external studies. We will compare music streaming data to other types of online content consumption and recommendation datasets, to identify common patterns and highlight those that are unique to music consumption. Then, we will take interest in the differences that can occur between behavioral data from streaming platforms, and declarative data collected in surveys about music preferences and listening habits. Finally, we will discuss the importance of making music streaming datasets public, and we will underline the challenges associated with doing so without compromising user privacy.

2.1   Users

Later on in the manuscript, we will talk about the processes sustaining the formation of our music preferences, which include social, geographical and cultural factors. In this section we provide an overview of a subset of the total population of Deezer users, and see what information can possibly be used to know more about them, beyond their music preferences.

At the time of writing this manuscript, Deezer counts about 16M active users, located in more than 180 countries, and among them around 7M are paying users. Like most platforms, Deezer collects data concerning its users. This data includes straightforward information, like self-declared gender and age, country of registration, as well as less obvious things like the type of subscription or the device used for streaming, through which some assumptions can be made on the social status or wealth for example. In the following paragraphs, we use a set of 50,000 random users as an example.

Before diving into the numbers, we have to keep in mind that a significant number of users choose not to disclose personal information, so all numbers should be taken with a grain of salt. For example, in a random sample of 50,000 users, only 31.1% have specified their gender, 27.1% their date of birth, and
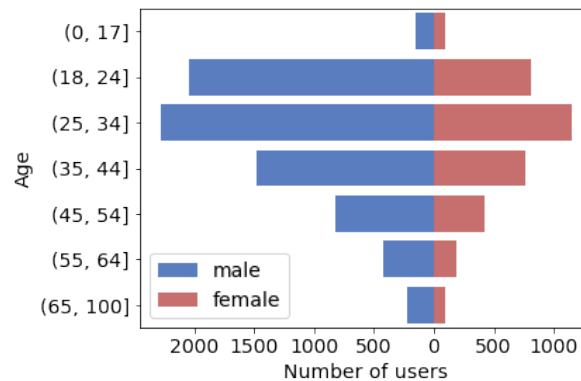
Figure 2.1: Deezer users' demographics (gender and age) of 50,000 random Deezer users.

59.4% their country of residence. However, as the platform now requires these fields to be filled in, we can assume that among active or more recent users, this information is generally available. Even with this, the accuracy of the provided data remains uncertain.

Figure 2.1 illustrates the gender and age distributions of Deezer users who have declared these information. Of these users, 26.7% are female and 65.4% are male. The users' are 33.89 years on average, with a median of 32 years. The most represented demographic overall is young males (18-35 years old). Deezer's largest markets are France, Brazil, and Germany. For users whose country is unknown, or to pinpoint a more precise location, listening history can be analyzed, as each stream is linked to an IP address. Location detection will be discussed further in the section on user-item interactions.

Other types of platforms often collect a lot of data related to the users' social or economic capital. On social networks apps, users usually fill in their university, place of work, share places they have visited, and connect with other users. Shopping websites can use the users' purchasing history to make assumptions on their financial and social position. In comparison, music streaming services collect way less personal data. However some assumptions about the user can be made, for example through variables like the subscription plan and the type of device used to stream music. Deezer has several subscription plans: freemium, premium, duo, family, student. Some users may have temporary access to the premium account through a 3-6-12 months voucher. The type of device, and sometimes connected devices like Bluetooth speakers are registered for each stream. By combining all of this data, some assumptions can be made about the users. To put it simple, a user having a self-paid premium account and using an iPhone is probably wealthier than a person with a freemium account and an Android phone.
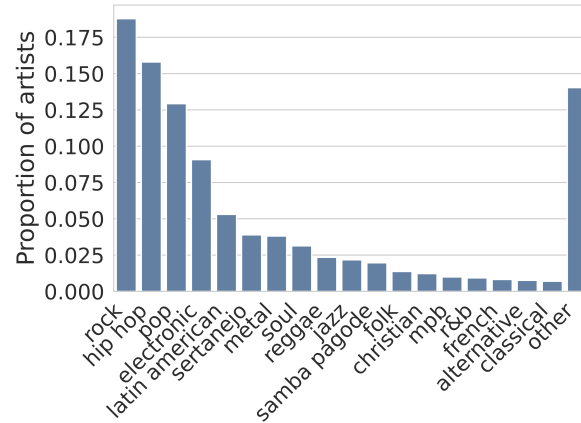
Figure 2.2: Macro genres distribution among favourite artists of 1M active Deezer users.

2.2 Music items

Streaming services usually have huge catalogs of music items : tracks, albums and artists. These items are supplied to streaming platforms by music majors, such as Universal, Warner, Sony and independent distribution services, like Distrokid for example. Each item is usually accompanied by metadata — data that describes the music item. For a track, metadata traditionally includes the related album and artist(s) IDs, release date, release country, information about the provider. Additionally, metadata can contain tags related to the item's genre, mood or country. On Deezer, for example, there are two types of genre tags : macro genres, which are 25 and represent a broad classification, as well as a list of 250 more specific genre tags, where several tags can be associated to one music item. Figure 2.2 represents the distribution of macro genres among the favourite artists of 1M active users from our paper Matrosova et al. (2024b).

Unfortunately, such tags are not provided systematically. In order to make up for the lack of provider tags, platforms (and researchers) often use a variety of techniques, including human annotations, audio analysis and web scrapping (Sordo and Serra, 2014; Humphrey et al., 2013). We will discuss in more detail the existing techniques and challenges of music annotation in Chapter 4. For now we will simply show an example of how tags can be ambiguous or incomplete, on the example of the dataset used in one of our research papers Matrosova et al. (2024a). This dataset contains three months of streaming logs of a total of 30,000 Deezer users, more specifically 10,000 users from each of three countries where Deezer is an important actor in the streaming market — France, Germany and Brazil — in 2019. For each track, we extracted three different country-related tags, from two different information sources: the artist's active country and country of origin (two distinct variables, coming both from internal

Country	Label Source	Labeled Streams	Local Streams Among Labeled	Local Streams Among All
France	Deezer - Activity	76 %	50 %	38 %
	Deezer - Origin	75 %	34 %	26 %
	MusicBrainz	76 %	38 %	29 %
Germany	Deezer - Activity	60 %	40 %	24 %
	Deezer - Origin	62 %	30 %	18 %
	MusicBrainz	69 %	33 %	23 %
Brazil	Deezer - Activity	41 %	48 %	19 %
	Deezer - Origin	36 %	37 %	13 %
	MusicBrainz	38 %	38 %	14 %

Table 2.1: Percentages of (i) labeled streams, (ii) local streams (among the labeled streams) and (iii) local streams (among all streams) in the Deezer dataset used in Matrosova et al. (2024a), by country and label source. A labeled stream corresponds to a stream of a music track that is tagged with a country label. A local stream corresponds to a stream where the user and the streamed artist have the same country label.

Deezer tags), as well as the artist's country according to MusicBrainz[3] (a public music database). Table 2.1 presents the proportions of labeled streams and local streams (among the labeled ones, and among all streams) in this dataset, according to the three label sources.

As we can see, none of the labeling sources provides complete label coverage. Across the three considered countries and label sources, between 24% and 64% of the streams are unlabeled. Also, the proportions of local consumption vary depending on the label source. For example, the artist's country is identified in about 75-76% of streams made by French users, depending on the label source, while for streams from Brazil, this coverage drops to only 36-41%. Country labels are just one example of metadata, but we can assume that the situation is similar in the case of genre and other tags. In the next section about user-items interactions, we will discuss the inequalities in the items' popularity distributions. Later on in Chapter 7, we will explain how those popularity distributions and lack of labels can, together, lead to misinterpretation of users' preferences.

2.3   User-item interactions

The main data that will be of interest for us along this manuscript is the data encoding the interactions between users and items, as it is the most helpful to

---

[3]musicbrainz.org

understand people's behavior on the service, as long as (possibly) their preferences, and to perform recommendation.

There are two main types of activity we can observe on a streaming platform — streaming, which is often considered as an implicit marker of preference, and liking, which can be viewed as an explicit marker of preference.

Additionally, skips and bans can be viewed as implicit and explicit markers of distaste. However, they are less often used in research or for recommendation : bans are quite rare, as they only occur in a context of algorithmic recommendation, and for skips it is hard to know if the reason was the user not liking the song, or if the song simply does not fit into the current user's listening context.

Representing user-item interactions

In the following we will denote by U the set of users, and I the set of music items. Let $V(u)$ be the set of liked or streamed items of user $u \in U$.

One way to represent user-items interactions is through a bipartite graph (Figure 2.3a): let $G(U, I; L)$ be this graph, where U is the set of vertices representing the users, I is the set of vertices representing the items, and L are the edges linking users and items: there is an edge $(u, i) \in L$ if the user u has liked or streamed the item i. For a vertex u in U, $V(u)$ are the vertices in I that are connected with u by an edge. For each item i, let $W(i)$ be the set of users connected to i, and $d(i) = |W(i)|$ its degree.

Another way is to view user-items interactions as a sparse matrix (Figure 2.3b): let M be this adjacency matrix. A matrix is called sparse when most cells are empty, meaning the majority of users interact with only a small subset of the total available items. M can be binary, then $M[u, i]$=1 if the user u has liked or streamed the artist i, and zero otherwise. For streams, it can also be non-binary, then $M[u, i]$ will contain the number of times a user u has streamed an item i (weighted graph).

Likes distributions

We can first take interest in the distribution of likes, on the example of the dataset we use in Matrosova et al. (2024b), which includes liked artists and songs of about 1M randomly selected users, who have been active (have streamed at least once) during October 2022.

On most music streaming platforms, users can explicitly 'like' songs, albums, and artists, which then appear in their personal 'favorites' collection. However, not all users engage with the liking feature. Among these 1M randomly sampled Deezer users, 87.1% have explicitly liked at least one artist, and 88.9% have liked at least one song. All together the users had liked 586,512 artists and 10,822,633 unique songs.

	u_1	u_2	...	u_n
i_1	1	0	...	1
i_2	0	1	...	0
...	...	...	...	...
i_m	1	1	...	0

(a)   (b)
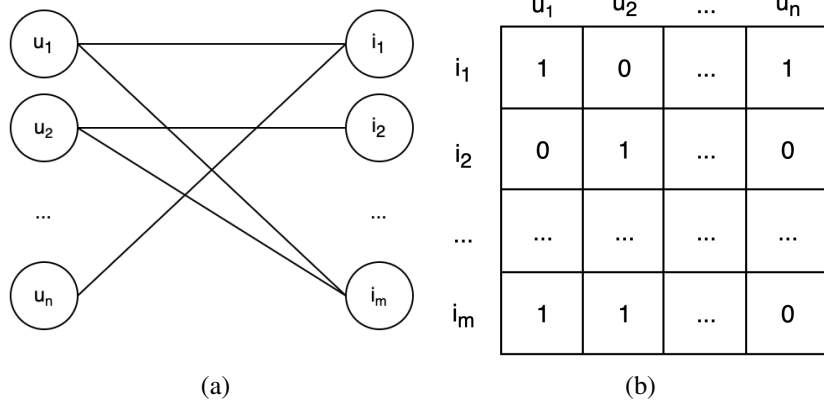
Figure 2.3: User-items interactions represented as (a) a bipartite graph $G(U, I; L)$ and (b) its adjacent matrix M.

(a)   (b)
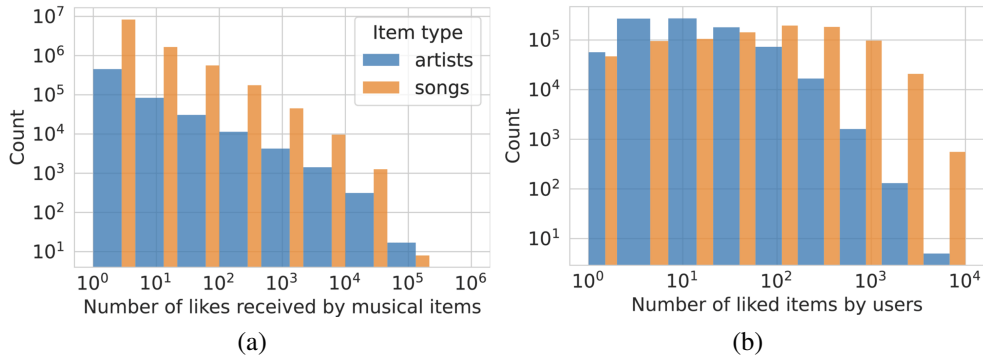
Figure 2.4: Heavy-tailed empirical distributions in the items liked by 1M users. (a)Distribution of artists' and songs' number of 'fans' (i.e. users who coined these artists/songs as 'liked). A large proportion of items is liked by only a few users, while some items are very popular (hundreds of thousands of fans). (b) The distribution of the number of given likes per user follows here again a heavy-tailed distribution, with some users liking ten thousand more items than other users. The proportion of users liking many items drops faster for artists than for songs.

The distribution of users according to the number of artists they have liked follows a heavy-tailed distribution (Figure 2.4a). Half of the users have liked 10 artists or less, with an average of 26 liked artists per user. Some outliers exist, such as one user who has liked 7,497 artists. Users tend to like songs more than artists, with 215 favorite songs by user on average. The user experience on the platform contributes to this gap between explicitly liking artists and songs: indeed, the like button can be easily hit on a song while the user is listening to it, while liking an artist requires the user to specifically go to the artist's page. Also, liking a track is a practical move, as it adds the track into a playlist 'Favourite tracks' which can be listened as a regular playlist, while liking an artist may pursue solely a bookmarking purpose. However most users have still liked at least some artists as it has been a mandatory step during the on-boarding process for the past few years on the platform.

The distribution of music items according to the number of users who like them similarly follows a heavy-tailed distribution (Figure 2.4b). For artists, the median value is equal to one — which means that at least half of them have been liked by only one user — while the average is around 38. The most popular artist has been liked by 86,877 users among the million users which were randomly sampled. We can thus see a huge disparity between the artists, with a few extremely popular artists that attract lots of users, and the majority of artists that are almost unknown. The songs follow a similar popularity distribution, with a median of 1, an average around 18, and a maximum of 75,453 likes.

We did not collect users' favorite albums in this specific dataset, but it is generally observed that albums tend to be liked less than tracks or artists. For example, in a dataset from the RECORDS project, 95% and 92% of users have liked at least one song or artist, respectively, while only 84% have liked at least one album. A Deezer study suggests that people listen to fewer albums than 5-10 years before. It it possible that this habit, that was common among generations raised with LPs and CDs, might today be more typical of 'connoisseurs'. Even fewer users have liked at least one playlist (77%). However, the same Deezer study indicates that the popularity of playlists has increased in recent years, progressively replacing albums in user preferences.

Streams distributions

A fuller knowledge can be extracted from streaming activity, not only because streams occur more often than likes and thus represent more data (every day, Deezer collects 195M rows of streaming logs, which corresponds to 130G of data), but also because they are logged in by the platform along with all the possible information about the context in which the stream was made (more than 50 attributes are recorded for each stream). We will observe the streams' distributions on the example of another Deezer dataset, taken from Matrosova et al. (2024b), describing the streaming activity of 60,000 users from April 1st

	Median	Mean	SD	Max	Total
Day	13	19	20	430	615,373
Week	45	62	59	1,503	2,717,914
Month	123	160	143	3,477	7,714,374
Year	521	655	564	9,274	33,966,616

Table 2.2: Distribution of the number of streams per user (and in total), over a day, a week, a month and a year.

(a)   (b)
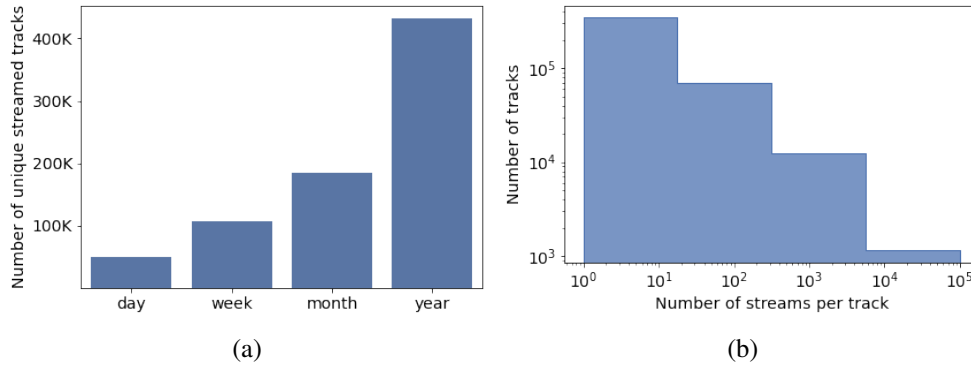
Figure 2.5: (a) Total number of unique tracks streamed by 16M users over different time periods. (b) Distribution of the number of streams per track over a year.

2022 to March 31st 2023.

The intensity of streaming activity varies widely among users (Table 2.2). An average user streams 19 (unique) songs per day, 62 per week, 160 per month and 655 per year. However, some users have an extremely intensive usage of the platform, with up to 430 songs per day, on average, for one of them. With the existence of such active users, the total number of streams quickly builds up, reaching a total of 40M streams for 60,000 users over a year (Figure 2.5a). In the same way as for likes, the distribution of the number of streams per musical item follows a long tail distribution. For example, over a year, more than 100,000 items have been streamed less than 10 times (all 60,000 users considered), while the top 1,000 tracks have been streamed between 10,000 and 100,000 times each (Figure 2.5b). If we look at it in terms of proportion of streams, 12M streams — which is equivalent to 1/3 of all streams — concern the top 1,000 songs.

Streaming context

For each single stream a user makes, numerous attributes are logged in by the streaming platform, including :

- the timestamp

- the duration
- the IP address
- the type of network connection (WIFI, LAN, mobile)
- the (lat, long) coordinates inferred from the geolocation of the IP address (as provided by a third-party service)
- the type of device
- the in-app context in which the stream was originated (there are about 15-20 basic contexts — such as playlist_page, album_page, flow, etc. — that can be grouped in three main classes: organic (the user searched for it themselves), editorial recommendation and algorithmic recommendation)

Duration.   Often users do not listen to a track till the end, and in this case it might be complicated to say if they liked it or not. A common practice in research and recommendation is to only take in account streams that lasted 30 seconds or more (Datta et al., 2018; Anderson et al., 2020, 2021).

Geographical location.   Geographical location is often a key variable in research on music geography, such as studying regional preferences or the spatial propagation of music. The location of an internet-connected device can be estimated through its IP address, which is assigned by ISPs and mapped to regions via external databases. The accuracy of IP geolocation varies: WiFi/LAN connections, which use more stable IPs, tend to provide reliable location data, while mobile networks, with dynamically assigned IPs, are less reliable. In France, around 35% of streams come from WiFi/LAN connections, typically in home or work settings. However, focusing only on these streams may introduce bias, as connection type can correlate with factors like age, access to a computer, or occupation.

Context.   The user can access music on the platform in many different ways. For each stream, Deezer collects hundreds of specific contexts that can be split in three main types. If the user found the track autonomously, for example by using the search bar, or through their own library of playlists, this stream is considered 'organic'. According to Villermet et al. (2021), organic streams are the most common, with 80% of the users who access at least half of their plays autonomously. Editorial streaming involves selections curated by human experts or editors employed by the platform. They are usually take the form of playlists created around moods, activities, genres, or cultural events. It is the least popular way of streaming, with only 8% of the users who stream 38% of editorial content. Last but not least, algorithmic streaming involves music recommendations generated by the platform's recommendation algorithms based on a variety of factors, such as the user's previous listening habits, similar users' preferences, and other data-driven insights. Unlike editorial playlists, algorithmically created

playlists aim to propose personalised selections for each user. 11% of the users mostly stream through algorithmic recommendation, with 58% of algorithmic streams. It may seem that editorial and algorithmic recommendations encompass only a small section of all streams. However, these playlists are often used in the specific context of discovery, and once users have liked a song that was recommended in a playlist, they will most likely access it through their own library the next time. In the same logic, a user can run into a song they already like in a playlist. The context tags only represent the local context of one specific stream, but does not necessarily correspond to the way the user has discovered the song initially.

2.4   Specificity of music streaming data in the context of broader online consumption

After exploring music streaming data through the example of Deezer, to provide some perspective we think it is useful to situate it within the broader context of online consumption, to determine if techniques from other fields can be applied to music, and to uncover the unique characteristics of music streaming data. Many other platforms, such as movie and video streaming services or e-commerce sites, also involve users interacting with large catalogs of items, and it would be interesting to compare music streaming data with several other types of online consumption datasets.

For the sake of simplicity in the scope of this section we chose to focus on the MovieLens (Harper and Konstan, 2015) dataset as an example, for several reasons. First, it is an open-source dataset which has been widely used in research on RS (He et al., 2017; Zheng et al., 2021b). Second, it provides a large set of user interactions with a catalog of movies, which, like music streaming platforms, involves media consumption based on individual preferences. This makes it an good candidate for a comparison with data from a major music streaming platform, even though it is important to keep in mind that similar analyses could be conducted with other datasets from various domains.

More specifically, in this section we will rely upon the 'MovieLens 100K dataset' [4], developed by the GroupLens research group at the University of Minnesota, which contains 100,000 ratings from 1,000 users on 1,700 movies, as well as movie metadata (e.g., genre tags), and user information. We will compare it to a subset of 1,000 users from the Matrosova et al. (2024b) dataset.

Number of items and interactions.   The first thing that makes the two types of data different is the catalog size: the MovieLens dataset counts a total of 27K
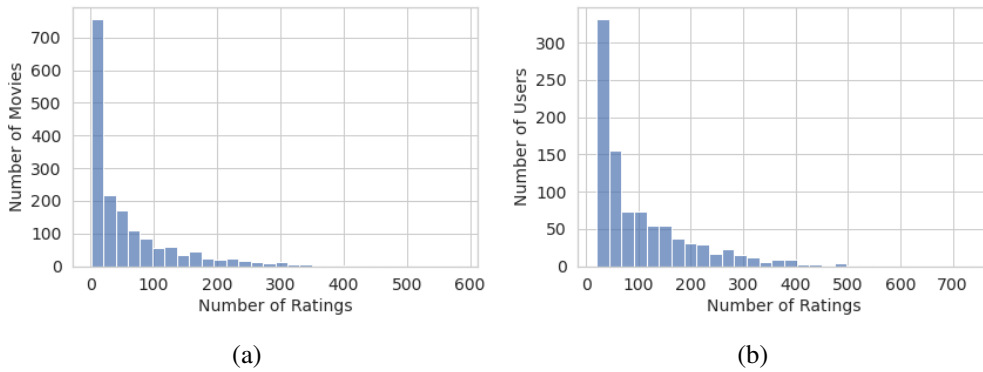
---

[4]Movielens 100K dataset

(a) (b)

Figure 2.6: Long-tailed distributions in the 'MovieLens 100K dataset' ratings. (a) Distribution of number of ratings per movie. A large proportion of movies have been rated by only a few users, while some movies are very popular (hundreds of ratings). (b) The distribution of the number of given ratings per user. Most users have only rated a few movies, while a few are extremely active raters.

items (674K on IMDb [5], the world's most popular source for movie ratings), while the Deezer catalog counts around 120M unique tracks. Consequently, the ratio between the number of users and items is also different. A thousand users on MovieLens have rated 1,700 movies, meaning users and items have approximately the same order of magnitude, while the same amount of users on Deezer like 10,000 artists, and 133K songs on average. And that is only considering likes, not to mention the vastly larger number of streams. Users on Deezer like songs approximately twice as often as MovieLens users rate a movie.

Naturally, the number of interactions with music are more frequent than with movies, not only because there are more available items, but also because listening to a song simply takes less time then watching a movie or a show, and music listening is often practiced to accompany other activities, while watching a movie is usually a self-sufficient occupation. The vast amount of data generated by music streaming, both due to the size of the catalog and the frequency of user-item interactions, presents significant computational challenges. These include the need for efficient storage, processing, and retrieval of data, as well as the development of scalable algorithms capable of handling the large volume of interactions Bertin-Mahieux et al. (2011); Hidasi et al. (2016).

Interactions distributions. Similarly to interactions with music, interactions between users and movies follow a long-tail distribution (Figure 2.6a). However, there are way more disparities between users and between items on Deezer than on MovieLens. While MovieLens users rated between 1 and 700 movies,

---

[5]IMDb

with a median of 65 ratings per user, Deezer users have liked between 1 and 4000 songs, with a median of 65 song per user. Even though the average user makes a comparable number of interactions on both platforms, the distribution of interactions is much more spread out (or skewed) on Deezer, showing a greater variation in how actively users engage with items, whereas MovieLens interactions are more consistent across users.

A similar situation can be observed in the items' popularity distributions (Figure 2.6b). The number of ratings per movie ranges from 27 to 583, with a median of 59, while the number of likes per song ranges from 1 to 90, with a very low median of 1.6. MovieLens shows a narrower and more balanced distribution, meaning that user interactions are more evenly spread across items (movies). Even lesser-known movies still get a decent number of ratings, reducing the disparity between popular and less popular items. Deezer, on the other hand, exhibits a highly skewed distribution: a few popular songs dominate user interactions, while most songs remain under-interacted with, reflecting a much more pronounced popularity gap.

This pronounced long-tail distribution on Deezer means that the data is particularly sparse, which implies several challenges when manipulating such data. Sparse datasets can make it more difficult for models to learn meaningful patterns or create effective recommendations for less popular users or items. Techniques like matrix factorization (Matrix Factorization (MF)) or collaborative filtering (Collaborative Filtering (CF)) might need additional methods (like regularization or data augmentation) to handle the sparsity effectively.

Different representations of preferences.   In the MovieLens dataset interactions consist of ratings, providing a straightforward and nuanced viewpoint on both the users' preferences and dislikes. Ratings are widely used in various domains (e.g., online shopping platforms) to reflect varying degrees of satisfaction.

However, when it comes to music streaming, there is no universal standard for representing user preferences. 'Likes', which may serve as an explicit marker of preference, are binary and lack the nuance of ratings. Alternatively, streaming frequency can act as a proxy for a user's connection with a song, serving as an implicit indicator of preference strength. Unlike movies or products, which are seldom consumed repeatedly within short time frames, it is common for users to listen to the same track multiple times, sometimes within the same day. As noted by Sguerra et al. (2022), the frequency with which a song is listened to correlates with the level of arousal it provides, and users tend to increase listening as they grow more attached to a track, until they reach a point of 'saturation'. The number of repetitions, or the total time spent listening to a song, could thus be used as an implicit form of rating.

As we can see, data from music streaming services require more interpretation compared to explicit ratings. Researchers must decide how to model

user-item interactions — whether based on binary likes, play counts, time spent listening, or a combination of factors. The choice of representation can significantly impact the outcomes of future research and RS, potentially influencing how accurately preferences are captured and predictions are made.

2.5 Enhancing streaming data with surveys

Streaming data are observational data that provide a rich, quantitative insight into user behaviors, and can reveal what users listen to, how often, and their patterns of music consumption over time. However, this type of data alone lacks the qualitative depth that is essential for understanding the broader social, demographic, and psychological contexts influencing these behaviors. Surveys can complement streaming data by providing this missing context, offering insights into why users may prefer certain types of music, how these preferences relate to other types of cultural consumption, to their social position and origins, and more broadly to various social factors that may influence their consumption habits.

Coupling observational data such as streams with self-declared information collected in surveys is not a very common practice, as it requires to both have access to users' streaming history data and have means to contact these people massively, but some researchers managed to do so. For example, Anderson et al. (2021) made users from Spotify answer to psychology tests online in order to investigate the link between music preferences and personal traits. However, most of the time studies use either streaming data or surveys, not both.

Recently, colleagues from the RECORDS project (Renisio et al., 2024) collected traces of online music consumption combined with survey and interview data from the same users, in order to compare the declared and observed listening behaviours. For the study, 100,000 Deezer users had their streaming history data extracted and analyzed. Out of these, about 20,000 responded to the survey. Furthermore, about a hundred of these survey respondents also participated in a detailed individual interview.

The first interesting finding was the nuances between people's declared preferences (what they reported in surveys and interviews) and their observed listening habits (as indicated by their individual streaming history data on Deezer). The integration of digital trace data allowed the researchers to see real-time, actual music consumption that often differed from what participants reported in surveys.

One of the most notable findings was the over-reporting of certain genres and artists in survey respondents, when compared to other music genres (including classical music, and inside jazz and french hip hop, the older, most legitimated/consecrated artists of theses fields). While many participants claimed to frequently listen to classical music, their streaming histories showed that they

actually engaged with this genre far less frequently than other genres they did not declare they listen to. Conversely, some genres like electronic dance music that are widely listened to were under-reported in surveys. Some individuals reported listening to a variety of genres, but their streaming histories showed a narrower range of music preferences. Even though classical music might have been over-declared partially because users may listen to it on other platforms or media than Deezer (because of a poor representation of this specific genre on classic streaming platforms), Renisio et al. (2024) claim that these discrepancies between declared and observed preferences could also be attributed to social desirability bias.

The preferences and listening practices declared by respondents conform to their representation of themselves, their taste and their listening habits, and these representations are embedded within their social or cultural context. This bias is particularly significant in contexts where the social representations of certain music genres are associated with higher social or cultural capital. Another factor contributing to the difference between declared preferences and actual streaming is recall bias. Participants may not accurately remember their listening habits, or may generalize their answers in a way that does not reflect their actual, specific listening behaviors.

Furthermore, the study highlighted significant differences between the solicited users who responded to the survey and those who did not. For example, respondents and non-respondents tended to stream different types of music. Respondents often listened to more English or American rock music, often older bands, whereas non-respondents favored contemporary French rap and R&B, indicating a cultural and possibly generational divide.

Also, based on the music preferences noted, the researchers inferred that non-respondents might include a higher proportion of younger individuals or those from different cultural or socio-economic backgrounds compared to respondents (Figure 2.7). For example, the previously mentioned preference for contemporary French genres might suggest a younger demographic, which is typically less likely to engage in surveys. The same goes for educational levels compared to respondents, based on the correlation typically seen between educational attainment and certain music preferences. Respondents' preference for genres like classical music and older rock might indicate higher educational levels, whereas the non-respondents' preferences could imply a different educational profile.

These differences in social characteristics and preferences between respondents and non-respondents raise important questions about the biases that can occur in survey-based studies, especially considering that a lot of studies on music preferences are made on pre-selected populations (students for example (Rentfrow and Gosling, 2003; Brown, 2012; Langmeyer et al., 2012)). Moreover, this kind of human bias may not only concern surveys, but also data col-
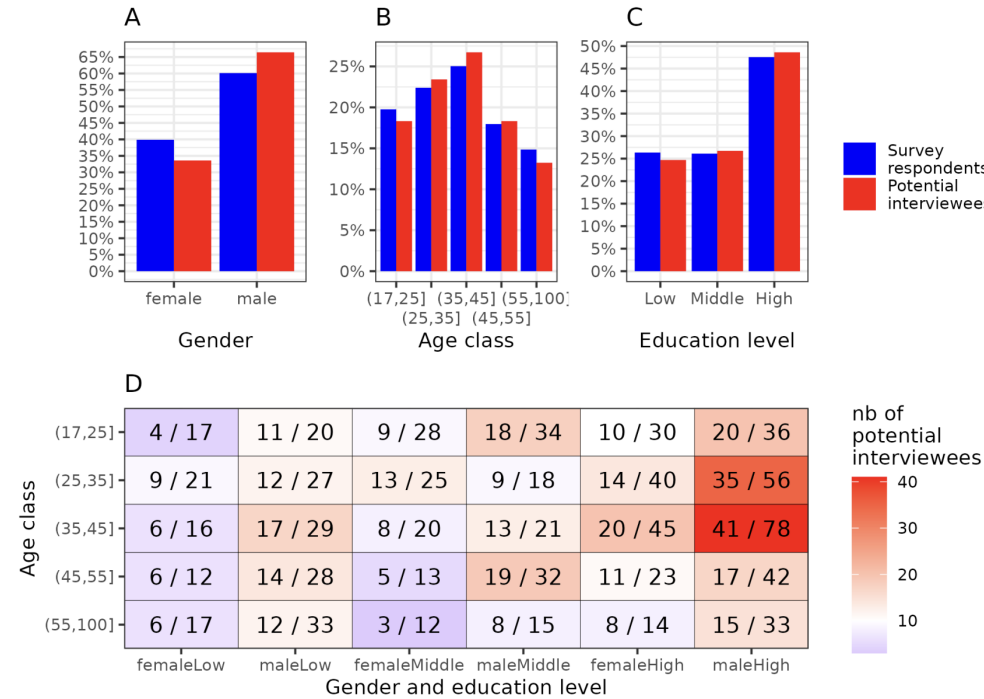
Figure 2.7: Differences in social characteristics of online survey respondents, according to their willingness to take part to an interview (from Renisio et al. (2024)).

lected online — depending on the service/website, or on the way the sampling is done, the users may represent different socio-demographics. Simply focusing on users of streaming platforms already introduces bias, as they typically represent younger demographics from developed countries.

In research about music consumption, only a few datasets are made public, as they are most of the time collected through private streaming services. Most papers then either rely on the same few publicly accessible datasets, possibly reproducing the same biases over and over, or using private datasets, which limits the possibility of reproducing their results by peers. Thus, in order for a more qualitative and reproducible research, opening datasets containing users' online behaviour seems necessary. We will discuss the challenges and limitations of this process in the next section.

2.6 Anonymization challenges in streaming data

As we have seen in the previous sections, streaming data contains a lot of different information that is more or less private and may compromise the identity of users. Accordingly, it should not be disclosed as it is. Most often, datasets are anonymized by simply hashing users' unique IDs on the platform. This method may be sufficient, but it all depends on what kind of data is disclosed.

First, there is the time and geolocation that the stream is tied to. A major work by De Montjoye et al. (2013) explored the privacy implications of individual human mobility data. They collected 15 months of data from 1.5 million individuals, recording spatio-temporal points based on mobile phone interactions, specifically whenever a user initiated or received a call or a text message. The study found that just four spatio-temporal points, selected at random, were sufficient to uniquely identify 95% of individuals in the dataset. Their dataset counts an average of 114 interactions per month per person, which is a little less than streaming activity (160 monthly streams on average per user on Deezer for example). Therefore, streaming data, if shared with the timestamps and locations, can be compromising. Based on the same principle, different variables like age, genre, type of used device etc. may, combined together, identify users in a unique way.

Moreover, the interactions with the music content itself can be unique for each user. Narayanan and Shmatikov (2008) applied de-anonymization techniques to the Netflix Prize dataset, which consists of anonymized movie ratings from 500,000 subscribers, demonstrating that minimal knowledge about an individual could easily identify their record in the dataset. With knowledge of just 2-8 movie ratings and some approximate dates (with a possible 14-day error), an adversary could most of the time uniquely identify a user's record within the dataset. Two exact ratings and dates (with a 3-day error tolerance) were sufficient to uniquely identify 68% of the records in the dataset. With 8 movie

ratings, where two could be completely wrong, and dates that might have up to a 14-day error, the adversary could uniquely identify 99% of the records in the dataset. Once again, considering people more often stream or like songs than they leave ratings to movies, users' interactions with music on streaming platforms are most likely unique.

What are the possible ways to anonymize such data? Traditional anonymization methods include data masking (i.e. hashing users' IDs), generalization (i.e. aggregating items by broader categories) direct suppression of sensitive data, of mixed, aggregation techniques like k-anonymity (Sweeney, 2002). However, these techniques seem inefficient when applied to high-dimensional datasets like movie (or music) streaming data. The uniqueness and sparsity characteristic of high-dimensional data mean that even when direct identifiers are removed, the unique combinations of attributes can still uniquely identify individuals.

Additionally, the high dimensionality adds to the challenge, as increasing the number of attributes in a dataset amplifies the likelihood of each record being unique or nearly unique. In response to these limitations, Narayanan and Shmatikov (2008) advocate for more sophisticated privacy-preserving techniques such as differential privacy. This approach consists in integrating randomness into the data release process, ensuring that the presence or absence of any single individual in the dataset does not significantly affect the overall outcome of data queries. For example, in the case of the Netflix Prize dataset, the date or exact ratings could be changed slightly. In the case of music, we can imagine changing some songs in one's streaming history to similar tracks.

However, adding noise to data inherently reduces the accuracy of the data, which means a middle ground must be found in order to protect users' privacy while preserving enough patterns from the initial data to perform the intended task.

An important thing to note — some of the data, on Deezer and other platforms, are publicly available by default, and collectable through APIs, like for example liked items. Thus, when thinking of anonymization, it is important to not only look at a dataset alone, but also take in account the context of this publicly available information. We will discuss this in more detail later on in the manuscript.

Chapter 3

Understanding and modeling musical taste

"Musical taste is the full mix of musical and cultural dimensions—from the macro level of genre, style, and era to the micro level of distinct musicological attributes—that at any given moment and in any particular configuration correspond to an individual's liking and appreciation." Gasser (2019)

Streaming platforms are particularly interested in understanding their users' music preferences, as it is a necessary step in order to make coherent recommendations. Though, capturing musical taste patterns and understanding why people like the music they consume is a challenge that was taken up by scientists from many different fields long before the emergence of streaming services.

A lot of the research conducted prior to 2010 was constrained by data limitations. Today, the availability of large volumes of streaming data enhances our comprehension of musical preferences, which in turn helps in the advancement of MRS. The concept of musical taste and its possible origins is vast, and it is impossible to describe all the existing studies that deal with it, at least within the scope of this manuscript. That is why this chapter is intended as a starting point into the topic, without claiming to be exhaustive.

First, we will examine the connection between musical taste and various aspects of human identity, drawing from research across sociology, psychology, and cultural studies. We will start with sociological studies that show how one's social background, including factors like social position and education, can influence their musical preferences. Then, we will look at psychological research to understand the relationship between what psychologists in this field call personal traits, and the music people prefer, exploring whether our music choices reflect our personalities or shape them. Finally, we will discuss how cultural differences affect music preferences, highlighting how music tastes can vary significantly across different regions and cultures. Through this discussion, we

aim to shed light on how music preferences are not just personal choices but are shaped by a complex interplay of social, psychological, and cultural factors.

In a second section, we will explore the changing nature of musical taste. Why and how does it evolve during different periods of our life? How do we choose music depending on our activity, entourage, or even weather? Also, since the spread of streaming services, how do recommendation algorithms influence our music choices?

Last but not least, this chapter focuses on the methods used to capture and quantify people's music taste, a task that is often more complex than it initially appears. To cite a few examples, differences between declared and observed preferences, the lack of consensus to label music styles, and the difficulty of interpreting users' online behaviour are some of the struggles that researchers can run into. We will explore how these methods can be adjusted to navigate and overcome these obstacles.

3.1 Primary origins of music preferences

3.1.1 Social background

Highbrow and lowbrow culture

In 1984, Bourdieu (1984) writes *'La Distinction: Critique Sociale du Jugement'*, a pioneering work in sociological literature. The book stands out for its unprecedented approach in examining the intricate relationship between cultural preferences and social stratification, which Bourdieu defines through a multi-dimensional analysis of social, cultural, and economic capitals 3.1. This book was among the first to use a rich array of data, dissecting cultural preferences across various domains such as music, painting, and even clothing, food, interiors, etc., thereby offering a nuanced understanding of how culture serves as a means of social reproduction.

'La distinction' is based on a detailed survey on the cultural practices and preferences of 692 (in 1963), and later an additional 1217 (in 1967-1968) people from Paris, Lille and a small provincial town in France. In the case of music, some pieces were pre-selected by the author, in a way to present a gradient within different genres : for example, classical music ranges from the compositions of highly regarded composers, like Bach's 'The Well-Tempered Clavier', to more commercialized forms such as Viennese waltzes. The participants were then asked questions about the selected music items.

First, respondents were asked to name the composers of a list of 16 music pieces :

- George Gershwin — Rhapsody in Blue
- Giuseppe Verdi — La Traviata
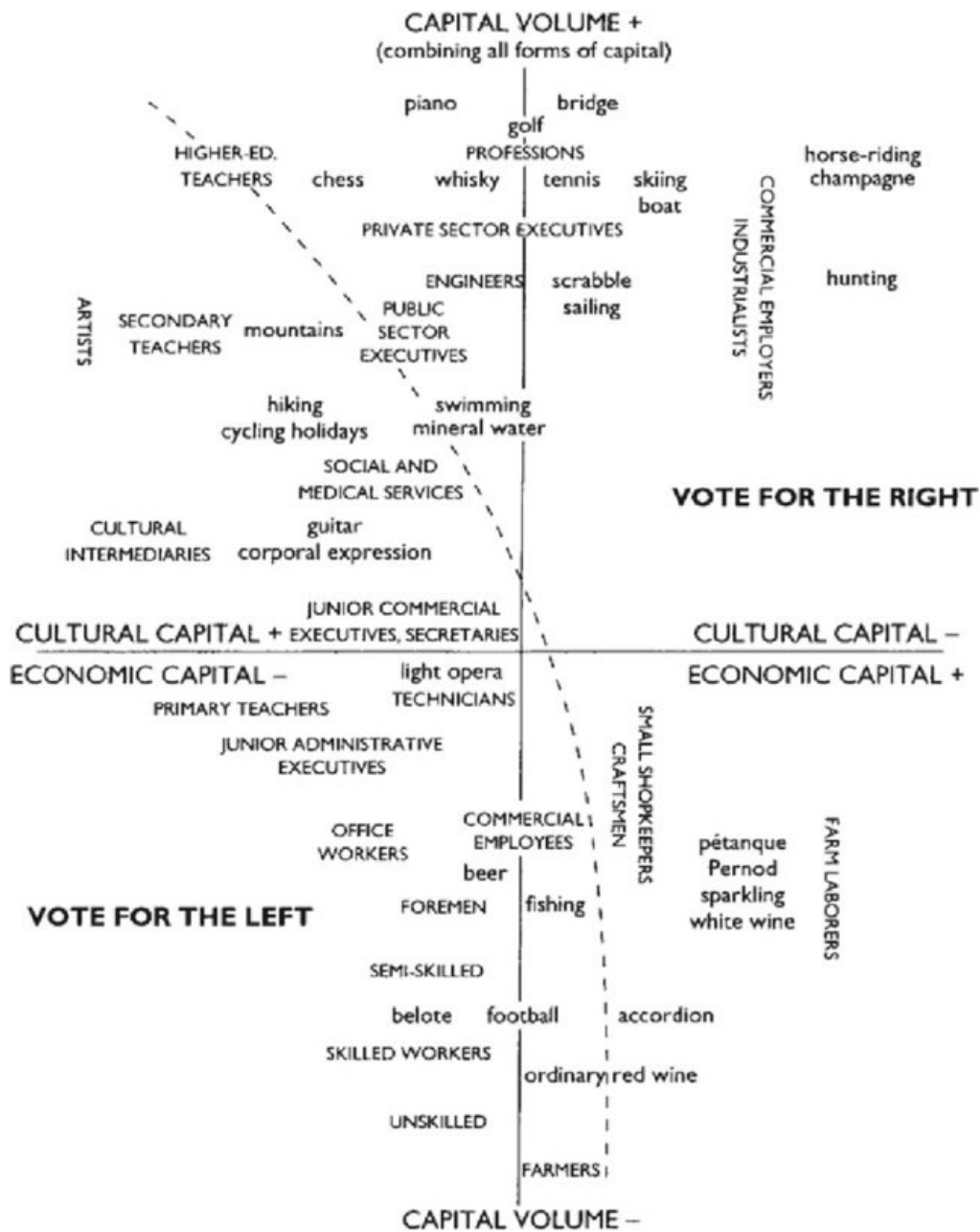
Figure 3.1: Social class stratification according to Bourdieu (1984). He discerns three main groups: the dominant class (high level of cultural and economic capital), the middle classes (lower level of overall capital), and the working class (low level of capital). Within each group, some smaller categories exist, for example the *'intellectual segments of the dominant class'* are people with high level of overall capital, and high cultural capital in particular (e.g., higher education teachers).

- Maurice Ravel — Concerto for the Left Hand

- Wolfgang Amadeus Mozart — A Little Night Music

- Georges Bizet — L'Arlésienne

- Aram Khachaturian — Sabre Dance

- Igor Stravinsky — The Firebird

- Nikolai Rimsky-Korsakov — Scheherazade

- Johann Sebastian Bach — The Art of Fugue

- Franz Liszt — Hungarian Rhapsody

- Maurice Ravel — The Child and the Spells

- Johann Strauss II — The Blue Danube

- Richard Wagner — Twilight of the Gods

- Antonio Vivaldi — The Four Seasons

- Johann Sebastian Bach — The Well-Tempered Clavier

- Pierre Boulez — The Hammer without a Master

A mere 67% of individuals with only a primary education (French CEP or CAP diplomas) could identify two out of sixteen composers, in contrast to just 7% among those with a degree higher than a bachelor's. In a more striking example, none of the surveyed workers or employees could identify more than twelve of the sixteen composers , while this task was completed by more than half of the artist and teacher demographic. This illustrates a clear disparity in cultural knowledge across different educational and occupational backgrounds. Activities like practicing an art or playing a musical instrument, which often require cultural capital acquired outside of school, also show a strong correlation with social class.

Then, through a set of questions about music preferences, Bourdieu identifies three main categories of taste, each associated with different social and educational backgrounds:

- Legitimate Taste: This category corresponds to the highest educational levels and the dominant social class. It includes classic and highly regarded works such as *'The Well-Tempered Clavier'* and *'The Art of Fugue'* by Johann Sebastian Bach or *'Concerto for the Left Hand'* by Maurice Ravel. This taste extends to emerging legitimate arts like cinema, jazz, and even certain forms of *'chanson'* (french pop music), as like Léo Ferré and Jacques Douai, recording French artists at the time the survey was conducted. Legitimate taste is most prevalent among those with the highest education.

- Middle Taste: This taste encompasses less known works of *'major'* arts, and major works of *'minor'* arts. Examples include George Gershwin's *'Rhapsody in Blue'* and Franz Liszt *'Hungarian Rhapsody'*, or artists like Jacques Brel and Gilbert Bécaud. Middle taste is more common in the

middle classes than in either the lower classes or the *'intellectual segments of the dominant class'* (Bourdieu, 1984, p. 22).

- Popular Taste: Characterized by a preference for either *'devalued'* classical music, due to its widespread popularity, such as Johann Strauss's *'The Blue Danube'* or *'La Traviata'* by Giuseppe Verdi, or *'light'* music by artists like Mariano, Guétary, or Petula Clark (that Bourdieu interestingly calls *'completely devoid of ambition or artistic pretense'* (Bourdieu, 1984, p. 22)). This taste reaches its highest frequency in the lower classes.

While Bourdieu's work offers interesting insights into the relationship between cultural tastes and social stratification, its methodology and scope present notable limitations. First, the classification of music within the study lacks a coherent structure, oscillating between broad genre categorizations and specific music pieces or composers, without offering a unified framework for understanding musical preferences. Bourdieu has been criticized for his 'miserablism' — a vision of lower classes through the lack (of culture) instead of difference. And indeed, in the case of music, rating the cultural knowledge through a fixed list of composers or music pieces might create bias and show only a part of the picture.

Second, the temporal context of Bourdieu's research significantly limits the applicability of his findings to contemporary audiences. The classification of music as *'legitimate'* and *'popular'* made sense at the time the book was written, however, it might not hold the same validity in contemporary times. Bourdieu himself states that, between the two waves of survey, the only results that changed were those about music, specifically *'chanson'* (french popular music), which is *'subject to more rapid renewal'* (Bourdieu, 1984, p. 665). The evolution of cultural norms and the blurring of lines between *'high'* and *'low'* art have challenged these classifications, a point that will be explored further in this section.

Finally, even though Bourdieu's methods can be applied more broadly, *'La distinction'* only focuses on French society. Other countries may show different patterns between musical preferences and social distinctions, because of their particular social structure.

Omnivorism and univorism

In 1992, Peterson (1992)'s *'Understanding audience segmentation: From elite and mass to omnivore and univore'* challenges the traditional view of cultural stratification in terms of audience segmentation in the arts, particularly music. Peterson sought to move beyond the elite-mass distinction, as conceived by Bourdieu, proposing a new framework of 'omnivores' and 'univores' to better capture the complexity of cultural consumption patterns.

The study was based on the 1992 Survey of Public Participation in the Arts,

conducted by the United States Bureau of the Census, which offers information on participation in the arts, such as ballet, opera, plays, museums, and concerts, paired with demographic information including age, sex, race, marital status and education level, of 18,775 Americans aged 12 and older. Peterson defines nineteen distinct groups of occupations, similarly to Bourdieu. To examine and quantify the relationships between different occupational groups and their music preferences, a log-multiplicative model was used, which is a type of regression analysis that converts the original data into a series of multiplicative terms expressed on a logarithmic scale.

The key finding of Peterson's research is the identification of what he calls cultural omnivores and univores. In the case of music, omnivores are people who listen and are open to different genres, while univores are those who stick to a specific music style. Contrary to Bourdieu's result, stating that high-status groups prefer elite cultural forms such as symphonic music and opera, the study showed that individuals in higher occupational groups appear to have eclectic tastes. They enjoy a wide range of music genres, encompassing both elite and non-elite forms, such as country and rock. This challenged the traditional view of a cultural elite disengaged from popular culture, revealing a more diverse and inclusive approach to cultural engagement among the higher social strata. In contrast, lower occupational groups demonstrated more specific and limited tastes.

Peterson's study introduced a new paradigm in understanding cultural tastes and preferences. By moving beyond the simplistic binary of elite versus mass to a more subtle omnivore-univore classification, his research provided a better understanding of the contemporary cultural landscape. Later on, a similar study in France (Coulangeon, 2005) examined the distribution of musical preferences across not only different social groups but also generational lines. Multiple Correspondence Analysis on answers of 4,074 participants of the 1997 French Cultural Participation survey revealed that the omnivore-univore classification was observed mostly among younger generations, confirming that it is indeed an emerging cultural consumption pattern.

Taste and distaste

In 1996, Bryson (1996) extends prior studies by concentrating on dislikes instead of preferences. She bases her study on data from the US 1993 General Social Survey, which includes a set of questions on musical taste, where 1606 respondents were asked to evaluate each of 18 music genres on a five-point Likert scale, ranging from 'like very much' to 'dislike very much.' This approach allowed Bryson to derive a measure of musical exclusiveness by counting the 'dislike' and 'dislike very much' responses for each respondent. Statistical methods, such as Ordinary Least Squares regression analysis, were used to understand the relationship between musical dislikes and various socio-demographic factors,

including education, political liberalism, and racial attitudes.

Like Peterson, Bryson found that higher education correlates with a broader acceptance of diverse musical genres, challenging Pierre Bourdieu's theory of higher social status leading to more exclusiveness in cultural tastes. Moreover, the study showed a strong correlation between political liberalism and musical tolerance: more educated and liberal individuals were generally more accepting of a wide range of musical styles.

However, the study revealed a nuanced aspect of cultural tolerance among more educated individuals. While they generally showed a broader acceptance of diverse musical genres, indicating a higher level of cultural openness, there was a concurrent trend of exclusion towards certain music styles. Specifically, genres like heavy metal, gospel and rap, often associated with lower educational levels, were excluded from their range of preferred music.

The study also showed that racist tendencies increased the likelihood of disliking genres whose fans are disproportionately non-white. By taking interest in dislikes, Bryson's work brings a new perspective to Perterson's omnivore-univore theory, showing that while cultural tolerance can indicate higher social status, it also maintains class-based and racial exclusions in cultural preferences.

An individualistic approach

While previous research predominantly attributed specific behavior to different social groups, Lahire (2008) introduces an interesting shift in perspective, questioning how cultural practices and preferences vary on an individual level. He suggests that individual choices in culture go beyond class distinctions, and people's interactions with culture are more personal and varied than class-based models suggest. This approach does not contradict Bourdieu's elite-mass theory or Peterson's univore-omnivore model, but rather complements them.

First, rubbing shoulders with people from a different social or cultural background, for example, interacting with people from different social or cultural backgrounds at school, at work, or within personal relationships is not uncommon and may expose individuals to art forms that are atypical for their social class.

Lahire also views cultural consumption as dynamic and situational rather than static and exclusively tied to class identity. Depending on one's mood, company, or social setting, preferences may vary. However, the role of context is an aspect previously overlooked by sociologists (we will explore the importance of context in music consumption later in this chapter).

Finally, he observes that individuals often exhibit a wide range of cultural practices and preferences. They may have a penchant for a certain aesthetic not influenced by their social circle, which might be explained on an intra-individual level. This naturally leads us from the social model of taste to a more psycho-

logical definition, which is the subject of the following subsection.

3.1.2 Psychology and personal traits

Survey-based pioneering studies

Some of the intra-individual preferences and perceptions of aesthetics can be linked to personal traits, rather than by the surrounding social context. This subject has been of interest for psychologists, starting from the late 90s. One of the first studies that aimed to explore the link between music preferences and personality traits was *'The Do Re Mi's of Everyday Life: The Structure and Personality Correlates of Music Preferences'* Rentfrow and Gosling (2003).

1,704 students from an American university had to complete different personality tests, as well as questionnaires about self-esteem, depression, self-views and cognitive abilities. In particular, the Big Five Inventory (Big Five Inventory (BFI)) was used, which is a popular tool for assessing the major dimensions of personality, commonly referred to as the Big Five traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Developed to measure these traits in a straightforward and efficient manner, the BFI includes 44 items that respondents rate on a Likert scale, reflecting how much they agree with statements about themselves. Here are a few example of statements that reflect the different personality traits:

- Openness to Experience: 'I have an active imagination.'
- Conscientiousness: 'I am always prepared.'
- Extraversion: 'I am the life of the party.'
- Agreeableness: 'I sympathize with others' feelings.'
- Neuroticism: 'I get upset easily.'

Through another experiment from the same study, authors identified four music preferences dimensions. A more detailed discussion of this segmentation of music can be found in Section 4.2.1, here we will simply take it as a given. To explore the relationship between music preferences and personality, scale scores were calculated for each dimension, and then analyzed the correlations between these dimensions and the scores obtained from the various personality tests. Table 3.1 shows the found patterns.

Interestingly, the 'Intense and Rebellious' dimension would be expected to be correlated with emotional stability, depression, and self-esteem, as it is associated with emphasizing negative emotions. However, there were no such correlations with this or any other music dimensions, suggesting that chronic emotional states might not strongly influence music preferences.

---

[1]Tendency to express thoughts and feelings as soon as they come to mind (from the acronym B.L.I.R.T. for 'Brief Loquacious and Interpersonal Responsiveness Test' (Swann Jr and Rentfrow, 2001)).

Dimension	Genres	Positive correlation	Negative correlation
Reflective and Complex	Blues, Jazz, Classical, Folk	Openness to new experiences, self-perceived intelligence, verbal ability, political liberalism	Social dominance orientation, athleticism
Intense and Rebellious	Rock, Alternative, Heavy Metal	Openness to new experiences, athleticism, self-perceived intelligence, verbal ability	-
Upbeat and Conventional	Country, Soundtracks, Religious, Pop	Extraversion, Agreeableness, Conscientiousness, conservatism, self-perceived physical attractiveness, and athleticism	Openness to New Experiences, social dominance orientation, liberalism, verbal ability
Energetic and Rhythmic	Hip-hop, Soul, Funk, Electronic	Extraversion, agreeableness, blirtatiousness[1], liberalism, self-perceived attractiveness, athleticism	Social dominance orientation, conservatism

Table 3.1: Music preferences dimensions, corresponding music genres, and their correlation with personality traits found in Rentfrow and Gosling (2003).

Similar studies were replicated in the US (Zweigenhaft, 2008), Canada (George et al., 2007; Miranda and Claes, 2008), Netherlands (Delsing et al., 2008), Germany (Langmeyer et al., 2008), Brazil (Pimentel and Donnell, 2008), Malaysia (Chamorro-Premuzic and Furnham, 2007) and Japan (Brown, 2012), overall leading to similar results. However, a few cultural specificities were also observed, which we will discuss later in this chapter.

Scaling up using online data

In all previously cited studies, the measurement of musical preferences relies on self-reported likes of musical genres. This approach can be problematic due to the lack of consensus on genre categorization, variability in genre interpretation among participants, and the questionable representation of actual listening behavior. Moreover, the reliance on college student samples might lead to biased results due to the social influences prevalent among young adults' musical choices.

The spread of the Internet and social media has launched a massive wave of data collection. People became willing to fill in personal information for the whole world to see. Massive datasets, including socio-demographic information, likes, online tests provided a propitious setup for scaling up previous studies and giving new answers to old research questions.

Nave et al. (2018) conducted two studies aiming to predict personality traits from musical preferences, using data collected online.

The research included a large, diverse group of 22,252 users from 153 countries, who used MyPersonality, a Facebook app. To gather data, the study used the International Personality Item Pool questionnaire for personality traits and evaluated musical preferences in two ways: first, through listening to 25 different 15-second music clips categorized according to the Mellow, Unpretentious, Sophisticated, Intense, and Contemporary (Rentfrow et al.'s five-factor model) (MUSIC) model (Rentfrow and Gosling, 2003); and second, through music-related Facebook 'likes'.

The prediction was made using LASSO regression (Tibshirani, 1996). For the 15-second music listening data, participants' preference ratings for the musical excerpts were used as direct inputs. The large Facebook 'likes' dataset was reduced through Singular Value Decomposition (Singular Value Decomposition (SVD)). Demographic variables like age and gender were also incorporated in the model. The model's accuracy was evaluated using Pearson's correlation.

Even though Facebook 'likes' of artists outperformed the 15-second music excerpts in predicting personality traits, both methods showed significant correlations with personality. While both methods similarly predicted openness and extraversion, traits like neuroticism and agreeableness were less consistently predicted across the two studies.

This result is particularly interesting because it is among the first to reveal a divergence between 'declared' and 'observed' preferences: as Facebook 'likes' are public, and users might 'like' certain artists to align with peer preferences or convey a particular self-image, rather than because they genuinely enjoy the music, they may not necessarily be correlated with the reactions to 15-second music excerpts, which were unfamiliar and lacked any social context. Unfortunately, the comparison between the two methods was only made in terms of their ability to predict personality traits, and authors did not confront the two types of data directly.

Overall, the study's findings align with Rentfrow et al. (2011)'s five-factor MUSIC model, demonstrating that personality traits are stronger predictors of music preferences compared to gender and age alone. This challenges the conclusions of North (2010) and Schäfer and Mehlhorn (2017), who suggested otherwise. However, the found correlations between personality traits and music preferences were moderate, indicating that personality only partially explains individual differences in musical tastes.

Behavioural data from streaming platforms

In contrast to earlier methods, Nave et al. (2018) introduced two key innovations: first, by using behavioral data where participants actively listened to and rated music based solely on the audio, free from labels like genre or artist name. This avoids the subjective interpretation of genre labels, which can vary widely between individuals, ensuring that everyone is evaluating the same musical content.

Second, he incorporated online data, in the form of Facebook 'likes', allowing the collection of huge amounts of data compared to traditional methods like surveys and interviews. However, this data is still somewhat declarative, as these 'likes' are public and might be influenced by how individuals want to be perceived. With the rise of streaming platforms, researchers now have access to real-life listening histories, offering the most authentic data yet. This eliminates the need for controlled lab listening sessions or self-reported preferences.

In 2021, Anderson et al. (2021) aimed to predict personal traits through a big and diverse set of metrics derived from streaming data, including mood, genre, and behavioral metrics, marking a significant shift from previous research methods.

Like in several previously discussed papers, this study uses the BFI to determine personal traits. 5,808 US Spotify users participated, their answers to the test were collected, as well as all their activity on the platform over a three-month period (age, gender, free/premium Spotify plan, streamed music and overall in-app behaviors). Listening data was mapped to Spotify's 66 genres and 25 mood categories (according to Gracenote (2016) classification), normalized to percentage terms per user. In addition to that, authors computed 123

derived metrics to capture users' music listening behaviors.

These metrics go beyond simple genre preferences and dive into aspects like diversity and discovery, such as genre entropy, which measures the variety of genres a user explores, or track discovery rate, tracking how often users listen to new songs. Listening habits were quantified by metrics like skip rate, showing how frequently users skip songs, and time of day concentration, which indicates whether a user listens consistently at the same times.

Platform usage was also measured, such as the diversity of devices used (e.g., phone, computer, speaker). Other metrics captured contextual listening behaviors, including playlist following rate and the extent to which users listen to music from their formative years. Finally, the study examined audio attributes like tempo, loudness, and danceability to map a user's acoustic profile, as well as mood and emotion metrics, to track engagement with music that elicits specific emotional responses(e.g., lively or sentimental).

In total, 211 mood, genre, demographic, and behavioral metrics were used to predict personality traits. In comparison to previous studies, which relied on a limited selection of artists, genres, or short music clips paired with basic scales like Likert ratings or binary like/dislike options, this study marks a turning point in methodology, proposing an unprecedented amount and diversity of metrics, which not only focus on the music itself but also on the ways it is consumed.

Given the varying distributions of the predictors, the numerical values of each predictor were transformed to achieve a standardized distribution. The appropriate standardization technique was selected based on the distribution of each feature — for example, a log transformation was used for values that span several orders of magnitude, like the number of plays in the last three months.

Having many predictors increases the risk of overfitting. Two common methods were considered: LASSO regression and ridge regression. LASSO regression is preferred for its interpretability and ability to reduce model dimensionality. Ridge regression, on the other hand, performs better when predictors are highly correlated but does not eliminate any predictors. Since the two methods performed equally well, authors ended up using elastic net regularization, which combines both techniques, and outperforms them individually.

To account for potential nonlinear relationships between variables (e.g., differences in behavior between free and premium users), the study also performed random forest regression. Both models were optimized by tuning their hyperparameters using grid search to minimize the root-mean-square error. Despite nonlinear models occasionally outperforming the linear ones, the overall performance between the two was similar, and the best-performing results were reported.

The study found that personality traits, based on the BFI, could be predicted from Spotify users' music listening behaviors with moderate to high accuracy, outperforming Nave et al. (2018)'s model based on Facebook 'likes'. Emotional

Stability and Conscientiousness were the most predictable traits, with emotion-
ally stable users favoring soothing or emotionally satisfying genres like Blues
and Soul, while avoiding intense or aggressive music such as Emo. People high
in Openness tended to explore diverse and less mainstream genres. Extroverts
leaned on social music choices, listening to others' playlists and enjoying ener-
getic genres like Reggaeton, while agreeable users preferred mellow music like
Jazz, steering clear of aggressive genres like Punk.

It was also found that habitual behaviors, such as skip rates and music dis-
covery patterns, played an important role in predicting personality, with emo-
tionally stable users skipping less, and open users discovering more new music.

These findings demonstrated greater predictive accuracy compared to previ-
ous studies, highlighting the value of real-life music streaming data for research.
However, the study was conducted solely on Spotify users from the U.S., which
may not be representative of a broader, global audience. Music preferences and
listening habits can vary across cultures, and Spotify users themselves represent
a specific demographic that may not reflect the general population, potentially
limiting the generalizability of the findings.

The role of lyrics

In addition to musical attributes, lyrics are also an important part of a song,
and can drive different emotions. Several studies take interest in lyrics specifi-
cally, using natural language processing (Natural Language Processing (NLP))
techniques. For example, Mishra et al. (2021) explore differences in emotional
language among fan communities of different music genres, linking these differ-
ences to the emotions contained in the lyrics, and suggest that extreme emotional
expressions in certain genres could serve as a cathartic release for fans. Another
study (Alaei et al., 2022) shows that individuals' favorite songs' lyrics reflect
their attachment style.

3.1.3   Cultural and geographical environment

Alongside social background and personal traits of character, the place where
we grow up and live can play a role in shaping our musical taste. For one, we
are emotionally more sensitive to our mother tongue than to any other language
(Caldwell-Harris, 2014), suggesting a stronger bond with music with lyrics in
our native language.

Musically speaking, timbre, tonality, rhythm are often region specific (Gómez
et al., 2009), and being exposed to music with certain patterns in childhood
makes us more likely to enjoy these patterns in adulthood (Pereira et al., 2011).

Also, not all areas offer the same opportunities for music discovery and cul-
tural activities : for example, the presence of cultural infrastructure like concert

halls, music schools or conservatories, and the diversity of the people living in that area, may also play a role in the shaping of one's musical taste.

Historically, since the late 1960s, geographers have taken an interest in mapping musical preferences across different regions of the world. These studies, however, have often lacked a structured approach and tended to focus on specific regions or music genres. For instance, DeHart (2018) examined the global popularity of metal music, Ford (1971) traced the spread of rock'n'roll in the US, and Oduro-Frimpong (2009) analyzed the local adaptation of U.S. rap music in Ghana. Music geography has various objectives, including the delimitation of musical regions (Nash, 1968), tracing the origins and diffusion of musical phenomena (Ford, 1971; Carney, 1977), and examining the evolution of a music style in response to its geographical context (Oduro-Frimpong, 2009; Oh and Park, 2013). Comprehensive reviews and categorizations of these, and other studies have been attempted by Nash and Carney (1996) and Carney (1998), who sought to organize the field into different thematic axes.

Also, unlike sociology and psychology, where a lot of studies rely on surveys, geographers mostly rely on data from/about industry stakeholders such as artists, radio stations, record stores, and music labels. This is especially the case for studies that were made before the advent of streaming technologies, when localized data on music consumption was not yet available.

Cultural differences in music perception

Do we all experience music in the same way, no matter our cultural background? Are our listening habits innate or learned? Research suggests that our experiences with music are heavily influenced by cultural exposure.

For instance, McDermott et al. (2016)'s study conducted on Tsimane people — an isolated native Amazonian population — revealed that they did not show a preference for consonance, unlike populations familiar with Western music. This suggests that exposure to musical harmony significantly shapes our aesthetic responses to music.

Further research by Cowen et al. (2020) explored how Western and Chinese music evoke distinct subjective experiences, identifying 13 different types of subjective reactions in both cultures. This study indicates that while specific emotions triggered by music are consistently recognized across cultures, broader affective features such as valence and arousal are less universally experienced.

Additionally, a study by Liu et al. (2023) shows that native speakers of tonal languages excel in melodic discrimination but have a harder time with beat perception compared to speakers of non-tonal and pitch-accented languages. This suggests that linguistic background can influence specific musical abilities, further indicating that our musical experiences are not universal but rather shaped

by a combination of innate abilities and cultural influences.

Also, as mentioned in the previous section, some studies on the relationship between personal traits and musical taste have been replicated in different countries and reveal some cultural differences. While these studies have found that many patterns in music preferences are similar across different cultures, they have also revealed notable differences.

For instance, the positive relationship between openness and liking reflective music seems to be universal, while the positive correlation between energetic and rhythmic music and extraversion, which was found in western countries, was not found in Japan (Brown, 2012). The study by Delsing et al. (2008) also highlighted how the same genre can be perceived differently depending on the country.

They found that trance and techno were associated with pop music in the Netherlands, while in the U.S. it was perceived as a genre of its own. Gospel, on the other hand, was associated with elite culture in the Netherlands, while perceived as conventional in the United States. These differences can be explained by the popularity of these genres in each country : while gospel originates and remains a well-known and popular genre in the US, it is not part of the common culture in the Netherlands. The same goes with electronic music, which is quite mainstream in the Netherlands, and probably considered more niche in the United States.

Given these differences in how music can be perceived depending on the region, it is crucial not to generalize findings on musical taste based on studies conducted in one specific area. This also underscores the importance of diversifying research locations to better capture the full spectrum of musical preferences across cultures.

Preferences specific to geographic areas

As mentioned previously, for a long time, music geography studies relied on data from music distributors, as data about listeners was difficult to obtain. With the rise of the Internet, however, such data has become more accessible, whether by reaching people through surveys or directly accessing vast amounts of geolocated listening data from music streaming platforms. This shift in data collection allowed the emergence of research that specifically focuses on the musical preferences of people in various locations.

For example, Mellander et al. (2018) aimed to explore music preferences across 95 large metropolitan areas in the U.S. using an online survey, which included Rentfrow and Gosling (2003)'s Short Test of Music Preferences and the MyPersonality test, with approximately 120,000 participants. This sample size is notably large compared to previous geographic studies of music preferences. Through factor analysis, correlation, and regression analyses, the authors

examined how music preferences relate to economic, demographic, and psychological variables.

The results indicate that regions favoring sophisticated and contemporary music tend to be more affluent, educated, and liberal, whereas areas preferring unpretentious and intense music are typically less advantaged, more working-class, and conservative. However, the study focuses only on large U.S. metropolitan areas (with populations over 500,000), potentially overlooking patterns in smaller or rural areas where different dynamics may be at play. For instance, recent streaming data based research by Lee et al. (2024) suggests that both collective and individual music diversity increases with population size. It is possible that factors other than diversity also differ between densely and sparsely populated regions. Additionally, the data was collected between 2001 and 2013, which may not fully capture the impact of more recent changes in music consumption trends, especially with the rise of streaming platforms.

Another study by Way et al. (2019) analyzed music consumption patterns across different U.S. states using data from over 16 million Spotify users. By examining the most-streamed artists and genres, they uncovered significant regional differences in musical preferences.

For example, genres like ranchera and mariachi were more popular in southern states with larger Hispanic populations, while gospel and soul had a stronger presence in the southeastern states. In contrast, states with more urban populations, such as New York and California, exhibited a greater diversity of genres. While state-level musical diversity — the range of genres consumed within a state — varied significantly, the diversity of individual listeners' tastes remained similar across states. However, as already noted, Lee et al. (2024) suggests that individual diversity increases with city population size, meaning that when aggregating listeners by state, which includes both smaller and larger cities, these differences in individual diversity might be blurred.

Currently, streaming is predominantly used in Western countries, and much of this research focuses on them, or even solely on the U.S.. There are challenges in conducting this type of regional analysis, especially in smaller countries, due to difficulties in accurately geolocating mobile Internet users. A recent study by Lesota et al. (2021) has a more global approach, and explores the place of local music and U.S. music in different countries, analyzing both the number of streams and the number of artists using Last.fm data. However, as we will show in Chapter 7, there is potential bias in such studies as users of a niche service like Last.fm may not represent the broader population of a country. This calls for caution in generalizing these findings to the entire population.

The effect of globalization

With technological advancements, people around the world now have access to more or less the same music. Globalization extends beyond music — it affects

nearly every aspect of our lives. But how exactly does it influence the music industry? Do geographical differences in music preferences progressively fade away?

From the 1960s through the early 1990s, this seemed to be the case (Achterberg et al., 2011). In countries like France, Germany, and the Netherlands, American music increasingly dominated the charts, pushing local music aside. However, in the 1990s, a shift occurred, and local music began to make a strong comeback. One major explanation is the geopolitical and cultural changes following the end of the Cold War in 1989. As American influence waned, European countries experienced a resurgence of national identity, reflected in the growing popularity of local music. Some governments even passed laws to support this trend, like France's 'loi Toubon,' which requires a minimum of 40% of music with french lyrics on the radio and other media.

Additionally, shifts in the music industry and technology contributed to this change. As production costs dropped, it became easier for local artists to produce and distribute their music. MTV, which had initially promoted a more global (primarily American) music agenda, also began to focus more on locally oriented content, helping fuel the rise of national music scenes.

Achterberg et al. (2011)'s study, which analyzed data up until 2006, predates the rise of streaming platforms. Did the advent of streaming reverse the trends they observed? While U.S. music still holds a presence in many countries' charts, Bello and Garcia (2021) claim that the share of local music and overall diversity kept increasing in the charts of most countries in the last years. By analyzing large-scale datasets from Spotify and iTunes, the study examined trends in song, artist, and label diversity across 39 countries from 2017 to 2020. Their findings reveal a growing cultural divergence, as national charts have become more distinct, with more unique, and increasingly local, songs and artists populating the charts.

Both these studies focus on chart data, but do charts accurately reflect what people choose to listen to? A recent paper by Lesota et al. (2022) shifts focus to actual music consumption patterns to examine the role of local and U.S. music in different countries. The authors used a subset of the LFM-2b dataset from 2018-2019, covering 12,875 users from 20 countries, and introduced two key measures to quantify music consumption. The first measure looks at the share of streams in each country that come from local, U.S., or other foreign artists. The second examines how much of the music produced by artists from a specific country is consumed domestically versus internationally. According to the study, U.S. music holds a strong global presence, accounting for around 40% of music consumption in many countries, while local music's popularity depends on the country, ranging from 20% (in Germany) to 80% (in Brazil).

While focusing on actual consumption data is a step in the right direction, several limitations in the study make its results less conclusive. First, the sam-

ple size is relatively small, with around 500 users per country on average, and as few as 115 users in some cases (like Turkey or Japan), making it hard to generalize the findings. Additionally, the users come from Last.fm[2], a relatively niche platform, meaning the audience may not represent typical listeners in each country, including the preference for local music. Finally, the authors excluded unlabelled streams — streams from artists whose country of origin could not be identified. Since it is unclear how many artists and streams were affected by this exclusion, and considering that artists from the U.S. and other Western countries are likely better labeled, the results may be biased in favor of these regions. We will come back to this study and its limitations in Chapter 7, extending the discussion to recommendation fairness for local music recommendation.

A final concept worth mentioning is glocalization (Hebert and Rykowski, 2018). While it is relatively straightforward to estimate the amount of produced and consumed music by local artists within a country, using charts, surveys or streaming data, evaluating the full impact of globalization is far more complex, as it manifests in various ways. For instance, music trends originating from the U.S. (or other countries) often spread and are adopted by local artists worldwide. A prime example is K-pop Oh and Park (2013), a genre that incorporates classic elements of Western pop — from the music itself to its visuals and music videos — yet it was absorbed into Korean local culture and re-exported globally, including back to the U.S.. Similar trends can be observed everywhere: reggaeton beats, originally from Puerto Rico, are now used by artists across the globe; rap has been localized in nearly every country, performed in countless languages. Many genres have traveled overseas, giving birth to styles like Japanese funk or Russian rock, which borrow the original codes of the genre but eventually evolve into distinct styles of their own. These patterns are much harder to quantify, but they are essential to consider when attempting to measure the music globalization process.

3.2 Musical taste as a dynamic concept

Previously discussed literature mostly considers music preferences at a given moment. However, our musical preferences are subject to change. We can discover new music through emergent artists, fresh releases and trends, or through new environments that we find ourselves in, offline and online. We might revisit old favorites driven by nostalgia, or get bored of an album that a short time before was our favourite. Our music selections can also vary with the activity we are engaged in, the place we are at, the time of day, the season, or the company we keep. In this section, we will explore how and why musical taste can evolve, both on the long and the short term.

---

[2]Last.fm

3.2.1   Aging

In the field of psychology, many studies explore the relationship humans have with music at different ages, from early childhood to later adulthood. In their literature review on music preferences of different age listeners, Hargreaves et al. (2006) try to dig to the root by understanding how and why do aesthetic preferences form in general. They refer to two main theories: the arousal-based approach, mostly associated with Berlyne (1973), and the cognitive approach, with researchers like Martindale & Moore (Martindale and Moore, 1989; Martindale et al., 1990).

Berlyne's 'inverted-U' hypothesis suggests that people prefer stimuli, including music, with a moderate level of complexity, enjoying it most when it hits an optimal level of arousal. Hargreaves et al. (2006) apply this theory to musical taste, explaining how preferences evolve with age. Similar to developing a taste for more complex flavors in food, children start with simpler music because it's easier to process, but over time, exposure to diverse genres helps them appreciate more intricate music. Just like how our tolerance for spicy food grows, our ability to enjoy complex music expands with experience, though too much complexity can still be overwhelming.

The cognitive approach, on the other hand, focuses on how the brain categorizes and recognizes patterns in music. People prefer music that fits familiar mental categories. As we age and our musical 'library' expands, our preferences shift to match our growing cognitive understanding of genres. This approach argues that it's not just complexity but also how well music aligns with our mental expectations that determines preference.

Initially, the cognitive approach was presented as competing to the arousal theory. However, Hargreaves et al. (2006) propose to consider the two theories as complementary, pointing out that arousal potential can also influence a piece of music's typicality within its genre. For instance, the music of Stockhausen, known for its high arousal potential due to its complexity and novelty, is considered atypical for classical music, which could explain its lesser popularity compared to more traditional composers like Beethoven.

In support of these theories, Hargreaves (1984) experimentally confirms the idea that exposure plays a key role in how we develop preferences for music, particularly in relation to complexity. The study consisted in two experiments involving a total of 99 participants. In the first experiment, 59 adults from different backgrounds, including adult education students and psychology undergraduates, were exposed to two musical pieces: an 'easy-listening' track and an avant-garde jazz track. These were played three times within a single session, and participants rated their liking and familiarity of each piece after every listening. The second experiment involved 40 undergraduate students aged 18–22, who listened to three different music pieces: a pop song, a classical piece, and an avant-garde jazz track. Over three weekly sessions, each piece was played

four times per session, and participants again rated their familiarity and liking. The study found that while participants preferred simpler, familiar music initially, their liking for more complex pieces, such as avant-garde jazz, increased with repeated exposure. This suggests that repeated listening can enhance one's ability to process and enjoy more complex music, aligning with the idea that exposure plays a key role in evolving musical preferences as we age.

However, several studies support the idea that 'open-earedness' — a tendency to be more receptive to a wide range of musical styles — decreases between childhood and adolescence, possibly due to a growing understanding of cultural norms and what is considered socially acceptable.

For instance, Hargreaves (1982) aimed to explore the evaluation criteria and language used to describe aesthetic reactions in children of different ages. A total of 127 children between the ages of 7 and 15 were asked to make freely formulated statements about the differences or similarities between 9 pairs of musical pieces. The authors claimed that sensitivity to stylistic categories of music increased with age, as older children more frequently used specific genre names to describe music.

LeBlanc and Cote (1983) conducted a study on 354 fifth- and sixth-grade students, asking them to rate their preferences for 36 excerpts of traditional jazz. Fifth-graders showed significantly higher preference ratings overall compared to sixth-graders, which the authors linked to a slight decrease in openness or enthusiasm for the music with age.

LeBlanc et al. (1996) examined how music preferences for art music, traditional jazz, and rock vary across age groups in a sample of 2,262 participants ranging from 6 to 91 years old. Participants listened to 18 excerpts with similar tempos across the three genres and rated them using a five-point scale. Rock music was consistently liked across all age groups, while art music and jazz were appreciated less during middle school years but showed peaks of higher preference among younger children and college students, suggesting that 'open-earedness' declines during adolescence and partially rebounds in adulthood.

Despite their reputation and high citation rates, all of these papers have significant limitations. For example, Hargreaves (1982) did not account for whether the children appreciated the music, and the use of more genre names as they grew older could simply reflect their expanding vocabulary. Although covering various tempos and both vocal and instrumental performances, LeBlanc and Cote (1983) only considered jazz music — it seems difficult to draw conclusions about 'open-earedness' based on the appreciation of a single genre. LeBlanc et al. (1996), though it has the most robust methodology, with a large and varied participant sample and an understandable choice of music genres, still has limitations: participants were primarily from middle-class backgrounds, which may not reflect broader trends, and the number of considered genres was too small. Additionally, the number of music excerpts in all studies was limited.

More recently, Louven (2016) offered a critique of Hargreaves' 'open-earedness' and the work that followed, highlighting the lack of consensus on the term's definition. He developed the 'Osnabrück Open-Earedness Index' to empirically measure 'open-earedness,' defined as the willingness to engage with unfamiliar or disliked music. In the experiment, 961 participants, ranging from children to adults with varying levels of music education, listened to 17 diverse music samples (classical, pop, avant-garde, ethnic). Participants controlled how long they listened to each piece, and preference ratings were collected afterward. The open-earedness index was calculated by comparing listening times for disliked music with overall listening duration, with higher scores indicating greater tolerance or curiosity.

Results showed that age had no significant impact on 'open-earedness'. However, again, the range of genres was limited, participants were predominantly middle-class, and the metric used is debatable — there was no consideration of the amount of liked or disliked, for example. In summary, it is unclear from existing research whether there is a fundamental change in our perception of music as we age, and the subject requires more robust studies.

One finding however persists across different studies and even fields: musical taste seems to primarily form during adolescence. Several survey-based studies report that music plays a particularly important role in teenagers' life. Teens spend 20% of their time listening to music, compared to 13% for adults (Bonneville-Roussy et al., 2013), and consider listening to music more important than indoor activities, such as chatting with parents or reading North et al. (2000).

Music plays a key role in fulfilling adolescents' emotional needs North et al. (2000); Saarikallio (2007). Listening to music is associated with emotional expression, relieving boredom, managing stress, and helping them get through difficult times. Females, in particular, report using music more for emotional regulation, while males are more concerned with using music to create an external impression or social identity.

According to North and Hargreaves (1999), adolescents use music to form their social identity by aligning their musical preferences with their self-concept and using these preferences to communicate their values and characteristics to others. Music acts as a 'badge' that helps teens express who they are and what they stand for. Adolescents not only judge others based on the music they like but also choose musical styles that reflect their self-image. This alignment with a particular music genre can boost self-esteem and influence how they are perceived socially, such as their likelihood of having friends or being seen as successful.

Probably because of this strong connection to music during adolescence, the patterns and preferences developed in these years tend to persist into adulthood. Numerous studies show that people favor music released during their

adolescence or early adulthood. A pioneering study by Holbrook and Schindler (1989), involving 108 participants aged 16 to 86, had respondents rate their preferences for 28 popular songs from 1932 to 1986. The study found that musical preferences follow an inverted U-shaped curve, peaking when participants were around 24 years old.

A more recent study by Jakubowski et al. (2020) explored the 'reminiscence bump,' where music evokes strong autobiographical memories. Involving 470 participants (ages 18 to 82), who rated 111 popular songs from 1950 to 2015, the study found that the reminiscence bump peaked at age 14, with participants reporting stronger memories tied to music from their adolescence. Interestingly, younger participants showed a 'cascading reminiscence bump,' displaying increased liking for music from their parents' youth.

A 2019 study based on streaming data Way et al. (2019) found similar results. It explored how a listener's age predicts the age of the music they consume, showing that while trends affect listeners of all ages—about 28% of tracks streamed across age groups were recent hits—listeners tend to favor music from their adolescence, between the ages of 10 and 20. These findings align with psychological research showing that adolescence is a key period for forming musical identity, reinforcing the lasting impact of music from this stage of life.

3.2.2   Change of environment

Relocation

Changes in music preferences may not only naturally occur with aging, but could also be triggered by a big change in life, like relocation. Not a lot of research explores the impact of environmental changes on individuals' musical preferences, however, it is addressed locally in some ethnic and migration studies. Several studies show that migration often triggers a sense of nostalgia, and music plays a significant role in this experience (Khorsandi and Saarikallio, 2013). It can intensify and stimulate nostalgic memories, and is used as a tool for coping with loneliness and finding meaning in life. Music also helps migrants construct social imaginaries and express themselves (Pistrick, 2017).

A more recent study by Way et al. (2019) examines how taste evolves following relocation based on online traces. Using streaming and location data from 16M U.S. Spotify users, the research analyzes shifts in musical preferences as individuals move from one state to another. First, authors calculated similarities among the top 10,000 most-streamed artists in the U.S. and grouped them into 200 clusters, or genres. Each user was then assigned a 200-dimensional vector representing their streaming activity across these genres. To assess the influence of interstate relocation on musical taste, the study summed the streaming data in each of the 200 defined genres from listeners of each state, creating a profile of 'typical' musical preferences for each state.

Migrations were identified using location data. The initial phase of the study focused on short-term relocation effects. Summertime is the most common period to relocate in the US, so the authors chose to analyse musical preferences from March to May and from September to November for users whose main streaming location has changed between these two periods. Then, a comparison was made between pairs of people who listened to similar music from March to May, with one person who has moved and one person who has not. The results have shown no significant differences between movers and non-movers, meaning that relocation has little effect on the evolution of musical taste, at least during the first months after the relocation.

Assessing long-term relocation effects posed challenges due to the relatively recent widespread adoption of streaming services. The researchers circumvented this by examining users who spent both Christmas and Thanksgiving in the same state, different from their usual location, hypothesizing that these users likely lived in those states previously. By comparing the musical tastes of relocated users to both their original and new states, and to those of non-relocated individuals of the same age and gender, it was found that while the musical preferences of relocated individuals more closely aligned with their original state (64%), there was a slight shift towards the preferences of their new state compared to non-movers (57.5%).

The study is however limited as it only considers relocation within the US. The change in music listening habits between different states might not be significant, as there are no language barriers, for instance. Studying individuals who move to another country could potentially show more variation in music preferences. However, a lot of people who move, especially those fleeing their home countries, come from places where international streaming services aren't widely used. They often download music, use sites like YouTube or local streaming services, which makes it hard to track what they listened to before moving. But, as some streaming services are becoming more widespread, there's hope that we will be able to get such data in the near future.

Mid-term variations: travels, lockdowns

Recent study by Kim et al. (2024) suggests that temporal routine changes are also capable of influencing music listening habits. The authors focused on two specific events — travel and the COVID-19 lockdown — to explore how they impact music consumption, particularly in terms of diversity. The study analyzed over 100 million streams from Deezer, examining the listening behavior of 44,794 users from nine countries (France, Germany, United Kingdom, Brazil, Australia, Russia, South Africa, Morocco, Mexico), with sample sizes ranging from 1,000 to 10,000 users per country.

To identify when users traveled, the researchers used their geographical location, determined from the IP addresses linked to their streams. The city where

a user streamed the most during the year was designated as their 'home city.' A change in the primary streaming location was considered travel. The geographical distance between a user's home city and the cities they visited was then calculated to analyze how travel influenced their musical preferences. For the COVID-19 lockdowns, the study focused on the period from March to April 2020, when strict lockdowns were imposed globally.

Each user's musical preferences were represented as a 'taste vector' using a Song2Vec model, which mapped each song to a vector based on its context within playlists. A user's taste vector was the average of the vectors for the songs they listened to during a specific period. The researchers also computed regional and global music profiles by aggregating the vectors of users from specific regions or across all users globally.

For users who traveled, the study compared their music preferences during the travel month (or shortly after) with their preferences over the six months leading up to the travel. To isolate the impact of travel, they paired users with similar music preferences before the travel period, then compared the changes in musical tastes between those who traveled and those who did not. For the COVID-19 analysis, the study compared users' taste vectors during the lockdown months (primarily March to June 2020) with their vectors from the six months before. A higher cosine distance indicated a greater divergence from previous preferences, suggesting more musical exploration or a shift toward new genres or styles. In both cases, users' taste profiles were also compared with regional and global vectors.

The study revealed that travel was strongly associated with musical exploration. When users visited new cities or countries, their music preferences diversified, often shifting toward regional music rather than global or mainstream trends. The greater the geographical distance from their home city, the more significant the change in their musical tastes. Those who traveled further were more likely to listen to music different from what they typically consumed. Additionally, users with more non-conforming musical tastes (those whose preferences already deviated from the global average) showed even greater exploration when traveling compared to users with more mainstream preferences. During the COVID-19 lockdown, users' musical preferences also diversified significantly, similarly gravitating toward regional content as seen with travelers.

Although the study suggests that events like travel or lockdowns had long-term effects on musical preferences, this conclusion is based on data from only one month following the event. Therefore, it remains uncertain whether these changes are truly long-term, and they may be more accurately described as mid-term shifts, at least until proven otherwise. While the methodology and findings appear robust, the study would benefit from larger sample sizes (some countries were only represented by 1,000 users) and data from different sources, as Deezer users may represent a particular demographic that is not fully generalizable.

3.2.3   Short-term variations: mood, activity, context

Music is a powerful tool for mood regulation, offering various strategies that individuals use to manage their emotions. Saarikallio (2008) identified seven key strategies for mood regulation through music listening — entertainment, revival, strong sensation, diversion, discharge, mental work, and comfort — based on a large set of adolescents' self-reported data. Research on music's role in reducing anxiety shows it can lower physiological markers of stress, such as heart rate and cortisol levels (De Witte et al., 2020). Ferwerda et al. (2015) placed 359 participants in various emotional states using film clips and asked to rate different emotionally laden music pieces based on their likelihood of listening to them, considering their emotional state, and their personality traits. Individuals with higher scores in openness, extraversion, and agreeableness tended to listen to happy music when feeling sad, aiming to improve their mood, while those scoring high in neuroticism preferred music reinforcing their sadness.

Because music is effective in regulating emotions, it can be used in different contexts and for varied purposes, with preferences often shifting depending on the specific listening situation. However for a long time research presented a 'pharmaceutical' model of music consumption, considering listeners as passive recipients of musical stimuli, and did not explore their active role in choosing music based on the situational context of listening.

North and Hargreaves (1996) were one on the first who tried to understand how musical preferences shift depending on the listening context. The study involved 393 psychology undergraduates who were presented with one of 17 different hypothetical listening situations (such as 'jogging,' 'at a nightclub,' or 'in church'). Participants rated the importance of 27 musical characteristics (such as 'loud,' 'sad,' 'invigorating,' or 'relaxing') in terms of how much they would like music with these qualities in each context. Results showed that musical preferences were highly context-dependent; for example, participants preferred 'loud' and 'invigorating' music in active situations like jogging or parties, while they preferred 'relaxing' and 'quiet' music in more subdued settings like bedtime or church. A factor analysis revealed key underlying dimensions of these preferences, such as arousal, sensuality, and melancholia, suggesting that music is chosen not just to match emotions but to enhance the mood of the situation.

In contrast to the hypothetical scenarios used in North and Hargreaves (1996), where participants were asked to imagine how they would choose music in different situations, Sloboda (1999) focused on real-life behavior through self-reported data. The study used the Mass-Observation Project, where 249 respondents, mainly adults from diverse backgrounds, provided detailed descriptions of how they actually used music in their daily lives, during various activities such as driving, housework, or unwinding after a stressful day. Like North and Hargreaves (1996), he found that music preferences vary by context, with more invigorating music preferred in active situations and relaxing music in quieter

settings. However, he also showed that music is often used to transform moods, such as alleviating stress, which goes beyond simply matching the context's emotional tone. Also, while situational factors influenced music choices, individual preferences and emotional needs played a significant role, leading to variation in the type of music people chose even within the same context.

These studies were conducted at a time when music listening required more effort and planning, such as carrying CDs or tapes and accessing playback devices. However, with the popularity of mp3 players in the 2000s and, subsequently, smartphones and streaming services in the 2010s, most people in the Western world became able to access any music at any time and place, dramatically changing how music accompanies daily activities.

The 2018 study by Volokhin and Agichtein (2018), also survey-based, reflects this evolution in music consumption by demonstrating how the widespread availability of personal devices like smartphones has expanded the range of activities that can be accompanied by music. Their study shows that music is now an integral part of everyday life for activities such as commuting, working, exercising, and even eating. Unlike earlier times when people were more dependent on the music provided in public spaces—such as cafes or stores—today individuals often choose to listen to their own music via headphones or mobile speakers, giving them control over the sound environment and allowing them to tailor the music to their personal preferences or emotional needs.

Most studies, like those discussed previously, rely on self-reported data, which raises questions about the accuracy and how closely it reflects real behavior. Unfortunately, detecting individual listening contexts automatically, based on streaming behavior for example, is a difficult, if not impossible task.

A few studies have attempted to detect context automatically, likeKaminskas et al. (2013) and Cheng and Shen (2014) using GPS data from mobile phones to identify users' geographical locations and surroundings, however, their primary goal was to provide context-aware music recommendations rather than analyzing users' listening behavior in specific contexts.

The analysis of changes in IP addresses combined with timestamps associated directly with streams, as in Way et al. (2019) and Kim et al. (2024), could offer a way to infer activities like commuting or being at work or home, but determining the exact user's activity or mood remains a challenge. A mix of self-reported contexts, combined with streaming histories, could provide a more accurate picture of user behavior in different situations. However, collecting this kind of data on a large scale is logistically difficult, as it requires a substantial number of users from streaming platforms who are willing to report their exact activities over an extended period.

Another possible method would be to analyze user behavior in context-oriented playlists, where the context can often be inferred from the playlist's title (e.g., 'Work Focus' or 'Relaxation'). By examining which tracks users

engage with or skip, we could identify patterns in the music users prefer for specific activities. However, since these playlists are curated to fit a particular context, the music selection might not be diverse enough, and users are likely selecting from a pre-filtered set of tracks. This can introduce bias, as the playlist may influence user choices, limiting our understanding of their broader listening preferences in those contexts.

One situation in which behavioural data can be used to understand context-related variations in music preferences is if the context is known to be common to all users in a given time and place. An example of such context is weather. Anglada-Tort et al. (2023) conducted a large-scale study analyzing over 23,000 songs that reached the UK Top 100 charts between 1953 and 2019, combining it with weather data (e.g., daily temperature, hours of sunshine, rainfall) from the UK Meteorological Office, and music feature data (e.g., energy, valence, tempo) extracted from Spotify's API. Music features were aggregated on a monthly level and reduced to two key components representing high-arousal positive music and low-arousal negative music using Principal Component Analysis (Principal Component Analysis (PCA)). They explored potential nonlinear relationships between weather and music features using generalized additive models, controlling for confounding factors like seasonal trends. The results suggest that songs with high energy and positive emotional valence, such as upbeat and danceable tracks, are more likely to rise in popularity during warm, sunny weather, while rainy days were associated with a preference for lower-energy music. This methodology allows to reveal the existence of mood-regulation mechanisms at a population level.

3.2.4 The role of streaming

The availability of music on the internet, starting with (often illegal) downloading and later the widespread use of streaming platforms, has dramatically changed how we experience and engage with music. In the past, people purchased individual albums, and thus had to chose wisely what they wanted to pay for and listen to. By 2023, physical records accounted for only 17.8% of global recorded music revenues, and considering the persistence of illegal downloading and streaming, the actual share of physical media in overall listening is likely even smaller. Streaming, on the other hand, represented 67.3% of global music revenues in 2023.

With a streaming platform account, users gain immediate access to vast music catalogs of tens of millions of tracks at no extra cost. To assist listeners in navigating these enormous libraries, platforms use various recommendation methods, including editorial (similar to traditional radio) and the novel algorithmic recommendations. These technological advances can not only shift consumption patterns but also have the potential to shape our music preferences over time.

The transition to streaming services

Datta et al. (2018) was one of the first to investigate how the adoption of music streaming services, impacts individual consumption patterns. They aimed to understand if streaming generates additional music consumption, affects the variety of music consumed, and influences the discovery of new music. The data, collected from a third-party service tracking platform choices and listening behavior across various platforms, included over 123 million plays for 4,033 users over a 2.5-year period, encompassing both streaming and ownership-based consumption.

First, users who began using Spotify during the study period were identified and labeled as 'adopters.' To isolate the effect of the adoption of streaming platforms, these users were compared to 'control' users who did not adopt streaming services. The identification was based on users' platform usage, ensuring that adopters were recognized only if they demonstrated a clear shift to Spotify after a period of non-use. To account for pre-existing differences between adopters and non-adopters, the study employed a matching procedure based on a range of observed characteristics, including demographic information and pre-adoption music consumption behaviors. The two groups were compared using a difference-in-differences approach, attributing changes in music consumption behaviors post-adoption to the effect of streaming. Authors analysed both user-level fixed effects, which account for time-invariant individual characteristics, and time-varying effects that might influence music consumption trends generally.

The study measured a user's music consumption as the number of songs they played across all platforms, including both streaming services like Spotify and ownership-based platforms such as iTunes and locally stored media players. To ensure meaningful consumption, the authors excluded any song that was played for less than 30 seconds or skipped before reaching halfway through the track. They observed a 49% increase in overall music consumption across all platforms, on average per user, six months after the adoption of streaming services. It also led to consumers expanding their listening across a larger variety of music, for example number of unique genres increased by 43%. Additionally, streaming facilitated the discovery of new music, with users encountering an average of 27 new artists per month. Despite a general trend towards less repeat consumption for newly discovered music, there was more engagement with users' top discoveries, indicating valuable exploration facilitated by streaming.

Also, streaming seems to have changed the relationship people have with different music items. In a 2020 commercial study made by Deezer [3], 54% of 8000 respondents declared listening to less albums than 5-10 years ago. Instead, 40% of people preferred playlists. This preliminary study offers an interesting

---

[3]Deezer study on albums consumption.

insight, however, it would be valuable to confirm these findings through actual consumption data.

Different use cases of streaming services

All users may not have the same ways to use streaming services, and thus can be affected differently by them. Villermet et al. (2021) categorize users from Deezer into four distinct groups based on their predominant method of discovering and consuming music on streaming platforms. These methods include algorithmically generated playlists, editorially curated playlists, and organic discovery through direct searches or previously liked songs.

The categories and their proportions are as follows: rather 'Algorithmic' users represent 11% of the sample, showing a preference for automated recommendations; 'Editorial' users, who prefer human-curated playlists, make up 8%; 'Organic' users, who mostly use search functions and browse their existing collections, account for 19%; and 'Very Organic' users, who heavily rely on manual selection and browsing, form the majority at 62%. This distribution highlights a continued preference for traditional, organic methods of music consumption among the majority of users.

Notable differences were found in the diversity of music consumption among the four user groups. Algorithmic users often experienced lower song dispersion and predominantly listened to niche artists, suggesting a push towards more unique content by algorithms. In contrast, editorial users showed similar song dispersion but favored popular artists, reflecting a human curatorial bias towards well-known music. Organic users displayed moderate diversity in both song dispersion and artist popularity, balancing mainstream and less-known choices. Very organic users exhibited the highest diversity, exploring a wide array of both popular and niche tracks, indicating their use of the platform as a broad digital library.

It is important to note that the study finds correlations between users' preferred streaming modes and their music consumption patterns, but does not establish causalities — for example, users who frequently rely on algorithmic recommendations tend to listen to less popular artists, though it is unclear if the recommendations shape their preferences or reflect existing tendencies.

Algorithmic recommendation

Internet services provide access to an unprecedented volume of data, and to navigate this vast information landscape, users can rely on RS. However, while these systems aim to personalize content to our tastes, they can also lead to the creation of so-called filter bubbles. (Pariser, 2011; Nguyen et al., 2014; Haim et al., 2018). These bubbles occur when the RS filter information to such an extent that users are predominantly exposed to content that aligns with their

existing preferences, thereby limiting exposure to new and diverse types of information.

Anderson et al. (2020) focus on music algorithmic recommendations, investigating whether they narrow down user preferences into filter bubbles or if they encourage a broader exploration of musical content.

Authors use a dataset collected directly from Spotify, including the listening history of over 100 million users and their interactions with millions of songs in a one-year time-frame. Musical diversity was quantified using song embeddings that capture the similarity between songs based on user listening patterns. These embeddings enabled the researchers to assign a diversity score to each user, distinguishing between organic and algorithmically-driven listening. The study employed statistical analyses to compare these diversity scores and their relation to user engagement metrics. Furthermore, a randomized experiment was conducted to assess the effectiveness of different song ranking algorithms on user satisfaction, measured through song streams and skips.

The study found that algorithm-driven listening is typically less diverse compared to organic listening. This suggests that while algorithms are good at suggesting tracks similar to past user behavior, they might limit exposure to a broader array of music styles and genres. Generalists (users with broad and diverse music preferences) and specialists (users who prefer a narrower selection of music) respond differently to algorithmic recommendations. Specialists benefit more from relevance-based recommendations because these recommendations align closely with their existing preferences. In contrast, generalists do not benefit as much from relevance-based algorithms and may need more diverse recommendation strategies that encourage exploration beyond their past behaviors. Over time, users who increase their diversity in music listening tend to shift away from algorithmic recommendations and lean more towards organic methods of discovering music, such as direct searches or choosing from personally curated playlists. This shift implies that users seeking more variety in their listening habits may find algorithmic recommendations too restrictive.

However, Villermet et al. (2021) raise the question of whether the reduced diversity of content consumption, which contributes to confinement and 'bubble' dynamics, is actually caused by RS, or if it originates from users' pre-existing preferences or their online activities.

Apart from general diversity, the field of algorithmic fairness investigates and quantifies biases that may be produced by RS. In the context of music, one of the most extensively studied biases is popularity bias (Celma and Cano, 2008; Kowald et al., 2020). This bias is intrinsically linked to the long-tail distribution of music data, where a small number of tracks accumulate most of the plays, while a large number of other tracks are seldom heard. This tendency can skew recommendations towards already popular tracks, reinforcing their visibility and neglecting lesser-known music. Other biases in MRS include the preference for

U.S. over local music (Lesota et al., 2021), biases related to the gender of artists (Shakespeare et al., 2020), and biases that may arise from the user's side, such as the gender (Lesota et al., 2021) or personality of the listeners (Melchiorre et al., 2020). We will dive deeper in the subject algorithmic fairness in Chapter 7.

3.3 Conclusion

Understanding and modeling musical taste, its origins, patterns and variations, is essential for both the advancement of social sciences and the development of more personalized and effective MRS. Musical preferences are not just personal or aesthetic choices — they are shaped by a mix of social, psychological, and cultural factors. For researchers in social sciences, investigating these patterns offers a window into human behavior, revealing how social class, personal traits or cultural background influence what we listen to. From this perspective, musical taste becomes a valuable lens through which we can better understand broader questions about social structures, individual identity, and cultural norms.

At the same time, for streaming platforms, understanding musical preferences is crucial for improving user experiences. Music recommendations are central to how streaming platforms engage users, and the better these systems can capture and anticipate the nuances of musical taste, the more personalized and satisfying the user experience will be. However, as musical taste is such a subjective and dynamic concept, quantifying it remains challenging, and despite the vast amounts of data available through streaming services, it is not always easy to translate it into meaningful insights about individual taste.

To this day, there is a big gap between research in social sciences and the field of MRS. Although sociologists, psychologists, musicologists and even geographers have conducted extensive research on musical taste and its many influences, their findings have rarely informed the development of RS. Conversely, the algorithms powering these systems often overlook the rich theoretical frameworks developed in social sciences. Only recently have some researchers begun to bridge this gap, recognizing the potential for cross-disciplinary collaboration, like for example Laplante (2014)'s essay on how insights from psychology could improve music recommendation. Our own RECORDS project is another step in this direction, bringing together researchers from both social science and computer science to leverage streaming data in understanding music preferences and enhancing RS.

This chapter has aimed to contribute to bridging this gap. By reviewing the different aspects that shape musical taste, such as social background, personality traits, cultural environment, and diverse life events, it seeks to better understand the broader human behaviors that contribute to the formation of our

music preferences. Equally important, we have explored the methods available for quantifying and evaluating these preferences and the ways in which they evolve. Traditionally, surveys and interviews were used to this end, but these methods, even though they have their advantages, were limited by their reliance on usually small, specific populations and the declarative, thus possibly biased nature of the data. The rise of the Internet and streaming platforms has opened the door to massive collections of behavioral data, offering new opportunities to study music consumption. Yet, with these new data sources came new challenges — such as interpreting behavioral signals and managing large datasets — that require new approaches.

In Chapter 6, we will come back to the subject of musical taste, exploring how individual and shared patterns in music preferences can be represented through computational methods on streaming data, both for research and recommendation purposes.

To conclude, we summarized, in Table 3.2, the key studies discussed in this chapter, demonstrating the evolution of data and methodologies used in research on musical taste from different fields.

Discipline	Type of data	Paper/Book	Data	Population	Methods
Sociology	Declarative (survey)	Bourdieu (1984)	Identifying and rating of music pieces titles	1,909 French individuals (1963–1968)	Descriptive analysis, social class stratification
Sociology	Declarative (survey)	Peterson (1992)	Rating 19 music genres	18,775 U.S. adults (1992)	Log-multiplicative model, regression analysis
Psychology	Declarative (survey)	Rentfrow and Gosling (2003)	Rating 14 music genres + BFI	1,704 American university students (2001)	Factor analysis, correlation analysis
Psychology	Declarative (survey + online likes)	Nave et al. (2018)	Ratings of 25 15-sec music excerpts + artists Facebook 'likes'	22,252 users from 153 countries (2012–2015)	LASSO regression, Pearson's correlation
Psychology	Behavioral (streaming data)	Anderson et al. (2021)	3 months of Spotify activity logs	5,808 U.S. Spotify users (2021)	Elastic net regression, random forest, demographic analysis
Cultural Geography	Declarative (online survey)	Mellander et al. (2018)	Rating 14 music genres + MyPersonality test	120,000 U.S. participants (2001–2013)	Factor analysis, correlation, regression analysis
Cultural Geography	Behavioral (streaming data)	Way et al. (2019)	2 years of geolocated streaming data	16M Spotify users in the U.S. (2017–2019)	Genre clustering, cosine distance analysis
Music Information Retrieval (MIR)	Behavioral (streaming data)	Lesota et al. (2021)	3 months of streaming data + artists' origin country	12,875 users from 20 countries (2018–2019)	Stream share analysis, geographic artist origin analysis

Social Sciences	Behavioral (streaming data)	Kim et al. (2024)	1 year of geolocated streaming data	44,794 Deezer users across 9 countries (2019–2020)	Song2Vec model, cosine distance, geographic distance analysis
Psychology	Radio charts	Anglada-Tort et al. (2023)	UK chart data from 1953–2019	—	PCA, generalized additive models, weather correlation
Marketing	Behavioral (streaming + ownership data)	Datta et al. (2018)	2.5 years of streaming + ownership-based data	4,033 users, 123M plays (2015–2017)	Difference-in-differences, matching procedure
Psychology	Behavioral (streaming data)	Anderson et al. (2020)	Spotify listening history: song embeddings and diversity scores over 1 year	100M Spotify users (2018)	Song embeddings, diversity scores, randomized experiment
MRS	Behavioral (streaming data)	Villermet et al. (2021)	1 year of contextualized streaming data	9,000 Deezer users from France (2020)	Cluster analysis, song dispersion, listening pattern analysis

Table 3.2: Overview of data and methodologies used in research on musical taste from different fields.

Chapter 4

Representing, categorising, and labelling the musical space

When studying musical taste and music consumption, researchers usually have to make the choice to represent music in a certain way. As we have seen in the Chapter 3, most sociology and psychology studies on musical taste either directly manipulate names of musical pieces, like songs, albums, artists or composers (Bourdieu, 1984; Nave et al., 2018), or use categories, like genres or moods (Peterson, 1992; Bryson, 1996; Rentfrow and Gosling, 2003).

Recommender systems often represent music in a continuous, multidimensional space, where each song is represented as an embedding — a vector that captures latent features of these items. These embeddings can be derived from the music's inherent features (van den Oord et al., 2013) or from user-item interactions, where songs frequently shared in streaming histories by users are considered similar (Shakirova, 2017; Sánchez-Moreno et al., 2016). Several psychological studies also aggregate music by audience preferences (Rentfrow and Gosling, 2003; Rentfrow et al., 2012).

The choice of a given music representation in a study can have a big impact on its outcome, and can be considered as a research question on its own. Actors from different academic domains, as well as industry, take specific interest in ways to represent and segment the musical space.

For **musicologists**, categorizing music is essential for the systematic study of its historical and cultural significance. It aims to provide a structured framework for analyzing musical influences and trends over time, supporting scholarly research and contributing to the preservation of cultural heritage. These taxonomies also shape the way music is understood, for better or for worse: for example it can confine some of it in problematic and stigmatizing categories, like 'world music' (Feld, 2000) or 'urban music' (Forman, 2002), minimizing its actual diversity.

For **music creators**, categorization facilitates the marketing of their music.

Understanding their niche helps in targeting specific audiences more effectively and can guide the development of new musical styles by positioning themselves between existing genres.

For **music distributors**, such as music labels and streaming platforms, music classification helps in organizing large music libraries, making it easier to deliver music to audiences. Specifically for streaming platforms, a well-labeled catalog allows users to search for music more efficiently and enables the creation of customized editorial or algorithmic playlists and radio stations that match individual listeners' tastes, thereby increasing user engagement and satisfaction.

Researchers in **music information retrieval (MIR)** look for ways to automate music classification, usually to label huge datasets for research, or to serve previously mentioned actors from the industry, like streaming platforms or music labels.

Finally, some **psychologists** also propose their own ways to classify music in order to study correlation between music preferences and personal traits.

Several authors have summarized existing approaches (Kaminskas and Ricci, 2012; Knees and Schedl, 2013) from previous studies. In this chapter, we aim to make an overview of our own, that spans between several research domains, and discuss the advantages and limitations of different ways to represent and dissect the musical space. In the first section, we will explore ways to label music based on specific features, like audio or lyrics, from human annotations to automatic classifications. In the second part, we will see how the musical space can be built through taste aggregation, from psychological surveys to collaborative filtering (CF) for RS.

4.1 Using music features

An obvious approach to classify and label music involves using the knowledge we have about the its musical properties, like rhythmic patterns, instruments, specific techniques or scales. Additionally, classification can be based on textual analysis of the lyrics or insights into the artist's background, considering how their music connects with historical or cultural influences. In this section, we will explore the various features that can be used to determine a song's genre, mood, and geographical or cultural origin, whether through human analysis or in an automated setup.

4.1.1 Human annotations

Human experts are valued for their knowledge and contextual understanding of one or several music genres, based on extensive experience and often academic or professional background in musicology.

The most commonly used music taxonomy, by different experts and people in general, are music genres. Categorizing music into genres is a seemingly trivial exercise, that turns out to be a daunting task. In 1982, Fabbri (1982) defines musical genre as *"a set of musical events (real or possible) whose course is governed by a definite set of socially accepted rules"*.

The same year, Tagg (1982) proposes a set of precise characteristics to distinguish folk, art and popular music. In 1996, Frith (1996) states that *"popular music genres are constructed — and must be understood — within a commercial/cultural process, they are not the result of detached academic analyses or formal musicological histories"*. And indeed, today, the music industry is the most interested in genre classification.

Pachet et al. (2000) mention the following actors of the music industry who resort to music classification :

- Record Company Catalogs (Universal, Sony Music, EMI, BMG): These companies have vast archives and their classification often sets industry standards.

- Record Shops and Megastores (Virgin Megastore, Tower Records, Fnac): These retail environments categorize music to facilitate consumer browsing, influenced by both market trends and historical data.

- Music Charts (Billboard, Top 50, Cashbox): These organizations categorize music based on sales, radio play, and streaming statistics, impacting genre perceptions through popularity metrics.

- Musical Websites and Online Record Shops (Amazon, All Music, Sonic-Net, Mzz, Listen, Netbeat): These platforms provide digital categorization, often using detailed metadata and user-generated tags.

- Specialized Press and Books: Publications dedicated to music critique and history, which often provide in-depth analysis and genre classification based on stylistic and historical research.

- Specialized Web Radios: Internet radio stations that often focus on specific genres, offering curated playlists that reflect nuanced understanding of genre distinctions.

- Online collaborative databases (Wikipedia, Discogs): Individuals who contribute to collective knowledge platforms, offering categorizations based on community consensus and personal expertise.

Such a vast amount of different parties leads to a lack of consensus on the taxonomy to use. In the same article, authors compare 3 Internet genre taxonomies: allmusic.com, amazon.com and mp3.com. The first problem they observe is that the three sources propose a different number of existing genres : 531, 719 and 430 respectively. This is primarily due to the fact that genres often lack clear, universally agreed-upon definitions.

Today, genres are often represented by a hierarchical tree structure (Figure

4.1), going from macro-genres (rock, jazz, classical etc.) to micro-genres (chill-wave, seapunk, lo-fi house etc.) (Raimond and Sandler, 2012; Tzanetakis and Cook, 2002; Anderson et al., 2021; Way et al., 2019). Even macro-genres like 'rock' or 'pop' do not denote the same set of songs in different taxonomies, let alone more specific sub-genres. This hierarchical system raises the question of how deep the classification should go or at what point a genre should be distinctly recognized.

Additionally, the contemporary music scene is characterized by an omnipresent mixing of genres. On one side, niche artists are creating unique styles by merging multiple genres, often giving them very specific names. For example, today, the Every Noise at Once website[1] — an initiative of Glenn McDonald, former Spotify engineer — counts 1300 genres. On the other side, mainstream music frequently blends various genre influences but is often broadly categorized under the label 'pop'. It is also common for multiple genres to be assigned to a single album, song, or artist, as seen in platforms like Wikipedia and other music tags' sources.

In response to this lack of structure, Yves et al. (2007) created the Music Ontology, a framework designed for structuring and publishing music-related data on the web. It consists of a comprehensive set of classes and properties to represent information about musical works, their performances, and recordings, integrating with other ontologies for enhanced richness and flexibility. Later in a separate study, they demonstrated the framework's effectiveness using a unique, query-driven evaluation methodology (Raimond and Sandler, 2012).

This method involves aggregating a large set of real-world, music-related user queries and evaluating how well these can be expressed within the Music Ontology framework. The key measure used, termed 'ontology fit', quantitatively assessed how the ontology could represent the features found in user queries, thereby serving as an indicator of its practical utility in real-life applications. The ontology's design is particularly aimed at addressing real-world user needs in the music domain, as evidenced by its evaluation against a dataset of user queries. Despite its robust design and practical utility (it was used on BBC Music website and the DBTune project, for example), the Music Ontology still struggles with subjective, emotional, cultural and contextual specific descriptions of music.

Beyond the inconsistency of taxonomies, the demographic composition of human annotators presents a deeper challenge. For example, Glott et al. (2010) show that most Wikipedia contributors are male, from Western countries, and with high education level. We can assume that a similar demographic distribution exists among other music annotation actors. This skew can influence genre recognition and appreciation. For example, Western music genres tend to be described and categorized with greater precision and nuance, while non-Western
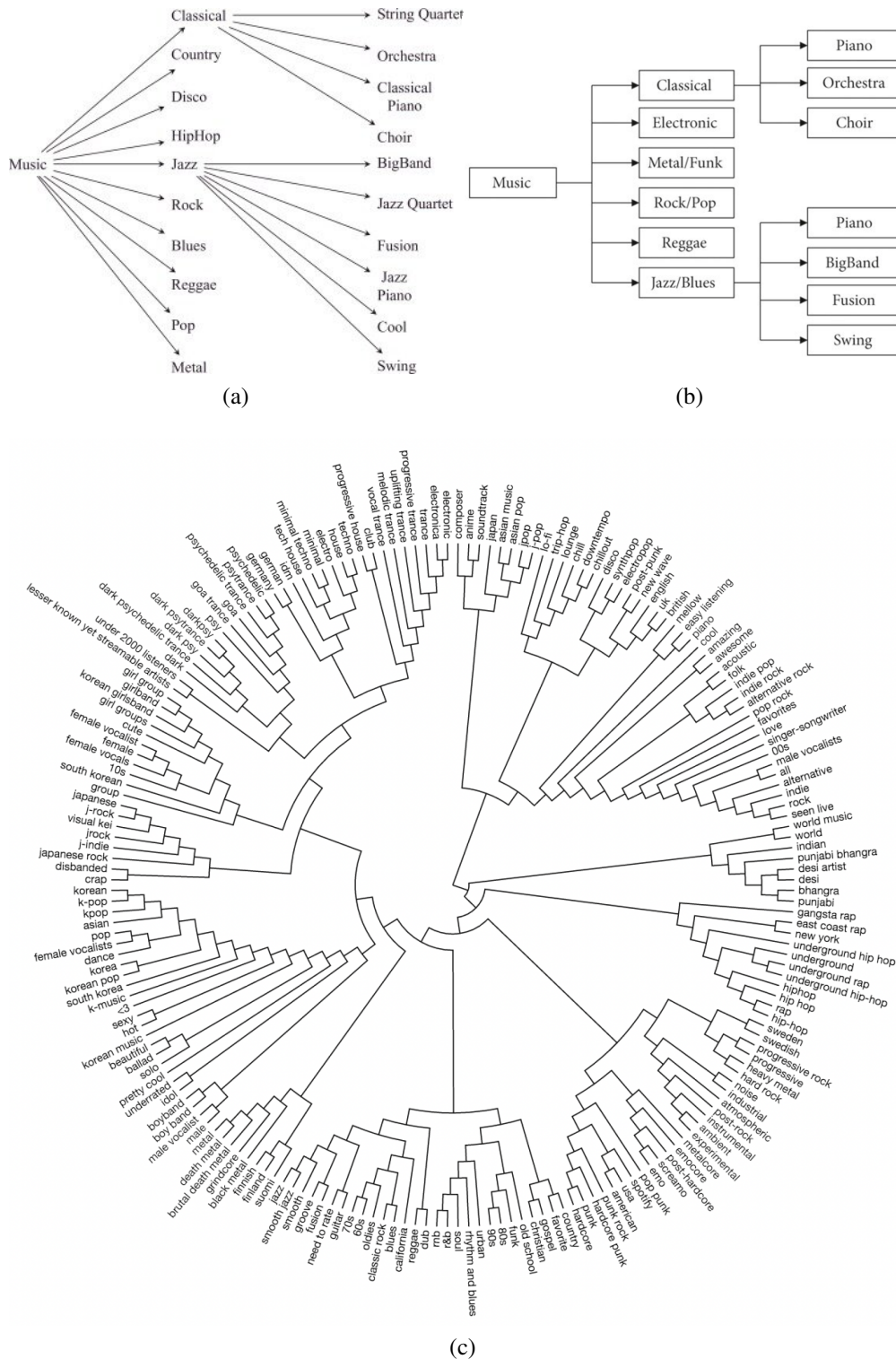
---

[1] Every Noise at Once

Figure 4.1: Examples of music classification in (a) Tzanetakis and Cook (2002), (b) Zhang (2021), and (c) Wallace (2015).

music might not be as accurately or comprehensively labeled.

These biases are further exacerbated by cultural differences in annotators. Epure et al. (2020) explore the challenge of accurately annotating music genres across different cultures, particularly those defined by language boundaries. The study combines language-specific semantic representations with several music genre ontologies, creating a cross-lingual, culture-specific music annotation system. Through this workflow, authors found that music genres are not consistently defined across languages. For example, a genre well-recognized in one language might not even have an exact equivalent in another, or it might be classified differently. Music genres often carried different connotations in different cultures, even if the same terms were used. There were also some disparities in the amount and detail of genre annotations across languages, often influenced by the popularity of certain music styles within those linguistic communities and the volume of content available on platforms like Wikipedia in those languages.

Related to these cultural inconsistencies is the difficulty in appropriately identifying and labeling an artist's country, which is another type of label that is often used to classify music. It raises the question of how this country should be chosen : based on where the artist was born, where they perform most often, or where they are most popular? Here again it often occurs that artists who are not from first-world countries are poorly labeled, and consequently have less visibility on different platforms.

Even when annotations are available, they tend to focus overwhelmingly on popular music, leaving a vast majority of less mainstream music from the 'long-tail' unannotated. This creates a vicious cycle: less popular music is poorly labeled or ignored altogether, making it harder for people to discover it, which in turn reduces its visibility and chances of being properly annotated. The problem is, most of the popular music is from Western countries, disproportionately made by males[2], and usually fits within the most mainstream genres. Therefore, this self-perpetuating cycle disproportionately affects music from niche genres, underrepresented regions and countries, as well as non-male artists.

On top of that, human-based classification systems present obvious scalability issues: manual classification is laborious and difficult to perform on larger datasets, especially with the continuous growth of music production and distribution. To address this challenge, Prockup et al. (2015) proposed a model to scale up human annotations from the Music Genome Project (Castelluccio, 2006) for predicting musical genres. They trained a machine learning model, based on logistic regression, using rhythm and timbre attributes labeled by experts, achieving an impressive average Area Under the Curve (AUC) (area under the curve) of 0.918. However, distinguishing between certain overlapping subgenres (e.g., Dance vs. Trance, or Hard Rock vs. Punk Rock) remained

---

[2]Between 2012 and 2017, only 22.4% of performers, and 12.3% of songwriters across 600 of the most popular songs were female. Smith et al. (2021)

challenging. In another model, they combined human annotations with audio features derived directly from the music signals, making the model even more effective. This leads us into the next section on the use of audio-based methods for scalable music labeling.

4.1.2   Automatic classification

Human annotations are highly impractical for large datasets, such as those found on streaming platforms for instance. Researchers in MIR explore ways to automate music labelling using audio features analysis. For music streaming platforms, this approach can be particularly valuable in 'cold-start' situations, where a track is new to the platform and has not been played sufficiently to gather enough user data to use for recommendation. By examining the intrinsic properties of the music itself — such as tempo, rhythm, and harmony — it becomes possible to gather information about a track, that can then be used to make preliminary recommendations and categorize music effectively even without extensive user interaction data. Also, lyrics analysis can be helpful in some cases, for example to detect the mood, or the language of a song.

Genre

Several studies have aimed to automate the classification of musical genres using audio signals to enhance music information retrieval systems, thereby reducing or eliminating the need for manual genre annotation.

Tzanetakis and Cook (2002) was a pioneering study in the field. Authors used 1000 tracks from various sources, including compact discs, radio, and online platforms, representing a total of 10 distinct music genres. They represented each track by a vector that combined timbral texture, rhythmic content, and pitch content features, all extracted from audio signals. These vectors were then used to train various classifiers, such as Gaussian mixture models (Gaussian Mixture Models (GMM)s) and k-nearest neighbors (k-nearest neighbors (k-NN)). The classification system they developed achieved an accuracy of 61% across the ten genres, a rate comparable to human performance in similar genre classification tasks Perrot (1999). Their study also highlighted the fuzzy nature of genre boundaries as a significant challenge, noting that the diversity within genres sometimes leads to misclassifications. For example, the genre of rock was identified as having particularly broad and overlapping characteristics with other genres, illustrating the complexity of accurately categorizing musical genres using automated systems.

Audio recording is about capturing the sound of the actual performance. Musical Instrument Digital Interface (MIDI) (Musical Instrument Digital Interface) recording or 'sequencing' is about capturing the actual notes of the performance. McKay and Fujinaga (2004) used MIDI recordings to extract features

like instrumentation, texture, rhythm, dynamics, pitch statistics, melody and chords. Neural network and k-NN based classifications were performed hierarchically, using different sets of features at different levels of the hierarchy, reaching a 90% classification success rate for 1049 tracks of 9 genres.

Cataltepe et al. (2007) experimented with different combinations of MIDI and audio data, testing various segments of music files and audio quality settings to see how these factors influenced genre classification. Their findings suggested that a hybrid approach using both MIDI and audio could enhance the performance of music genre classification systems, however, they do not outperform the results of McKay and Fujinaga (2004).

Baniya et al. (2014) address the complexities of automatic music genre classification within large datasets. Their work focuses on refining the classification process through improved audio feature extraction and classifier design. They explore a diverse set of audio features categorized into dynamics, rhythm, spectral, and harmony, applying advanced statistical techniques to distill these into a compact, informative representation. Their method involves evaluating each feature's effectiveness using the Minimum Redundancy Maximum Relevance (Minimum Redundancy Maximum Relevance (MRMR)) algorithm, which ranks features based on their importance for genre classification. The researchers use Support Vector Machine (Support Vector Machine (SVM)) classifiers to assess the efficacy of their feature selection method, finding that the MRMR-based feature selection outperforms PCA in terms of classification accuracy.

Recent studies on automatic genre classification focus on big datasets extracted from streaming platforms, and mostly use deep learning techniques. Bahuleyan (2018) use 10-second sound clips extracted from YouTube music videos of 7 different genres, on which they employ two distinct approaches. The first is a deep learning model using convolutional neural networks (Convolutional Neural Network (CNN)) trained on spectrograms of audio signals. The second involves traditional machine learning classifiers trained on hand-crafted audio features. The combination of the two approaches produced an AUC value of 0.894.

Over the years, music genre detection has advanced, using larger and more diverse datasets encompassing a broader range of genres. Despite these improvements, significant limitations persist. One common issue is that the music excerpts used for training models are typically short, potentially missing style variations that occur throughout a track. Additionally, most studies still rely on a taxonomy of ten genres or fewer, often choosing genres that are easily distinguishable from one another to simplify the classification task. This approach will not work in a real-life scenario, for example on the catalog of any steaming service. Finally, the inherent difficulties that humans face in consistently identifying music genres translate into similar ambiguities for automatic genre

classification.

Mood

Humans naturally associate specific music features with particular emotions or contexts, often independently of the genre. One effective way to categorize music is by identifying its 'mood.' This approach focuses on the emotional impact and the atmosphere conveyed by a piece, allowing for a categorization that resonates with the listener's experience regardless of musical style or structure.

Lu et al. (2005) is a pioneering study in predicting mood from audio features, using supervised learning. They used a dataset of 800 music clips selected for their representative emotional expressions, extracted from about 250 pieces of music, primarily from the classical and romantic periods.

The corpus was labeled prior to machine learning analysis using a detailed process involving three experts. These experts, knowledgeable in music theory and psychology, worked together to annotate the clips, categorizing them into four primary mood clusters: Contentment, Depression, Exuberance, and Anxious/Frantic. To ensure the accuracy of the mood labels, any music clip that did not receive unanimous agreement among the experts was excluded from the dataset. This method minimized subjectivity and ensured that each selected clip clearly represented its assigned mood, providing a strong foundation for training the hierarchical mood detection system effectively.

Three primary types of audio features are analyzed to classify music according to its mood: intensity, timbre, and rhythm. Intensity features focus on the energy across different frequency subbands of the music, reflecting the dynamic range and overall loudness that can indicate the emotional intensity of a piece. Timbre features capture the quality of the sound that distinguishes different types of sounds and instruments, independent of pitch and loudness; these include spectral shape characteristics like brightness and spectral contrast, which help in identifying the unique color or tone quality of the music. Lastly, rhythm features assess aspects such as rhythm strength, regularity, and tempo, which are crucial for understanding how the timing and pace of music convey mood.

A hierarchical classification framework is used to initially group music clips into broad mood categories using intensity features, which capture the energy and loudness of the music. This prepares the ground for a more detailed classification within these groups using additional timbre and rhythm features. To model the complex relationships between audio features and mood categories, GMMs are employed, using the expectation maximization algorithm to efficiently estimate the parameters of Gaussian components. Additionally, k-means clustering is applied at the beginning of the GMM process to help set initial parameters, enhancing the performance and convergence of the expectation maximization algorithm in fitting the model accurately to the data. Together,

these techniques create a robust system for recognizing and categorizing the emotional nuances in music, achieving an average mood detection accuracy of 86.3%.

The study set a strong baseline for further research in the field. However limited in terms of music diversity, focusing exclusively on classical and romantic music. Also they didn't keep the music pieces that were mood-ambigious, which is far from a real-life scenario, on music streaming platforms for instance.

13 years later, Delbouys et al. (2018) propose a sophisticated approach to predict music moods, using neural networks on a substantial and diverse dataset. This dataset included 18,000 tracks sourced from the Million Song Dataset and the Deezer catalog, notable for its size and diversity, marking it as one of the largest datasets used for multimodal mood detection to date.

The labeling was derived from the Million Song Dataset, which includes various descriptive tags, including those related to mood. These tags were converted into a two-dimensional space of valence (from positive to negative mood) and arousal (from calm to energetic mood), based on a psychological model by Russell (1980) commonly used in music information retrieval. To achieve this two-dimensional embedding, a database by Warriner et al. (2013) was used, which provided valence and arousal scores for thousands of English words. For tracks with multiple tags, the mean of these scores was calculated to assign a consistent mood descriptor to each track, allowing efficient labeling of the large dataset without the need for direct human annotation.

Authors considered both audio and lyrics as features that can determine the mood of a song. First, separate models were trained for audio and lyrics data. These models learned to predict mood from each type of data independently. Various deep learning architectures were employed, including CNNs for audio and different configurations of recurrent neural networks like LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Cho et al., 2014) for lyrics. After training unimodal models, a multimodal approach was used where the outputs of the audio and lyrics models were combined. This combination was done through fusion techniques such as mid-level and late fusion, aiming to leverage the strengths of both modalities. The fusion of modalities resulted in more accurate mood predictions compared to models using either audio or lyrics alone. The study found that audio models were particularly strong in detecting arousal, which relates to the energy and intensity of the music. Lyrics, while useful for detecting arousal to some degree, proved especially valuable for enhancing valence predictions, which relate to the emotional positivity or negativity conveyed by a track.

The innovative aspect of this study is shown by its application of deep learning techniques to a large and varied dataset, effectively enabling mood predictions across diverse music styles. Recently, a similar method — training CNNs on audio-based song embeddings — has been implemented for mood-specific

music recommendations in the Deezer app (Bontempelli et al., 2022). In this case they take in account the fact that the mood has to fit the overall user's preferences.

Overall, it seems like detecting the global mood of a song somewhat more realistic than to determine a specific genre. However, a common limitation to most papers on the topic is the cultural specificity in the perception of music. Research shows that the emotions attributed to a music piece can vary depending on the country and culture of the listener (Brown, 2012; Laplante, 2014; Cowen et al., 2020). Bhat et al. (2014) addresses this concern, aiming to classify both Western music and Hindi film music. The paper highlights that the threshold values for classifying moods based on extracted audio features such as intensity, timbre, pitch, and rhythm vary between Western and Hindi music. This suggests that cultural differences in music composition and listening practices might influence the perception of mood. For instance, what constitutes a 'happy' or 'sad' song in Western music might have different acoustic profiles compared to Hindi music, reflecting divergent cultural expressions of emotion through music.

Language and country

Genre and mood are not the only properties that can be used to segment a music catalog. In some cases, like for example to make country specific recommendations, determining the language or the artist's country can be relevant as well.

When lyrics are available in written form, natural language processing (NLP) techniques can be employed to detect the language Mahedero et al. (2005). If the lyrics are not available in text, they may first need to be transcribed from audio Gao et al. (2021).

Lyrics language can be used as a proxy to determine the artist's country, which is a frequently missing piece of metadata in big music catalogs, whether it's the artist's country of origin, their primary area of activity, or their main market. Although some languages are spoken in multiple countries, and artists might use languages like English to reach a global audience, the language of the lyrics can still serve as a valuable clue, especially when other data sources are unavailable.

However, it's important to note that NLP and transcription techniques tend to be more efficient for widely spoken languages, which could inadvertently reinforce existing biases. Artists from non-English speaking countries or those singing in less common languages often have smaller fan bases and lack detailed metadata. These artists, especially local ones from regions with less widely spoken languages, are typically the hardest hit by annotation gaps and also the most challenging to support with automated metadata completion techniques.

4.2 Using taste aggregation

4.2.1 Declarative data

A series of studies was made, mostly by psychologists, in order to classify music based on people's preferences and the emotions they get from listening to different kinds of music.

A significant early work by Rentfrow and Gosling (2003) involved six distinct studies that explored common beliefs about music, the underlying structure of music preferences, and the links between music tastes and personality traits. One of these experiments aims to categorize music based on people's musical taste, without any pre-existing theories or expectations about the number of dimensions or the nature of the underlying structure.

1,704 undergraduates from the University of Texas at Austin, were asked to complete a survey comprising 14 music genres: alternative, blues, classical, country, electronica/dance, folk, heavy metal, rap/hip-hop, jazz, pop, religious, rock, soul/funk, and soundtracks. Participants rated their preference for each genre on a 7-point Likert-type scale, with endpoints at 1 (Not at all) and 7 (A great deal). Principal-components analysis of the answers led to a four-factor solution that accounted for 59% of the total variance, each encompassing different genres of music:

- Reflective and Complex: This dimension includes genres like blues, jazz, classical, and folk music, which are known for facilitating introspection and are structurally complex.

- Intense and Rebellious: Rock, alternative, and heavy metal. These genres are characterized by high energy and often emphasize themes of rebellion.

- Upbeat and Conventional: Country, soundtrack, religious, and pop music. These genres tend to emphasize positive emotions and are structurally simple.

- Energetic and Rhythmic: Includes genres that are lively and rhythm-focused, often featuring rap/hip-hop, soul/funk, and electronic/dance music.

Several methods were used to to choose the appropriate number of factors to retain: scree test, Kaiser rule, parallel analyses of Monte Carlo simulations (Horn, 1965), and the interpretability of the solutions (Zwick and Velicer, 1986), all converging towards this four categories classification.

Rentfrow and Gosling (2003) introduced a novel approach to structuring music preferences, providing a comprehensive framework that has significantly influenced further research in music psychology. Over time, their experiment was replicated in various forms as evidenced by multiple studies (Delsing et al., 2008; Brown, 2012; Langmeyer et al., 2012; Schäfer and Sedlmeier, 2009; George et al., 2007). These replications revealed a broad consistency in results,

yet they also highlighted some inconsistencies in the number of factors identified and how music genres were categorized. These variations are attributable to several factors: cultural differences (as Rentfrow and Gosling (2003) initial studies were confined to the United States, while subsequent studies spanned different countries), demographic limitations (original studies focused on college students, who do not represent the broader population), and methodological discrepancies.

A significant issue with previously mentioned studies is their reliance on genre classification, which itself can introduce bias and limit the scope of results. Genres, being pre-established categories, might influence how respondents perceive and evaluate the music, due to preconceived notions or biases associated with those genres. Moreover, different individuals might categorize the same music under different genre labels, leading to inconsistencies in data collection and analysis. Furthermore, while these studies often categorize music based on general preferences, such as liking or disliking a piece, they typically overlook the specific emotions and feelings that the music evokes in listeners.

In 2011 and 2012, Rentfrow and peers Rentfrow et al. (2011, 2012) replicated their study with an updated methodology. Instead of asking to rate genres directly, they picked several music samples from different genres that the respondents had to listen to, and then rate them on different levels : auditory features, affect, energy level, perceived complexity etc. Several experiments where made for each paper. Both studies converged towards a five-factor solution: Mellow, Unpretentious, Sophisticated, Intense, and Contemporary (MUSIC).

In order to address the cultural specificities in music perception, Cowen et al. (2020) explore the diverse emotional responses that music evokes in listeners from the U.S. and China. The research investigated how music across these cultures triggers specific emotions and broader affective states like valence (the emotional value associated with a stimulus) and arousal (the intensity of emotion provoked by a stimulus). The primary aim was to understand the taxonomy of emotional responses to music and whether these responses are consistent across different cultural contexts. Specifically, it sought to determine how music induces a range of emotions and how these are conceptually organized across cultures.

1,591 U.S. participants and 1,258 Chinese participants listened to 2,168 music samples from modern and classical Western music as well as traditional Chinese music. They then had to describe their feelings using terms from a list of 28 emotions (sad, dreamy etc.) and along scales that captured 11 affective features (valence, arousal etc.). Statistical methods were used to analyze and categorize the types of emotional responses evoked by the music samples. Factor analysis helped uncover underlying relationships between emotional responses and music, while hierarchical clustering grouped similar emotional expressions. Regression analysis was used to investigate the correlation between musical fea-
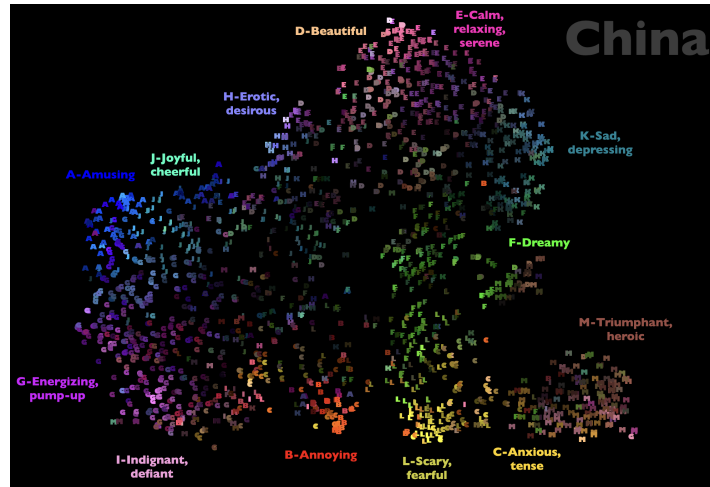
Figure 4.2: Interactive music emotions map, representing the 13 music emotions (in China) from Cowen et al. (2020). Each point corresponds to a music excerpt. The map is available on Alan Cowen's website[3].

tures and emotional reactions.

The research identified 13 distinct emotional experiences that music can induce: amazement, joy, beauty, nostalgia, sadness, peacefulness, power, tenderness, transcendence, tension, wonder, triumph, and longing (Figure 4.2). In comparison to the MUSIC model, this segmentation offers a more detailed taxonomy. Specific emotions were better preserved across cultures than generalized affective states like valence and arousal. Both U.S. and Chinese listeners reported similar emotional responses to a wide array of music samples, indicating some level of universality in emotional responses to music. However, certain emotions differed: for example, when U.S. participants reported feeling triumph, participants from China experienced feelings of beauty and transcendence. These results recall that cultural background might have an impact on one's perception of music, and thus has to be taken in account when labelling and classifying music.

4.2.2 Behavioural data

Unlike content-based features, context-based features do not require access to the actual music file. Therefore, systems such as music information systems can be developed without needing an acoustic representation of the music, simply by using a list of music items. However, this approach requires access to extensive, clear, and ideally noise-free user-generated data. If this condition is met, community data offers a valuable source of information on social context, capturing the 'collective wisdom of the crowd' without requiring explicit or direct human input.

With the rise of the Internet and streaming platforms, data on users' listening histories, likes and ratings has become accessible. This data can be used to compute similarities between musical items, positioning them in either a continuous or discrete space. This approach is central to my own work and publications, thus deserves special attention. We will address the choice of the data, normalization techniques, diverse similarity metrics, and clustering algorithms, specifying the nuances of each step, and giving examples of existing literature and/or practical examples on Deezer data.

Choosing the data and similarity criteria

The initial step is to decide which data to analyze, beginning with the choice of music items, e.g. tracks, albums, artists. For example, tracks are more specific but present computational challenges due to their high volume. On the other hand, artists and albums are broader categories that might simplify analysis. However, the same artist might have explored different styles, and different tracks from the same artists may address to different audiences.

Then depending on the source of data, a criteria for determining similarity must be established. For example, before the streaming era, Cohen and Fan (2000); Zadel and Fujinaga (2004); Schedl et al. (2005) and Schedl (2008) studied the co-occurrence of artists names on web pages, considering that the more often two artists are mentioned on the same web pages, the more similar they are. Pachet et al. (2001) used playlists from a French radio station and CD compilation databases to analyze co-occurrences between artists and tracks.

Cano and Koppenberger (2004) and Aizenberg et al. (2012) also used the co-occurrence of artists in playlists. Baccigalupo et al. (2008) introduced a distance function to account for how closely two artists appear in playlists. Later, the co-occurrence of tracks in playlists started serving recommendation purposes (Kim et al., 2018; Bendada et al., 2023). The advantage is that playlists are not only created by editorial teams, but also by simple users, which means the reflect a more realistic picture.

Last but not least obviously the behaviour of users of streaming platforms can tell us a lot about the music items they interact with. It includes streaming data — if a lot of people listened to the same tracks those tracks must be similar. This is a nice approach, but we have to be careful about the complexity because of huge amount of data. A stronger factor are likes — people didn't just listen, they have a special relationship with this musical item (Bendada et al., 2023; Matrosova et al., 2024b). Also it allows to reduce the number of items, and to not have to consider the temporality factor, like how far should we go in the streaming history.

Regardless of the chosen data, it can typically be represented as a mn matrix M, with items as rows, and terms, playlists, or users as columns (Figure 4.3). Depending on the nature of the data, the matrix can be binary or non-binary. For

	Playlists			
	p1	p2	...	pn
i_1	1	1	...	1
i_2	0	1	...	0
...	...	...	...	...
i_m	1	0	...	0

Items — Occurrences

	Users			
	u_1	u_2	...	u_n
i_1	0	1	...	0
i_2	1	0	...	1
...	...	...	...	...
i_m	1	0	...	0

Likes

	Users			
	u_1	u_2	...	u_n
i_1	8	5	...	21
i_2	0	5	...	1
...	...	...	...	...
i_m	7	0	...	0

Number of streams

Figure 4.3: Examples of interaction matrices.

binary data, we simply consider the occurrence of an event, such as whether a song is in a playlist or if a user has liked or listened to a song. For non-binary data, we count the occurrences, such as the number of times a user has listened to a particular song. From now on, if not specified otherwise, M will denote an item-user interaction matrix.

Normalizing the data

If we keep the matrix as it is, it can contain a certain amount of biases that we might want to diminish. For example, if we count the number of times a user streamed a certain song, not all users have the same intensity in usage of streaming platforms, so for one user 5 streams may be 50% of their weekly streams, while for another it will be 1%. This case, we want to make a column-wise normalisation of M. An even stronger bias might be caused by the popularity of the music items. As we know the popularity of music items follow a 'long-tail' distribution, which means that a very small amount of most popular items end up as the most cited/listened/liked. In the matrix, the row corresponding to a top artist will be very dense, while for most, less popular artists the row will be very sparse. A stream of a major artist then 'counts less' that a stream of a niche one. In order to diminish this bias we normalize the interaction matrix row-wise. There are two ways to normalise a matrix by vector norms.

L1 normalization, also known as **Manhattan normalization**, consists in dividing each element by the sum of the absolute values of all elements in the row (or column).

$$M_{\mathrm{L1}} = \frac{M}{\sum_{j=1}^{n} |M_{ij}|}$$

L2 normalization, also known as **Euclidean normalization**, normalizes the data by dividing each element by the square root of the sum of the squares

of all elements in the row (or column).

$$M_{\text{L2}} = \frac{M}{\sqrt{\sum_{j=1}^{n} M_{ij}^2}}$$

'

Dimensionality reduction

A problem that can occur with huge datasets, is that calculating similarity metrics on M as it is might be impossible because of the high complexity of similarity algorithms. In this case, item vectors can be reduced to **embeddings** — lower-dimensional vectors that capture the main aspects of their original matrix relationships and properties. Let's dive into the most common dimensionality reduction techniques.

Singular value decomposition (SVD) is probably the most widely used **matrix factorization (MF)** technique, one of it's applications being dimensionality reduction. MF techniques are widely used in RS to discover latent features underlying the interactions between users and items. These techniques consist in approximating M by the product of two (or more) smaller matrices that capture latent factors about users and items. SVD breaks down matrix M into three matrices as follows:

$$M = P\Sigma Q^T$$

where :

- P (left singular vectors) is an $m \times m$ orthogonal matrix, where columns are eigenvectors of MM^T.
- Σ (singular values) is an $m \times n$ diagonal matrix with non-negative real numbers are sorted in descending order, representing the square roots of the eigenvalues of M^TM or MM^T.
- Q (right singular vectors) is an $n \times n$ orthogonal matrix, where columns are eigenvectors of M^TM.

The idea is that, as the singular values in Σ follow a descending order, and eigenvectors in P and Q are sorted accordingly, the first eigenvectors contains the biggest part of information (and variance) of the initial matrix, and the last ones are the least informative. By discarding the smallest eigenvalues, and corresponding eigenvectors, the dimensionality can be reduced without loosing too much information (Figure 4.4). In order to know 'where to cut', i.e. choose the optimal number of components, a middle ground must be found between minimizing the number of dimensions, while maximizing cumulative explained variance ratio. There are several ways of achieving this:

- One common approach is to choose the number of dimensions such that a desired percentage of the total variance in the data is explained. For instance, retaining enough dimensions to explain 90% of the variance often
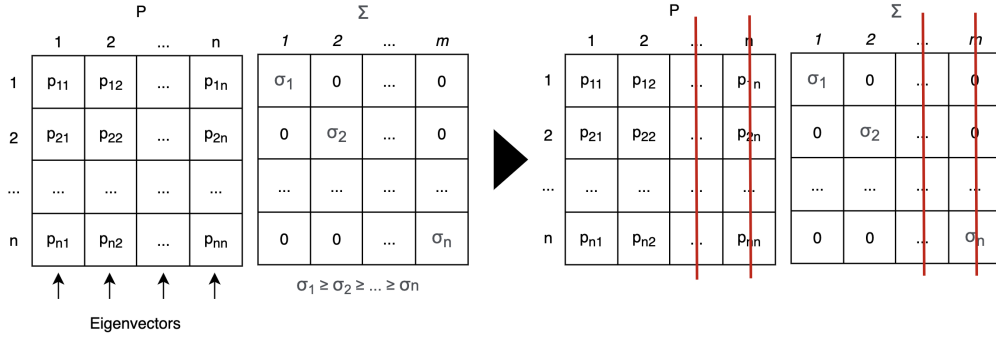
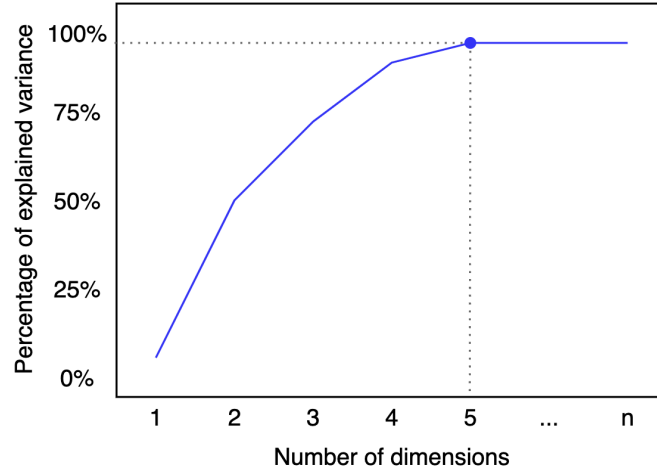Figure 4.4: Discarding the smallest eigenvalues and corresponding eigenvectors in SVD.

Figure 4.5: Illustration of the 'elbow' method to select the optimal number of dimensions/components in SVD.

provides a good balance between information retention and dimensionality reduction. This is calculated by summing the squares of the singular values.

- A scree plot, which shows the eigenvalues or singular values in descending order, can be used to visually assess where the values start to level off (often referred to as the 'elbow' method). The point at which the decrease in singular values becomes less pronounced is typically a good choice for cutting off the number of dimensions (Figure 4.5).

Here, we only take interest the items, i.e. the rows in M. In this specific case, SVD can be performed both directly on M, or on MM^T, as it decomposes into:

$$MM^T = P\Sigma P^T$$

where U contains the eigenvectors (also the left singular vectors of M, and Σ

is a diagonal matrix with the square roots of the eigenvalues of MM^T (which are also the singular values of M^2). Either way, the item vectors in matrix P form so-called embeddings, that represent the items' latent features. These embeddings can then be used to compute similarity metrics and/or to perform clustering algorithms.

Unlike recommendation tasks, where MF techniques like SVD are an end in themselves, usually, when the final goal is to simply understand the items similarities and correlations, SVD is used only as a step for dimensionality reduction. However, an interesting finding by Afchar et al. (2023) suggests that in the case of music data specifically — as it is particularly prone to exhibiting community effects linked to its many historical and cultural groundings — SVD embeddings are interpretable and can directly reveal existing item communities. They observed (on 6 different datasets) that embedding vectors tend to self-organize along lines that pass through the origin, a pattern that is no longer visible once similarity metrics are performed.

To explore this behaviour, authors used the degree-corrected block model (Degree-Corrected Block Model (DCBM)) (Karrer and Newman, 2011), a extension of the stochastic block model (Stochastic Block Model (SBM)) (Holland et al., 1983). The SBM is a probabilistic model for network data that groups nodes into blocks or communities, with the connection probability between any two nodes depending on their respective community memberships. The DCBM extends this by allowing node degrees to vary, acknowledging that nodes within the same community can have different levels of connectivity. This degree correction makes DCBM particularly suited for modeling networks like music streaming data, where both community affiliation and individual popularity (degree) play crucial roles.

Prior literature (Jin, 2013; Lei and Rinaldo, 2015) suggested that the diagonalization of DCBM matrices leads to the appearance of spikes. Afchar et al. (2023) not only confirmed this, but also demonstrated the reciprocal relationship, showing how such spikes in SVD embeddings are indicative of the community structures modeled by DCBM. Finally, they show that on real-life streaming data, the spikes indeed correspond to similar music items, i.e. communities.

Principal Component Analysis (PCA) is a technique that is closely related to SVD, however PCA is specifically aimed at reducing the dimensionality of a dataset, while SVD is a more general matrix decomposition technique which can be used for many other purposes. Because the primary goal of PCA to identify the directions in which the data varies most significantly, it uses the covariance matrix: it is a key tool in understanding these variances and the correlations between different variables in the dataset. The eigenvectors and eigenvalues are then computed on the covariance matrix using SVD, picking up the first few eigenvectors, which capture the most variance, as principal components /
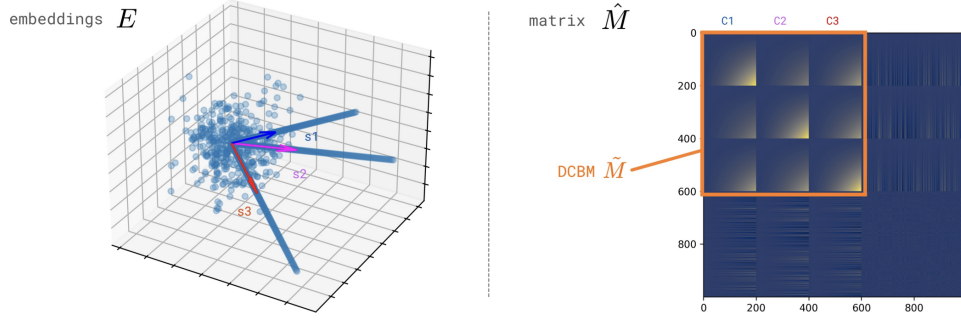
Figure 4.6: Parallel between spikes in SVD and communities in DCBM (illustration from Afchar et al. (2023))

dimensions.

The embeddings on their own are autonomous and represent the items in a space. Then they can be used to compute similarity metrics or clustering algorithms.

Computing similarity

Once the matrix M is set, each musical item is defined by a vector in a multi-dimensional space. One straightforward way to identify the distances between musical items is to compute the similarity between two item vectors. A lot of similarity / distance metrics exist for this purpose.

Euclidean distance is a measure of the straight-line distance between two points in Euclidean space. It is used by Slaney and White (2007) on music items' ratings, by Shavitt and Weinsberg (2009) on co-occurrences of peer-to-peer music sharing.

$$\text{dist}(M_i, M_j) = \sqrt{\sum_{k=0}^{n}(M_{ik} - M_{jk})^2}$$

Cosine similarity measures the cosine of the angle between two vectors, indicating how similar they are, with values ranging from -1 (completely dissimilar) to 1 (identical). Sometimes, cosine distance is used, which is derived from cosine similarity and is defined as $1-$ cosine similarity. For example, Knees et al. (2004); Anderson et al. (2020) use cosine similarity on music items rating vectors and songs' co-occurrence in playlists respectively.

$$\text{sim}_{\cos}(M_i, M_j) = \frac{M_i \cdot M_j}{\|M_i\| \cdot \|M_j\|}$$

Pearson's correlation coefficient measures the linear relationship between two vectors, giving a value between -1 and 1, where 1 means a perfect positive

linear correlation, -1 means a perfect negative linear correlation, and 0 indicates no linear correlation. Pachet et al. (2001) songs' co-occurrence in playlists. Sánchez-Moreno et al. (2016) use both cosine similarity and Pearson correlation for music recommendation, representing the items through users' and socio-demographic data.

$$P(M_i, M_j) = \frac{\frac{1}{n}\sum_{k=0}^{n}(M_{ik} - \bar{M}_i)(M_{jk} - \bar{M}_j)}{\sigma_{M_i}\sigma_{M_j}}$$

Jaccard Coeffcient measures the proportion of shared elements (for binary vectors only) relative to the total number of elements that have a 1 in at least one of the vectors. McFee et al. (2012) apply Jaccard coefficient on a matrix recording if users' have interacted with music items. Afchar et al. (2023) uses it on SVD embeddings of artists' co-occurrence in playlists.

$$J(M_i, M_j) = \frac{|M_i \cap M_j|}{|M_i \cup M_j|}$$

Clustering

Finally, music items can be automatically split into distinct categories, using either the original interaction matrix M or embeddings derived from it.

K-means (MacQueen et al., 1967) is one of the most popular clustering algorithms. It is a centroid-based algorithm that partitions a dataset into k distinct, non-overlapping clusters.

The algorithm starts by selecting k data points as the initial centroids (randomly, or though more effective centroid initialization techniques like k-means++). Then, it assigns each data point to the nearest cluster centroid, usually based on Euclidean distance, but other distance metrics can be used as well. Once all data points have been assigned to clusters, the centroids of these clusters are recalculate as the mean of all data points that belong to each cluster. The steps are reiterated until the centroids do not change significantly between iterations, or a maximum number of iterations is reached.

K-means is particularly efficient on large datasets, with a complexity of $O(mkt)$, where k is the number of clusters and t the number of iterations. Several techniques exist to identify the optimal value of k, however most of them require to perform the algorithm several times with different values of k beforehand, which can add computational time. To cite a few — the elbow method involves plotting the sum of squared errors against different values of k and looking for the 'elbow' where improvements in the sum of squared errors diminish; the Silhouette score measures how similar an item is to its own cluster compared to other clusters for different values of k, and the k that maximizes the average Silhouette score is considered optimal.

K-means algorithm has been broadly used with music data. For example, Kim et al. (2007); Shavitt and Weinsberg (2009); Pavitha et al. (2022); Mukhopadhyay et al. (2024) use k-means clustering for MRS. Pavitha et al. (2022) in particular compare the performance of different clustering algorithms and finds k-means to perform better than other algorithms. Way et al. (2019) use k-means among other clustering methods to detect music genres from streaming data and state that all of the used clustering techniques produce similar-scoring partitions of the data. Cai et al. (2021) propose a method based on k-means to study the evolution of music genres.

A very similar method, **k-medoids**, is based on the same principle, but it uses the most centrally located item in a cluster as the centroid, instead of using the mean of the items in each cluster. Privandhani et al. (2022), for instance, compares the results of k-means and k-medoids clustering on streaming data from Spotify, suggesting that k-medoids tends to be more robust against outliers and might provide a more representative clustering by focusing on the most centrally located objects in a cluster. Also the medoids can be used to extract knowledge about the clusters, as they can be interpreted as the most representative items of their clusters Matrosova et al. (2024b).

Another popular approach is **hierarchical clustering**. It is a group of clustering methods (Murtagh and Contreras, 2012), that can be categorized into two main types :

- **Agglomerative clustering** is a bottom-up approach. It starts with each data point as its own cluster. Thus, if there are m items, we begin with m clusters. Recurrently, the closest pair of clusters are merged into one, until all points are merged into a single cluster or until a specified number of clusters is achieved. This merging is based on a defined distance metric (similarity between individual items) and linkage criterion (distance between clusters). This approach is the most commonly used.

- **Divisive clustering** is a top-down approach. It starts with all data points in a single cluster. Recurrently, each cluster is split into two smaller clusters using a flat clustering method, like k-means for example, until each cluster contains only a single data point or until a specified number of clusters is achieved.

In both cases, the result can be visualized using a dendrogram, a tree-like diagram that records the sequences of merges and shows the distance at which each merge occurred (bottom-up approach) or how large clusters were divided (top-down approach). This visual summary can be particularly useful in the case of music data, as music genres are classically represented and understood in a dendrogram shape. The advantage of this method is that it does not require specification of the number of clusters in advance. Some disadvantages however exist : hierarchical clustering is sensitive to noise and outliers, and, especially for agglomerative clustering, the algorithm can be computationally expensive,

with complexity $(O(m^3))$, making it impractical for large datasets.

Li et al. (2011) organize music data based on user-assigned tags, artist-related style, and mood labels, extracted from Last.fm and All Music Guide[4] websites. They show that agglomerative clustering outperforms divisive clustering in this task. Way et al. (2019) use agglomerative clustering, among other clustering algorithms, based on streaming data from Spotify to identify different music genres. In this study, all clustering methods gave similar results. Pavitha et al. (2022) states that though agglomerative clustering is effective on streaming data, it suffers from its computational complexity, and the use of k-means is preferable with big datasets. With a nod to the previous section, a range of studies have used hierarchical clustering based on music features : for example Cilibrasi et al. (2004) effectively distinguished between musical genres and composers by transcribing music pieces into strings, and Le Bel (2017) uses agglomerative clustering for audio classification.

Other clustering methods exist, however they seem to be less popular when it comes to music data classification.

Spectral Clustering (Von Luxburg, 2007) uses the items similarity graph and segments it into several small groups with similar values. It starts by constructing a similarity graph, where each node represents an item, connected to others based on a similarity measure, often Gaussian similarity. A key step involves calculating the graph's Laplacian matrix, defined as $L = D - S$, where S is the matrix of similarities between items of size $m \times m$ and D is the degree matrix, sums all the similarities between item i and all other items. In the Laplacian matrix the smallest eigenvalues (excluding zero) and their corresponding eigenvectors capture the most significant structure of the data. These eigenvectors are used as principle components, and form new features that reposition the original data into a space where clusters are more distinguishable. The clusters can then be effectively identified using standard clustering techniques like k-means.

In comparison to traditional methods like k-means, that assume that clusters are separable linearly, spectral clustering can be helpful in identifying clusters with complex shapes and connectivity patterns. In the case of music items, we can assume that they can group into non-convex clusters, for example if a genre is a fusion between several others, items from this genre will be surrounded by items from other genres.

In practice, however, this specificity has not been much explored. Karydis et al. (2009) perform spectral clustering on music items from Last.fm in three different way : considering only the item-tag relationship, user-item relationship, and tripartite relationship among users, music items, representing the data as 3D tensors (which are multi-dimensional embeddings, so to say). The 3-dimensional approach showed a better Silhouette score, however, the spectral clustering was not compared to other algorithms. Way et al. (2019) state that

---

[4]All Music Guide

spectral clustering performs similarly to k-means and agglomerative algorithms on Spotify artists, whose similarity was derived from the frequency with which listeners stream two artists in succession.

On the other side, Darke and Below Blomkvist (2021), who use spectral clustering on Spotify data, where each song was represented as 30-dimensional vectors based on the songs' metadata, ended up with one cluster of 40,718 songs and 22 groups of $/approx$ 100 songs, suggesting that the method is not suitable for this kind of datasets.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is an algorithm that groups together closely packed points and identifies outliers in noisy datasets (Ester et al., 1996). It operates based on two key parameters: eps (epsilon), which defines the radius of a neighborhood around each point, and $MinPts$ (minimum points), which specifies the minimum number of points required to form a dense region. Each point in the dataset is classified as a core point, border point, or noise point based on these parameters. Core points have at least $MinPts$ within their eps neighborhood, border points have fewer than $MinPts$ but are in the neighborhood of a core point, and noise points are those that do not meet the criteria to be either core or border points. DBSCAN then iteratively expands clusters starting from core points, incorporating all directly reachable points, and thus effectively handles clusters of varying shapes and sizes while robustly filtering out noise and outliers.

The effectiveness of the DBSCAN algorithm for clustering music, particularly with streaming data, presents mixed findings. Several studies highlight its shortcomings: Nordström and Håkansson (2012) found that DBSCAN produced less favorable outcomes than k-means, attributing this to the varied densities of data points in the dataset, and Pavitha et al. (2022) concluded that DBSCAN is not well-suited for music classification. Conversely, Kużelewska and Wichowski (2015) introduced a modified version of DBSCAN that outperformed CF approaches in the context of music recommendation. This suggests potential for tailored versions of DBSCAN in specific applications despite its general limitations.

4.3   Conclusion

In this chapter, we have explored the ways in which musical space can be categorized, labeled, and represented, drawing from a diverse array of fields including musicology, the music industry, streaming platforms, MIR, and psychology. Each field brings its own priorities and methodologies to the classification and organization of music, influencing the approaches to music representation.

One of the key conclusions of this analysis is the recognition that no single method of music classification can fully encapsulate the complexity and diversity of musical expressions. The categorization systems that prioritize technical

musical features, such as rhythm and timbre, do not always align with those that categorize music by genres or emotional impact. This divergence underscores the challenge of creating a universal system of music classification that is culturally and contextually inclusive.

Human annotations obviously bring a depth of understanding that automated systems often struggle to achieve. Expert musicologists can discern nuances in genre, historical significance, and cultural context that are crucial for a comprehensive music taxonomy. However, the scalability of human annotation is limited; as music databases expand exponentially, relying solely on human expertise becomes impractical. This limitation is particularly significant given the pace at which new music is produced and consumed on digital platforms.

The Internet provides an excessive amount of data, ranging from tags to actual listening traces. One might think that with such an abundance of data, the possibilities for music classification and recommendation are nearly limitless. However, because this data is reflective of real-life human behaviors, it inherently introduces biases related to the popularity of artists or songs. As noted by Celma Herrada et al. (2009), the majority of music catalogs fall into the long tail; they are unpopular and therefore less likely to be well-documented or tagged. This scenario is known in RS as the 'cold-start' problem, which will be further explored in Chapter 5.

For these lesser-known and sparsely tagged tracks, audio-based solutions can provide significant aid. By analyzing the intrinsic properties of music—such as tempo, melody, and rhythm—these methods enable the labeling of large portions of a catalog, thereby making them more accessible to users. This increased accessibility not only enhances the user experience by diversifying the music they encounter but also generates more data on these tracks, potentially increasing their visibility and popularity. In turn, this can lead to a more equitable representation of diverse musical expressions within digital platforms, gradually mitigating the long-tail issue and fostering a richer, more inclusive musical ecosystem.

Chapter 5

Recommending music

Today, when we think about music streaming platforms, we almost immediately think about algorithmic recommendation. Algorithmically generated playlists are usually put forward by the platforms, in more and more different ways — mixes inspired by a certain genre, artist, song, mood, or the users' streaming history as a whole — the chances are maximized for users to find what they are looking for. Indeed, the music catalogs, interfaces, and even subscription plans proposed by most streaming services (Deezer, Spotify, Apple Music, Tidal etc.) are quite similar, making the quality of music recommendations a potentially important factor in the users' choice of platform. As more and more people join music streaming platforms, more data is available to improve algorithmic recommendations. Conversely, the more users resort to algorithmic recommendations, the more these algorithms might have an impact on the music they listen to.

At its core, a recommender system (RS) connects users with items, which means that an effective music recommender system (MRS) must be able to detect and use patterns in both user preferences and the musical space. In the previous chapters, we explored these two key dimensions. In Chapter 3, we focused on understanding, quantifying, and modeling musical taste, exploring how user preferences are formed, how they evolve over time, and how they can be computationally represented. This understanding is critical for tailoring recommendations to individual users. On the other hand, the system also needs to capture similarities between music items to make accurate suggestions. In Chapter 4, we examined how knowledge about music can be extracted — whether based on audio features, metadata, or user interaction — enabling the system to navigate the vast music catalog and make meaningful connections between music items.

With these foundations in place, we can now dive deeper into understanding not only how MRS work, but also how they can influence our music preferences. As we have seen, there are many ways to represent both user preferences and the

space of music items. Every decision in designing a RS — the data we choose to provide, how we structure and represent it, how the algorithms uses it — directly affects the recommendations users receive. Over time, this may contribute to reinforcing or altering users' musical tastes, for example driving users toward certain types of music and reducing exposure to niche or local content. The impact of these systems is not limited to individual user preferences: they can also shape broader trends in music consumption, promoting certain artists or genres while potentially sidelining others. This makes it essential to carefully consider how the technical choices made in MRS affect not only user satisfaction, but also diversity, discovery, and fairness.

This chapter is structured as follows. First, we will make an overview of the methods used by MRSs, highlighting the advantages and limitations of each method. Following that, we will address the specific challenges associated with music recommendation and ways to address them. These include the cold-start problem, which concerns not only music but any RS in general, the ratio of repetition and discovery there should be in the recommended music, the role of context, and finally the fairness of MRS and the possible biases that algorithmic recommendation can create or emphasize.

5.1   Recommender systems overview

A RS is an information filtering system that predicts and provides a set of recommended items for a user based on their known preferences or other data (Adomavicius and Tuzhilin, 2005; Ricci et al., 2011). Let:

- U be the set of users.
- I be the set of items available for recommendation.
- $V(u) \subseteq I$ be the set of items that user $u \in U$ has previously interacted with or expressed a preference for.
- $R(u) \subseteq I$ be the set of items recommended to user u by the system.

The goal of the RS is to find the set $R(u)$ that maximizes the likelihood of user u interacting positively with the items, given their known preferences $V(u)$. Mathematically, this can be expressed as a function $f : U \times I \to \mathcal{R}$ that computes a relevance score for each item $i \in I$ for a given user u, based on their past behavior or other data:

$$f(u, i) \to \mathrm{score}(u, i)$$

The set of recommended items $R(u)$ is then defined as the top k items from the set I that have the highest predicted scores according to the function $f(u, i)$:

$$R(u) = \arg\max_{i \in I} \left(f(u, i) \right), \quad |R(u)| = k$$

The primary recommendation approaches include Content-Based Filtering (CBF) — which focuses on the features of the items a user has interacted with, collaborative filtering (CF) — which leverages the behavior of similar users to predict preferences, and hybrid approaches, which combine both these methods.

5.1.1   Content-based filtering

The fundamental principle behind content-based filtering (CBF) is that recommendations are made by analyzing the features or attributes of the items that a user has interacted with in the past. For example, in a MRS, if a user frequently listens to tracks with similar genres, rhythms, or tempos, the system will recommend other songs with those same features (Pazzani and Billsus, 2007; Lops et al., 2011). To this end, the items and user profiles are represented through feature vectors. Each item i is described by a vector $\vec{v}_i \in \mathcal{R}^j$, where j is the number of features. In a music streaming service, the features might include taxonomies like genre or mood, or audio characteristics such as pitch, tempo, and rhythm Schedl et al. (2015). The profile of a user u can typically be computed as the average of vectors corresponding to music items $V(u)$ that the user has liked or interacted with in the past:

$$\vec{v}_u = \frac{1}{|V(u)|} \sum \vec{v}_{i \in V(u)}$$

The RS then suggests new items by computing the similarity between the user's profile vector $\vec{v}_u$ and other item vectors $\vec{v}_i$, using a similarity metric like cosine similarity:

$$\text{similarity}(\vec{v}_u, \vec{v}_i) = \frac{\vec{v}_u \cdot \vec{v}_i}{\|\vec{v}_u\| \|\vec{v}_i\|}$$

De Gemmis et al. (2015) presents the high-level architecture of a content-based RS as three main components: the Content Analyzer, the Profile Learner, and the Filtering Component (Figure 5.1).

- The **Content Analyzer** is responsible for processing unstructured information, such as text or audio data, to extract structured and relevant features. Its main task is to convert items, such as songs, into feature vectors $\vec{v}_i$ using feature extraction techniques. This structured data produced by the Content Analyzer is then used as input for both the Profile Learner and the Filtering Component.

- The **Profile Learner** collects data representative of user preferences and generalizes this data to construct a user profile. For instance, the user profile $\vec{v}_u$ is created based on the feature vectors of songs the user has liked. This generalization can be done through machine learning methods, which infer user preferences from past interactions.
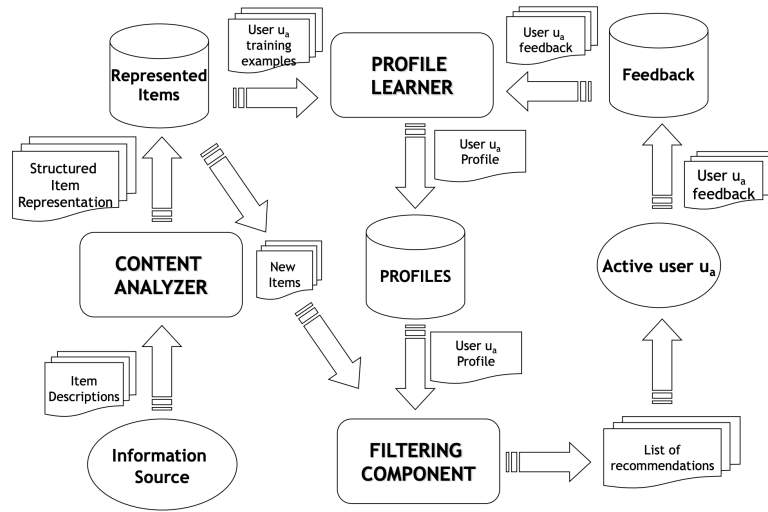
Figure 5.1: High level architecture of a CBF system by De Gemmis et al. (2015).

- The **Filtering Component** uses the user profile $\vec{u}$ to suggest relevant items by matching the profile against the items' feature vectors $\vec{v}_i$. The matching process involves calculating a similarity score (e.g., cosine similarity), to rank items based on how well they match the user's preferences.

In summary, the Content Analyzer takes unstructured data and converts it into structured feature vectors $\vec{v}_i$. The Profile Learner aggregates these vectors to create a user profile $\vec{v}_u$ using machine learning techniques. Finally, the Filtering Component matches this profile with new items by computing similarity scores, producing a ranked list of recommendations.

Advantages

- **No need for other users' interactions data**: as it does not rely on user-item interactions in order to determine the similarity between items, CBF only needs the user's own interaction history with items, making it effective even when user data is sparse.

- **Items cold start problem**: CBF relies on the items' intrinsic properties, making it efficient even for new items.

- **Interpretability**: It is relatively easy to understand and explain why a particular item was recommended, as the recommendations are based on item features.

Limitations

- **Limited discovery**: Users may be limited to recommendations that are very similar to what they have already interacted with, reducing the po-

tential for discovering new or diverse items.

- **Feature engineering**: The effectiveness of CBF relies heavily on the quality and completeness of the item features. Poor feature selection can lead to poor recommendations.

- **Users cold start problem**: While CBF can handle new items better than CF, it still struggles with new users that have not yet interacted enough with the platform.

5.1.2 Collaborative filtering

The fundamental principle behind CF is that users who have had similar preferences in the past will continue liking similar things in the future (Herlocker et al., 2000). This means that it is possible to predict a user's preference for an item by finding similar users or similar items and aggregating their ratings or interactions. The approach relies on large datasets of user-item interactions to identify patterns and similarities, enabling the system to make personalized recommendations. CF techniques can be split into two main types (Ricci et al., 2011). Memory-based approach includes user-based CF, which finds users with similar tastes to recommend items, and item-based CF, which finds similar items based on users' past interactions. Model-based approaches like matrix factorization (MF) leverage latent factor models to uncover hidden patterns in user-item interactions, providing a scalable way to handle large datasets and sparse interactions (Koren et al., 2009; Adomavicius and Tuzhilin, 2005).

User-based collaborative filtering

User-based CF predicts a user's interest in an item by finding similar users who have shown similar preferences in the past (Herlocker et al., 2000). The primary assumption is that if user u has a similar rating pattern to user u', then user u is likely to rate new items similarly to how user u' has rated them. Let:

- $M \in \mathcal{R}^{|U| \times |I|}$ be the user-item interaction matrix, where $M_{u,i}$ is the rating or interaction of user u with item i.

- $similarity(u, u')$ denote the similarity between two users u and u'.

The predicted rating of user u for an item i can be estimated by aggregating the ratings of similar users $u' \in U$, typically weighted by their similarity to u:

$$\hat{M}_{u,i} = \frac{\sum_{u' \in N(u)} similarity(u, u') \cdot M_{u',i}}{\sum_{u' \in N(u)} |similarity(u, u')|}$$

where $N(u)$ represents the set of nearest neighbors (similar users) to user u.

Item-based collaborative filtering

Item-based CF focuses on the similarity between items rather than users. Schafer et al. (1999) describe item-based CF as looking at the items a user has liked and recommending similar items, assuming that items rated similarly by different users are likely to be perceived as similar in quality or taste. Let $similarity(i, i')$ denote the similarity between items i and i'. The predicted rating of user u for an item i can be estimated by aggregating the ratings $M_{u,i'}$ of the user for similar items $i' \in I$:

$$\hat{M}_{u,i} = \frac{\sum_{i' \in N(i)} similarity(i, i') \cdot M_{u,i'}}{\sum_{i' \in N(i)} |similarity(i, i')|}$$

Here, $N(i)$ represents the set of nearest neighbor items to item i.

Matrix factorization

MF represents a more sophisticated approach to RSs by building predictive models based on user-item interactions (Koren et al., 2009). The idea is to 'fill-in' the gaps in the interaction matrix M, that correspond to a user that has not interacted with an item, by predicting hypothetical scores, based on the patterns observed in the interactions that have actually been made. To achieve this, we approximate M by decomposing it into two lower-dimensional matrices, P and Q, which represent the latent factors of users and items, respectively (Figure 5.2). These latent factors are intended to capture hidden patterns or traits in the data that help explain the relationship between users and items. In a RS, they capture the underlying reasons (e.g., genre preferences) why a user might like or dislike certain items. Mathematically, we decompose M as:

$$M \approx P\Sigma Q^T$$

where:
- $P \in \mathcal{R}^{|U| \times k}$ is a matrix representing users, with k latent factors.
- $Q \in \mathcal{R}^{|I| \times k}$ is a matrix representing items, with k latent factors.
- $\Sigma \in \mathcal{R}^{k \times k}$ is a diagonal matrix containing singular values, which scales the latent factors.

Once we decompose M, the product of the matrices P and Q^T approximates M, 'filling' the empty cells with predicted values. These predicted values represent the estimated ratings or interaction scores for user-item pairs that were previously unobserved.

Several techniques exist to decompose the interaction matrix M. One deterministic approach is using SVD-based methods, which factorize M into singular vectors and values. However, pure SVD assumes a fully observed ma-
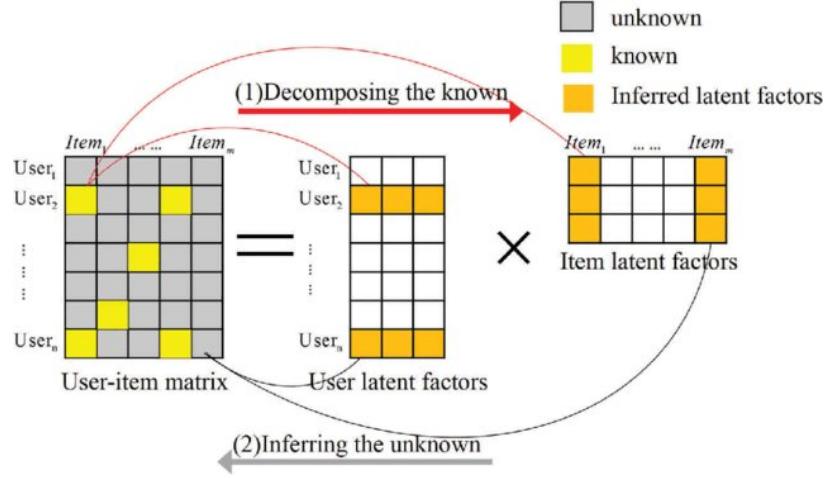
Figure 5.2: Illustration of MF by Liu et al. (2017).

trix, so variants like Truncated SVD are typically used in recommendation tasks
with sparse data (Koren et al., 2009). Other techniques, like alternating least
squares (Alternating Least Squares (ALS)) (Hu et al., 2008), iteratively opti-
mize user and item latent factors Zhou et al. (2008), while neural network based
MF methods extend traditional MF by leveraging deep learning to capture non-
linear interactions between users and items , (He et al., 2017). These usually
aim to minimize a loss function, such as the squared error, to find matrices P
and Q that minimize the error between the actual user-item interactions $M_{u,i}$
and the predicted interactions $P_u^T Q_i$:

$$\min_{P,Q} \sum_{(u,i)\in K} (M_{u,i} - P_u^T Q_i)^2 + \lambda(\|P_u\|^2 + \|Q_i\|^2)$$

where:

- K is the set of existing user-item interactions.
- λ is a regularization parameter that controls overfitting (which would lead
  to recreating the exact matrix M, without filling the empty cells) by pe-
  nalizing large values in the latent vectors.

In practice, MF is one of the most widely used techniques in RSs because
effectiveness in handling large-scale, sparse datasets, making it suitable for large
platforms, like those of music streaming (Koren et al., 2009; Zhang et al., 2019;
He et al., 2017).

Advantages

- **No need for item metadata**: Unlike CBF, CF does not require detailed
  information about the items themselves. It relies solely on user-item in-
  teractions, making it applicable to a wide range of domains.

- **Discovering unobvious patterns**: Since CF relies on the behavior of other users, it can recommend items that are not necessarily similar in content but have been liked by similar users. This helps in introducing users to new and diverse items that they might not have found through CBF alone.

- **Scalability with user base growth**: As the number of users and their interactions grow, the system can improve its recommendations by learning from more data. More user interactions generally lead to better accuracy.

Limitations

- **Cold start problem**: CF struggles with both new users and new items due to the lack of interaction data. Without sufficient historical data, it is challenging to make accurate recommendations.

- **Popularity bias**: CF tends to favor popular items that have been interacted with by many users. This can result in a lack of diversity in recommendations, overshadowing less popular but potentially interesting items.

- **Interpretability**: CF algorithms, especially those based on MF or deep learning, often act as black boxes, making it difficult to understand and interpret why a particular item was recommended.

5.1.3   Hybrid approaches

Hybrid RSs combine CF and CBF to provide more accurate and diverse recommendations (Burke, 2002). By blending CF to capture patterns in user behavior with CBF to recommend new items based on their attributes, hybrid systems can overcome common challenges like the cold start problem and data sparsity. Common hybridization strategies include:

- Weighted hybrids, where CF and CBF outputs are combined with assigned weights.

- Switching hybrids, which switch between CF and CBF based on available data.

- Feature augmentation, where one method enhances the input to the other.

Pandora's Music Genome Project is an example of a hybrid approach, leveraging both CF and CBF by analyzing up to 450 musical attributes and refining recommendations based on user behavior Lops et al. (2011). This method allows for a more personalized and exploratory experience. Hybrid models like Pandora's are known for outperforming pure CF or CBF models, balancing the recommendation of familiar items with the discovery of new content (Burke, 2007).

5.2 Challenges in music recommendation

5.2.1 Data scarcity

Recommendation systems often face challenges related to data scarcity, which can be divided into two main categories: the cold-start problem and the long-tail issue. These challenges impact the ability of RSs to provide accurate and diverse recommendations, especially in domains like music streaming where users may want to explore both popular and niche content.

Cold-start problem

The cold-start problem (Schein et al., 2002) is a common challenge for any RS Lam et al. (2008); Lika et al. (2014), including music recommendation Ferraro (2019); Schedl et al. (2018). It arises when the system lacks sufficient information about a new user or item, making it difficult to provide personalized recommendations.

- **New Users**: When a new user signs up for a music streaming platform, the system has no prior knowledge of their preferences due to the absence of user-item interactions. To mitigate this, streaming services often use onboarding questionnaires to collect information about users' favorite genres, artists, or songs during the sign-up process. Demographic information, such as age, gender, and location, can also be used to suggest popular music among similar user groups. Additionally, popular and trending tracks are often recommended as an introductory mechanism.

- **New Items**: For new tracks or artists, the system faces a cold-start problem as there are no prior interactions with these items. This is particularly problematic for emerging artists and niche content, as their music is less likely to be discovered. To address this, content-based, or hybrid RS can be used to leverage the intrinsic properties of the new item (e.g., rhythm, timbre, or metadata) to generate recommendations. For instance, systems like Pandora's Music Genome Project use manually curated features to classify and recommend new songs.

Cross-domain recommendation is another strategy that aims to improve recommendations in one domain by transferring user preference information from another domain (Zhu et al., 2021). Some commercial systems have successfully implemented cross-domain recommendations. For instance, platforms like Amazon use cross-domain techniques to recommend products across categories (e.g., 'customers who bought this book also bought this movie') (He and McAuley, 2016). Despite the potential benefits, cross-domain recommendations come with challenges, as the success of these systems often depends on the degree of overlap between the domains.

Lastly, active learning techniques can be used to directly tackle the cold start

problem by eliciting specific user feedback to enhance the system's understanding of user preferences (Elahi et al., 2016). Active learning can involve soliciting ratings on items predicted to be of interest to the user or exploring novel ways to interact with users to gather valuable data. However, traditional active learning methods often introduce biases, as users may only rate items they are interested in, leading to skewed data. Despite these challenges, each approach provides valuable tools for improving recommendations in cold start scenarios, although they often need to be combined or enhanced with personalization techniques to address the inherent limitations of each method effectively.

Unpopular items

The long-tail issue affects items that are not new but are simply less popular, including niche tracks and emerging artists. These items are less likely to be recommended by traditional CF systems because they lack sufficient interaction data to be ranked highly. However, these items might be of interest to users looking to explore beyond mainstream content. As such, this is not just a cold-start issue but a continuous challenge for items with limited engagement. To address this, RSs need to balance exploration and exploitation — offering users both familiar and novel content. Previously discussed techniques like content-based or hybrid RSs can also be used in this case. Additionally, techniques like latent factor models (e.g., MF) can capture hidden relationships between users and unpopular items, improving recommendations for items in the long tail.

5.2.2   Context

In Chapter 3, we explored how music preferences can vary significantly depending on the context, such as the time of day, day of the week, season, activity, mood, or even the weather. Streaming platforms have recognized the importance of these contextual factors, leading to a growing number of efforts to develop context-aware MRSs.

However, creating these systems poses two main challenges. The first challenge is accurately detecting the user's context; while some factors like weather or time are relatively easy to determine as they are consistent across users, others like mood or activity are highly personal and much harder to infer. The second challenge is personalization: determining what a specific user wants to listen to in a given context. This is particularly complex because individuals have unique preferences and different ways of responding to various situations. The core goal of algorithmic recommendations is to tailor suggestions to these individual preferences, making context-aware recommendations a demanding yet crucial task in the evolution of music streaming services.

Over the past decade, researchers have explored various methods to predict user context and adapt music recommendations accordingly, enhancing the per-

sonalization and relevance of music streaming services. For example, Baltrunas et al. (2011) developed a RS that detects in-car contexts by leveraging sensors and user inputs to gather information such as driving speed, traffic conditions, and driver mood, subsequently tailoring music suggestions to improve the driving experience.

Similarly, Kaminskas et al. (2013); Cheng and Shen (2014) investigated location-based recommendations by utilizing GPS data from mobile phones to identify the user's geographical position and surroundings, enabling the system to suggest music that aligns with the cultural and environmental aspects of the location, such as playing upbeat tracks in urban settings or tranquil melodies in natural landscapes.

Further expanding on contextual factors, Wang et al. (2012); Schedl et al. (2014) proposed systems that incorporate a diverse range of contexts including time of day, weather conditions, and user activities, by collecting data from sensors on mobile phones; these systems dynamically generate playlists that adapt in real-time to the user's changing circumstances, aiming to provide a more engaging and contextually appropriate listening experience. However, a key limitation of these studies is scalability; they often rely on specific sensors and contexts that may not work well across all users and environments, making it challenging to implement these solutions broadly on streaming platforms.

In recent years, auto-tagging (Choi et al., 2016) has emerged as a powerful approach to context-aware music recommendation, using machine learning techniques to automatically assign context-specific tags to tracks based solely on their audio content. Ibrahim et al. (2020) proposed a method that leverages playlist titles to label tracks with contextual tags, creating a large dataset of approximately 50,000 tracks across 15 different contexts (e.g.: 'car', 'happy', 'workout'). Using this dataset, they trained a CNN to predict the context in which a track is most suitable. This method overcomes some of the limitations of previous context-aware systems by bypassing the need to access sensor data from mobile devices or user interactions, and instead focusing on the audio characteristics.

While many strategies have been developed to detect contexts suitable for playing certain tracks, accurately determining a user's current mood, activity, or environment based solely on interaction patterns or mobile device data remains challenging. As a result, many streaming platforms, including Deezer, resort to user-assisted approaches to enhance contextual recommendations. For instance, Deezer's 'Flow Moods' feature (Bontempelli et al., 2022) allows users to manually select their current mood from predefined categories such as 'Chill' or 'Motivation.' This selection then informs the algorithm, which curates personalized playlists using contextual scores attributed to songs through audio auto-tagging techniques.

5.2.3   Balancing discovery and familiarity

The primary objective of any RS is to assist users in discovering new items that may align with their interests, based on specific requests or past preferences. Most recommendation algorithms operate on similarity metrics, using the users' preferences as input and delivering the most similar items as output. The accuracy and precision of these systems are typically evaluated based on how closely the recommended items match the users' known preferences. However, this approach can inadvertently lead to the formation of filter bubbles (Pariser, 2011), where users are repeatedly exposed to the same types of items, limiting their experience to a narrow range of content. To prevent this and to keep the user experience fresh and engaging, it is crucial to diversify the recommended items. Yet, it is equally important to avoid recommending items that are too dissimilar from the users' usual consumption, as this can confuse users and discourage them from engaging with the platform. In the context of music, the challenge becomes even more complex, as listeners often expect to be recommended tracks they are already familiar with and enjoy. Therefore, a balance must be found between the familiarity and novelty of the recommended content, ensuring that users remain comfortable, yet stimulated by new discoveries.

Discovery and diversity

As previously mentioned, most classic recommendation algorithms are based on similarity metrics, thus plenty approaches exist to find items similar to the users' preferences. As the usage of search engines and RSs increased, researchers realised the need to diversify the output data, which triggered the development of different diversifying algorithms.

The earliest strategies focused on post-processing methods, where relevance was primarily determined first, followed by a re-ranking process to incorporate diversity. Maximal marginal relevance (Carbonell and Goldstein, 1998) was among the first to use a greedy algorithm to balance relevance and diversity by iteratively selecting items that maximize a combination of the two (Carbonell and Goldstein, 1998). Later, more sophisticated post-processing methods like determinantal point processes Kulesza et al. (2012) offered a probabilistic model to optimize diversity across the entire set of recommended items, considering global item relationships rather than just pairwise similarities (Kulesza et al., 2012).

Moving forward, in-processing methods started to incorporate diversity directly into the model training phase. For instance, Wasilewski and Hurley (2016) proposed incorporating diversity as a regularization term within the loss function during model training. This approach allowed the RS to learn to balance relevance and diversity simultaneously. Another example is the work by Chen et al. (2020), which introduced an intent-aware diversity method, adjust-

ing the relevance of items based on their contribution to the overall diversity of the recommendation list, ensuring a more balanced output.

In more recent years, pre-processing methods have also gained attention, focusing on structuring the input to the recommendation models to inherently promote diversity. For example, Zheng et al. (2021a) proposed sampling strategies that prioritize less popular items during the graph neural network-based recommendation, ensuring that these items are adequately learned and represented in the final recommendations. Kwon et al. (2020) categorized users into distinct types based on their purchasing behaviors, and developed hybrid recommendation strategies to ensure a diverse set of recommendations tailored to each user type.

When it comes to music consumption, novelty and diversity are particularly important as they significantly influence the listener's engagement and enjoyment. Preference for musical novelty is tied to the brain's reward system, where new and unexpected musical elements activate regions associated with pleasure and reward (Salimpoor et al., 2015). This explains why listeners often seek out new music or variations of familiar genres to maintain interest and excitement. Exposure to a diverse range of music genres and styles can enhance cognitive flexibility, allowing listeners to appreciate and adapt to different musical structures and cultural contexts (Krause et al., 2015). The need for diversity in music is also linked to the concept of 'optimal distinctiveness,' where individuals try to balance the need for belonging with the desire for uniqueness, often achieved through the exploration of varied musical experiences (North and Hargreaves, 1995). Therefore, diversifying music recommendations seems crucial for streaming platforms in order to increase user satisfaction.

While most of the diversification methods discussed earlier can be applied to MRS, research in the music domain tends to focus more on personalization than on specific diversification techniques. Indeed, users have varying levels of tolerance and demand for novelty, making it essential for MRS to consider individual preferences. For example, Schedl and Hauger (2015) propose to group users into different categories based on individual diversity, mainstreaminess, and novelty scores, and then applying different recommendation algorithms tailored to each group to enhance accuracy. Lu and Tintarev (2018) developed a re-ranking algorithm that adjusts the diversity of a recommendation list based on the user's personality traits, demonstrating improvements in both diversity and user satisfaction. In Robinson et al. (2020)'s exploratory user study, participants made a clear difference between 'inner diversity' (within existing preferences) and 'outer diversity' (introducing new, unfamiliar genres), a distinction not well captured by current diversity metrics. The same study suggests that the need for novelty and discovery can vary not only between users but also within the same user, depending on their mood and context. To address this, giving users the ability to control the level of diversity in their recommendations could enhance

their experience. Such interactive interfaces, which allow users to adjust the novelty or diversity in their music recommendations, have been explored by Jin et al. (2018); Millecamp et al. (2018); Knees et al. (2020), offering promising solutions for aligning recommendations with individual user preferences.

Familiarity and repetition

It is not uncommon to listen to the same music several times within a month, a week a day or even an hour. Berlyne (1973)'s inverted U-shape theory suggests that as familiarity with a song increases, so does the listener's enjoyment, up to a certain point. Beyond this peak, further exposure may lead to a decline in preference due to over-familiarity. This pattern indicates that initial exposure to a novel song is often not the most enjoyable experience; instead, repeated listens are necessary for the song to fully resonate with the listener (Hargreaves, 1984). Moreover, familiarity has been shown to significantly influence music listening behavior. Ward et al. (2014) demonstrates that listeners tend to prefer familiar songs over unfamiliar ones, even when they express a desire to explore new music. This consumption pattern differentiates music from other types of content — people rarely re-watch the same movie or repeatedly purchase the same item online within a short time frame. Consequently, MRS must adapt to this specificity by finding a balance between recommending familiar and new songs.

Garcia-Gathright et al. (2018) conducted a comprehensive study to understand user satisfaction with music discovery on streaming platforms, particularly focusing on the role of familiarity in recommendations. Initially, they conducted face-to-face interviews with 10 participants, which revealed that users have varying goals when interacting with music recommendations, such as finding new music or listening to familiar favorites. These insights were further validated through a large-scale survey of 18,547 users, which measured satisfaction with both overall music recommendations and specific weekly playlists. The survey revealed that users were significantly more satisfied when the recommendations included at least one familiar track they loved, with 74.2% of those users reporting high satisfaction, compared to only 29.0% among those who did not find any familiar tracks they loved. The analysis of user interactions with the platform, such as saving tracks, skipping, and downstream listening, validated the insights gained from the user interviews and surveys.

Building on the idea that repetition of familiar tracks enhances user satisfaction, Manolovitz and Ogihara (2021) extend this concept to the repetition of new, unfamiliar songs as a critical factor in helping users develop a liking for them. Through an experiment involving 19 Spotify users, the authors investigated how repeated exposure to new tracks influences user engagement and retention. Participants were given playlists containing a mix of both new and repeated songs from their 'Discover Weekly' selections, and their listening be-

haviors were tracked during six weeks. The study found that just one additional listen to a new, initially liked song increased the likelihood of the user revisiting it by 10%. This finding demonstrates that a single exposure to a novel song is often insufficient for it to fully resonate with the listener, and repetition is key to fostering a connection between users and new music.

To conclude, no study, to our knowledge, has yet explored the optimal proportion of familiar songs or the frequency of repetition needed to maximize user satisfaction in MRSs. This remains an important area for future research, as understanding these factors could significantly enhance the effectiveness of personalized music recommendations.

5.2.4   Fairness

Fairness in RSs focuses on ensuring that the algorithms and models used do not introduce or perpetuate biases, providing equitable recommendations to all users. It is crucial to maintain users' trust, promote diversity, and ensure that all demographic groups are fairly represented and served by the recommendations. In recent years, this has become a growing subject of research, driven by the increasing awareness of the ethical implications and social impact of recommendation algorithms (Wang et al., 2023).

In the realm of MRS, fairness is an important issue due to the impact these systems have on both users and artists. The primary stakeholders involved in MRS fairness include platform users, item providers (artists and their labels), and the music streaming platforms themselves (Dinnissen and Bauer, 2022). For users, fairness typically involves ensuring that recommendations are unbiased and equitable across different demographic groups. Item providers, often the artists, are affected by how frequently their music is recommended, which can influence their exposure and revenue. Finally, the platforms, which facilitate the interaction between users and artists, must balance the interests of both parties to maintain a fair ecosystem.

Various biases can affect the fairness of MRSs. Popularity bias is one of the most common, where popular songs are recommended more frequently, potentially marginalizing lesser-known artists. Demographic biases also exist, where the quality of recommendations may differ based on user characteristics such as age, gender, or country, often resulting in a preference for mainstream users or certain demographic groups. Gender bias specifically affects both users and artists, with male artists typically receiving more recommendations than female artists, and women users experiencing lower recommendation quality (Shakespeare et al., 2020; Lesota et al., 2021). Understanding and addressing these biases is crucial for developing fairer MRSs that can cater equitably to the diverse needs of all stakeholders involved.

Artists

A lot of existing studies on MRS fairness focus on artists and more generally music distributors, in particular because their revenues are directly correlated with their exposure and popularity on streaming platforms.

Several qualitative studies aim to understand the artists' perception of the fairness of music streaming platforms, particularly focusing on how these platforms and their embedded RSs impact artists.

For example, Ferraro et al. (2021b) conducted semi-structured interviews with 9 music artists from different countries and levels of popularity. The study revealed that artists feel their profiles are often inadequately presented on music platforms, with a lack of contextual information and a bias towards older tracks. Artists expressed difficulty in reaching new audiences due to the prevalent popularity bias in recommendation algorithms, which favor more established artists. There was a unanimous call for greater transparency in how these algorithms function and why certain music is promoted. Some artists suggested implementing quotas for local music to help lesser-known artists gain visibility. Opinions were divided on whether larger repertoires should result in more frequent recommendations. Most artists preferred that new releases be promoted more heavily, emphasizing the importance of balancing the discovery of new artists with the promotion of existing popular tracks.

A similar study with 14 artists from the Netherlands (Dinnissen and Bauer, 2023) also highlights artists' calls for greater transparency in algorithmic functioning and a more equitable approach to promoting new and lesser-known artists. Both studies provide valuable insights into the artists' perspectives on MRS fairness, however the small sample sizes restrict the generalizability of the findings to the broader artist community. Additionally, while the samples included diverse nationalities and popularity levels, they may not fully capture the wide range of experiences and genres present in the music industry. Moreover, the participants' awareness of the interviewers might have influenced their responses, potentially biasing the findings.

Some studies with computational methodologies seem to converge with the artists' perceptions about the popularity bias in MRS, especially those based on CF techniques. For example, Celma and Cano (2008) show that item-based CF is positively biased towards popular artists, while CBF exhibit more diverse connections between artists of different popularity levels. Kowald et al. (2020) applied user-based, item-based and MF-based CF algorithms on the LFM-1b dataset and show that all of the methods tend to predominantly recommend popular artists, as evidenced by the positive correlation between artist popularity and recommendation frequency. Turnbull et al. (2022) show that among three algorithms, SLIM (Ning and Karypis, 2011), Multi-VAE (Liang et al., 2018) and WRMF Hu et al. (2008), applied also on LFM-1b, the most accurate model (SLIM) has the most popularity bias while less accurate models have less

popularity bias.

On the other hand, a simulated user experiment in the same study, based on data from Spotify, Amazon Music, and YouTube Music, shows little to no bias in commercial MRS.

In this experiment, twelve simulated user accounts were created on each streaming service, based on real user data from the Last.fm 1-Billion dataset. These accounts were divided into three categories (low, medium, and high mainstream users), with four accounts per category. Each simulated account followed and listened to the top ten most-listened-to artists for their respective user group. The protocol involved listening to the top songs of these seed artists, after which the accounts were logged out and returned to the next day to analyze the given recommendations. Recommendations generated by the streaming services were collected by sequentially noting the top artists from each generated mix (e.g., Daily Mix on Spotify) until ten recommended artists were recorded for each simulated user account. The average popularity of recommended artists was measured using Spotify's proprietary score and Last.fm user data.

The comparison between average popularity of recommended artists with that of artists in the user profiles showed minimal to no bias. Even though this experiment raises an important question about whether simulated 'sterile' recommendation experiments can accurately represent real-life scenarios, the proposed methodology is still far from reflecting the complexity and nuances of actual user interactions and behaviors on commercial streaming platforms. Indeed, the relatively small scale and simplified nature of the simulated user profiles may not fully capture the complexity of real user interactions with streaming services. The study also relies on a short-term dataset, whereas commercial services utilize extensive long-term user data and advanced algorithms to refine recommendations.

Apart from popularity, gender biases have also been a growing concern. For example, Shakespeare et al. (2020) seeks to determine if CF algorithms influence the exposure and representation of male and female artists differently. They conduct two separate experiments, both performing CF algorithms (UserKN-NAvg (Desrosiers and Karypis, 2011) and Non-Negative MF (Lee and Seung, 1999)) and selecting top 5 artists on the LFM-1b and LFM-360k datasets. The first experiment was run on a representative set of users, the second — on users with extreme preferences towards artists of a specific gender. Then, authors compared proportion of artists of different genders in the initial users' preferences and their recommended music.

In both experiments male artists tended to receive more recommendations than female artists, with a more pronounced bias in datasets with higher initial preference towards male artists. Several factors, mostly data related, can be at the origin of the observed gender bias. First, the researchers were only able to identify the gender for about 27-31% artists in the dataset, and approximately

82% of the identified artists were labeled as males. Moreover, a majority of the most popular artists were labeled as males (85% of the top 20% most popular artists), consequently, male artists receive more play counts and are thus more likely to be recommended by the algorithms. This creates a feedback loop that perpetuates the visibility and dominance of male artists in recommendations. Additionally, as CF algorithms are known to emphasize popularity bias, they can indirectly extend gender related bias.

Ferraro et al. (2021a) goes beyond simply exploring how gender biases manifest in music platforms and propose methods to mitigate them. In 'Break the Loop: Gender Imbalance in Music Recommenders,' the authors highlight how CF algorithms, like ALS, contribute to the uneven exposure of male and female artists. Through interviews with artists, the study uncovers a strong consensus for the need to address gender bias by promoting more equitable exposure of female artists. Ferraro et al. (2021a) introduce a progressive re-ranking method designed to gradually increase the visibility of female artists in recommendation lists, addressing the feedback loop that favors male artists. Their approach emphasizes a gradual shift to avoid user pushback while maintaining recommendation accuracy. By simulating feedback loops, they demonstrate that such re-ranking can improve gender balance over time without significantly impacting the quality of recommendations.

Finally, some recent attempts have been made to evaluate algorithmic biases on local music recommendation. Lesota et al. (2022) applied two algorithms, ItemKNN (Sarwar et al., 2001), and item-based CF algorithm, and NeuMF (He et al., 2017), a neural network based MF algorithm, on users from n different countries from the LFM-2b dataset. According to the study, item-based algorithms promote local artists, while MF algorithms reinforce popularity biases and thus promote U.S. music as it is the most streamed music in all countries in the dataset. However, the dataset and methodology used in the study present several problems, that we will debunk in Chapter 7.

Users

An equally important aspect is the fairness to users, which is particularly of interest for streaming services as users' satisfaction with the music recommended to them can be directly correlated with their engagement and retention on the platform. Fairness to users typically implies providing equally qualitative recommendations to users of different demographics (e.g. gender, age, country) and different music consumption profiles (e.g. different levels of mainstreamness or omnivorism).

In terms of demographics, several studies focus on the impact of MRS on users of different genders. For example, two studies from the same group of researchers, Melchiorre et al. (2021) and Lesota et al. (2021), investigate how male and female users are affected by algorithmic popularity bias. The researchers

evaluated several CF algorithms, such as ItemKNN, ALS, BPR (Rendle et al., 2009), SLIM, and MultiVAE (Liang et al., 2018), to assess their performance across male and female user groups from subsets of the LFM-2b dataset. They measured bias by quantifying performance disparities between user groups.

Their findings revealed significant gender bias, with most algorithms favoring male users, particularly in terms of accuracy metrics. Interestingly, the most accurate algorithm, SLIM, showed the greatest unfairness, whereas less accurate algorithms, like BPR, showed more homogeneous results across gender groups. The substantial difference in user representation, with 71.5% male users in Lesota et al. (2021)'s dataset and 77.9% in Melchiorre et al. (2021)'s dataset, is probably the main cause of the observed gender bias in RS performance, as algorithms tend to be optimized for the majority user group.

Authors proposed a debiasing strategy, involving resampling female users' data to equalize gender representation. Specifically, they increased the representation of female users by duplicating their data points until they matched the number of data points for male users. This strategy allowed to slightly improve the fairness for female users, without compromising overall accuracy.

Ekstrand et al. (2018)'s work, later reproduced by Neophytou et al. (2022), investigates not only gender, but also age bias of CF algorithms, using the MovieLens and Last.fm datasets. The study finds that younger users, particularly those in the 18-24 age group, often receive more accurate recommendations compared to older age groups. One possible reason for this bias is the over-representation and higher activity levels of younger users in the datasets, providing richer interaction data for algorithms to learn from. Additionally, the algorithms may inherently favor popular items, which align more closely with the preferences of the larger, younger demographic.

Several studies show evidence of significant variations in users' representation and music consumption between countries, which can be at the origin of algorithmic recommendation bias on music streaming platforms (Bauer and Schedl, 2018, 2019; Neophytou et al., 2022). Neophytou et al. (2022) show, on the LFM360K dataset, that users from countries with higher representation generally receive more accurate recommendations, similarly to users of different age or gender. Bauer and Schedl (2018, 2019) propose measures for local and global mainstreamness in the LFM-1b dataset, and find that different countries exhibit quite unique characteristics regarding what music is considered popular or mainstream.

Bauer and Schedl (2019) finds that countries with listening habits closely aligned with the global mainstream generally received more accurate recommendations compared to those with distinct local music tastes. In order to mitigate this bias, authors proposed to tailor the CF process by integrating country-specific mainstreamness models into it. They developed several mainstreamness measures using artist play counts and artist listener counts, applied both globally

and country-specifically. CF was then adjusted to focus on nearest neighbors within the same mainstreamness level and country, ensuring that recommendations were tailored to users with similar local preferences. This approach improved recommendation accuracy by reducing bias and better capturing regional music tastes, compared to applying the same recommendation algorithm on users altogether.

Study by Melchiorre et al. (2020) moves beyond simple demographics, and finds that music recommendation algorithms exhibit biases based on users' personality traits. Unlike gender, age, or country, personality traits are challenging to gather at scale on streaming platforms, making it difficult to study and address these biases. Nonetheless, it's important to recognize their existence, as personality traits are known to significantly influence music preferences Rentfrow and Gosling (2003).

Another thing that varies from user to user is their behaviour on the music platform, including their music preferences, which can be for example more or less mainstream or diverse, or the intensity of their interactions. Because of the long-tail distribution of artists, it is natural to expect CF algorithms to differently affect users with more or less mainstream preferences.

Kowald et al. (2020) work with a subset of 3000 users from the LFM-1b dataset and split them into three equal size categories based on their mainstreaminess score. This score is calculated as followed: the popularity of an artist is defined as the proportion of users in the dataset who have listened to that artist, and for each user, the mainstreaminess score is the average popularity of all the artists that the user has listened to. Low-mainstreaminess users (LowMS) were found to listen to at least 20% of the 80% least popular artists; medium-mainstreaminess users (MedMS) had scores around the median; and high-mainstreaminess users (HighMS) had the highest mainstreaminess scores.

Then the study tested six recommendation algorithms to evaluate the presence and degree of popularity bias in these different types of users. Random recommendation was used as a baseline, and performed equally for all user types. Expectedly, the MostPopular algorithm displayed the strongest bias, consistently recommending popular artists and significantly disadvantaging LowMS users. UserItemAvg and the KNN-based approaches (UserKNN and UserKNNAvg) showed moderate bias, also leaning towards popular artists, with LowMS users receiving consistently poorer recommendations. Non-negative MF performed the best for LowMS users compared to other algorithms.

Interestingly, for most of the tested algorithms, the accuracy generally followed a pattern where MedMS users received the most accurate recommendations, followed by HighMS users, and then LowMS users. This is not necessarily intuitive, as we would expect an algorithm that reinforces popularity bias, like MostPopular for example, to perform better for users with the most mainstream taste. However, as it is not the case, we can make the hypothesis that

the high accuracy for MedMS users is due to the fact that most algorithms perform better for the average users, and not for those who simply prefer the most popular artists.

In another study by Li et al. (2021), based on ratings of both music and videos from Amazon, authors indirectly categorized users into mainstream and non-mainstream groups based on the distribution of recommendation accuracy, more specifically root mean square error scores across users. The bins were distributed as follows: the top 10% of users with the lowest scores were considered mainstream, as the recommendations were highly accurate for them; the next 40% were a mix of mainstream and some non-mainstream users; the following 40% primarily consisted of non-mainstream users with higher scores, indicating less accurate recommendations; and the bottom 10% were the most non-mainstream users, whose preferences were the hardest to predict, resulting in the highest scores.

This method of categorizing users raises some concerns, as it is doubtful that a direct parallel can be drawn between accuracy and mainstreaminess, especially considering that some studies like the previously discussed paper byKowald et al. (2020) show that the best accuracy concerns average users, and not the most mainstream ones.

Despite these concerns, it is still worth discussing the debiasing method proposed in the study. This method is based on autoencoders, which are neural networks that learn to compress input data, like user reviews, into a smaller representation (encoding) and then reconstruct the original data from this compressed version. As they focus on the reconstruction of original input data, rather than relying solely on co-occurrence patterns typically used in CF, autoencoders might help reduce bias for less mainstream users by ensuring that the unique details of their preferences are preserved. The model outperformed traditional methods like MF and DeepCoNN (Zheng et al., 2017), significantly enhancing recommendation accuracy for the bins of users authors labeled as less mainstream, while maintaining or only slightly reducing accuracy for mainstream users. While it remains uncertain whether the user groups in this study truly correspond to varying levels of mainstream taste, the proposed debiasing method appears effective in adjusting recommendation scores for users with different music consumption patterns.

In addition to the popularity of the music consumed, the size of a user's streaming history or music library can also affect recommendation quality. We know that new users, who have had little interaction with the platform, face the cold-start problem, which intuitively leads to the assumption that the more we know about a user, the better the recommendations we can make. However, Ekstrand et al. (2018) finds a negative relationship between the size of a user's profile and recommendation accuracy. Authors suspect that users with more items in their profile have already rated the 'easy' items, so recommending for

them is a harder problem. Supposedly, this could also be related to the diversity of the user's musical preferences. So-called omnivores, who have larger and more varied libraries, might confuse recommendation algorithms due to the heterogeneity of their tastes. Unfortunately, to our knowledge, no studies have specifically examined bias in recommendations for univore versus omnivore music listeners, but this is undoubtedly an area worth exploring.

Balance between stakeholders

In an ideal scenario, MRS should to be equally fair both to users and to music providers, however, aligning the two can be challenging. Approaches to maximize user satisfaction can lead to unfair exposure for less popular item providers. Conversely, approaches which aim to ensure equitable exposure for all item providers can reduce user satisfaction because the recommendations may not align closely with individual user preferences.

Several strategies can be applied in order to find a middle ground that would satisfy both users and artists specifically on music streaming platforms Mehrotra et al. (2018). One strategy consists in balancing relevance and fairness by assigning weights to each factor, allowing the system to introduce fairness without significantly compromising relevance. Another way is to incorporate a degree of randomness by probabilistically choosing between prioritizing relevance or fairness in each recommendation. This method aims to maintain overall user satisfaction while ensuring fairer exposure for less popular items. Also, a minimum threshold of relevance can be fixed before considering fairness, thereby ensuring user satisfaction while still promoting fairness. Finally, recommendations can be personalized based on the user's individual tolerance for fairness, tailoring the balance between relevance and fairness to each user's preferences.

Part II

Personal Contribution

Introduction

At the beginning of this thesis, we formulated two main research questions to which we aim to contribute :

- **How can we model musical taste, both to better understand it and to use it for recommendation purposes?**

- **How can we measure the influence that recommendation systems may have on musical preferences?**

These questions are not only central to understanding individual and collective music consumption patterns, but they also have broader implications for the development of fair, transparent, and diverse RSs. The way musical taste is modeled impacts how users are exposed to new music, which in turn shapes their long-term preferences and listening habits. Similarly, the second question addresses an important ethical dimension: RSs have the power to reinforce or mitigate biases, both in terms of promoting mainstream content over local or niche music, and in how they influence the visibility of underrepresented artists and genres.

The preceding chapters have laid the groundwork for addressing these questions. We have begun by exploring the nature of streaming data, highlighting its unique features such as the large volume of user interactions and the diversity of musical items available. Then we have surveyed literature on musical taste across multiple disciplines, identifying how preferences can be measured—from cognitive psychology to social science perspectives. Additionally, we have examined methods for labeling and categorizing music, as well as representing the vast space of musical items in ways that can inform recommendation algorithms. Finally, we have discussed RSs, and specifically how they can influence music consumption, potentially introducing biases or amplifying existing preferences.

With this context established, we now turn to our contributions to these research questions through two specific studies I conducted, both of which culminated in published papers.

Chapter 6

Modeling music preferences from streaming data

At the outset of my PhD journey, our goal was to study the dynamics of musical taste — how preferences evolve over time and how geographical factors may influence them. As we began exploring data from Deezer, we quickly realized the volume was overwhelming. To meaningfully capture a user's preferences at a given moment, we needed a proxy — a summarized representation of their musical taste. This raised an important question: what exactly can be considered as an individual's 'musical taste' within all this data? How do we define and measure it?

We were inspired by a paper published in a very different field, De Montjoye et al. (2013)'s 'Unique in the Crowd', that is concerned with individual human mobility, as it can be measured and understood from ICT-based data, such as mobile phone data, transportation cards, georeferenced posts on social media platforms, etc. . Based on the location data of 1.5 million people over 15 months, obtained from a mobile phone operator, the authors of this study found that 95% of individuals in the dataset could be uniquely identified with only four spatiotemporal points, sampled at random in the location history data. Such a set of points can be seen as some kind of 'fingerprint', a unique set of information that is personal, distinctive, and possibly representative of one's mobility : even though they do not focus on the nature of the locations, authors touch on the idea that certain significant places, like home or work, can be frequented more often, making them highly identifiable.

This made us wonder: could we apply this concept to music streaming? Could we find a set of music items — artists or songs — that uniquely identify a user? And if so, might this unique combination of items reflect their musical taste? Drawing on Lahire (2008)'s definition of musical taste as a unique blend of influences and experiences accumulated through one's life course, we aimed to define a 'musical fingerprint' that could reflect a user's distinct taste. For

example, when I thought about my own personal music habits, I realized there probably aren't many Deezer users who listen to a mix of Russian pop, French rap, and Algerian blues, alongside some doom-metal from my teenage years and neo-soul from more recent times (here you can see a bit of Nault et al. (2021)'s snobivorism at play). So, I imagined that my musical fingerprint would reflect this unique combination of genres.

Since streaming history involved a large amount of fluctuating data and likely included music that users didn not always like, we initially focused on 'likes', specifically artist likes, assuming they better reflected core preferences — almost as if the users had done a pre-selection for us (later, we incorporated liked songs and then streaming data to explore user identifiability for anonymization purposes, though this is beyond the scope of this discussion). Our first goal was to understand how unique people's preferences actually were.

We developed a greedy algorithm designed to find the smallest combination of favorite artists that could uniquely identify a user. Surprisingly, we discovered that users could often be identified with a very small number of artists — on average, two artists in a dataset of 1M users, and even fewer when using liked songs or streams. However, this did not feel like a true 'fingerprint', representative of such a complex thing as a user's musical taste. The issue lays in the long-tail distribution of item popularity: liking one or two obscure artists was often enough to make a user identifiable, but these niche artists did not necessarily reflect the user's broader musical preferences — or at least we could not detect it, as most of these unpopular artists lacked genre labels/tags.

This limitation led us to rethink our approach. Instead of focusing on the smallest unique set of items, we aimed for a set that would be more representative of a user's preferences, regardless of its size. After all, our original vision for the fingerprint was to track the evolution of musical taste over time, compare preferences across geographical regions, and potentially use this information for recommendation purposes. Imagine, for example, how practical would it be if we could transfer a well-chosen set of artists when switching streaming platforms, rather than spending weeks retraining a new RS through extensive listening.

This shift in focus led us to adopt an approach very close to a recommendation setup. We aimed to find a set of items that could be used to recover, or predict, the user's known broader preferences through CF, i.e. user similarity. Do the two approaches lead to convergent or diverging solutions, i.e. are the items that best represent a user's preferences the ones that make them unique?

This work resulted in a paper entitled ***"Depict or Discern? Fingerprinting Musical Taste from Explicit Preferences"***, published in the *TISMIR journal* in January 2024.

The study was co-authored by Manuel Moussallam (Deezer Research), Thomas Louail (CNRS, Géographie-cités), and Olivier Bodini (Université Sorbonne Paris

Nord). We collaboratively designed the study. I conducted the data analysis and experiments. The manuscript was drafted by myself, Manuel Moussallam, and Thomas Louail. All authors contributed to the final version of the paper.

This research contributes to the broader literature on musical taste by examining it from two perspectives: as something unique and distinctive of each individual (in line with Lahire (2008)'s definition), or as something shaped by shared patterns and similarities within a group (similarly to Bourdieu (1984) or Peterson (1992)). We propose data structures and algorithms for both views, linking these different perspectives on musical taste to practical algorithmic recommendation.

Depict or Discern?
Fingerprinting Musical Taste
from Explicit Preferences

KRISTINA MATROSOVA [iD]

MANUEL MOUSSALLAM [iD]

THOMAS LOUAIL [iD]

OLIVIER BODINI [iD]

*Author affiliations can be found in the back matter of this article

ABSTRACT

The notion of personal taste in general, and musical taste in particular, is pervasive in the literature on recommender systems, but also in cultural sociology and psychology. However, definitions and measurement methods strongly differ from one study to another. In this paper, we question two different views on taste that can be retrieved from the literature: either something that is *distinctive* of an individual, or something that *essentially captures* the extent and diversity of their preferences. Relying upon a dataset that contains the complete list of musical items liked by individual users of a streaming service, as well as streaming logs, we propose two methods to compute *fingerprints* of their musical taste. The first one explicitly targets a *uniqueness* property, aiming at selecting items that uniquely identify a user in the crowd. The second approach focuses on a *representativeness* task that is fundamental in recommendation, i.e. building a summary depiction of the user's preferences that can be leveraged to propose other items of interest. We demonstrate that the two methods lead to conflicting solutions, hence highlighting the need to precisely acknowledge which point of view applies when addressing a computational question related to taste. We also raise the question of users' identifiability through their online activity on music streaming platforms, and beyond.

1. INTRODUCTION

An increasing proportion of people rely upon streaming services to listen to music, and large amounts of detailed, individual data collected by these services are becoming available to scientists. These data open the door to an improved understanding of spatial and temporal dynamics related to music consumption, such as the long-term evolution of people's listening behavior through the course of their life, or the geographical spread of different songs, artists and music genres at different periods of time. However, in order to study such high-level dynamics, it is necessary to have quantitative tools that are able to capture and expressively summarize these enormous amounts of listening data and musical preferences produced by millions of users.

Quantitative research on people's musical taste spans over many scientific disciplines. From a sociological standpoint, musical taste has been long studied as a self-declared, differentiating feature among individuals and social groups (Bourdieu, 1984; Peterson, 1992; Bryson, 1996). Psychological studies have been investigating correlations between musical preferences and personality traits (George et al., 2007; North, 2010). More recently, the concept has been used in the music recommender systems literature, as the distinctive part of the musical space from which a user is likely to enjoy a recommendation (Laplante, 2014; Ferwerda and Schedl, 2014; Uitdenbogerd and Schyndel, 2002). While in its general understanding, musical taste is an individual's set of musical preferences, when it comes to the literature we observe conflicting approaches that can be broken down into three dichotomies. The first one lies in the empirical data supporting the research – declarative information collected in questionnaire surveys or interview-based research, versus interaction traces assumed as implicit and explicit preferences that can be retrieved from online activity logs. The second dichotomy is related to the "resolution" of the information at hand: either aggregated (generally at the level of music genres), or directly at the "atomic" level of musical items, namely songs, albums and artists. Finally, the third dichotomy of musical taste is the focus on either its distinctive features – what in their taste makes individuals or groups different from one another? – or the focus onto its essence – what, among an individual's appreciations, best sums them up?

In recommendation, usage data and explicit preferences collected by platforms are used to derive average "taste profiles" from which new items can be sampled and proposed. There are also examples of recommender algorithms that treat each user as a mixture of profiles (Vargas and Castells, 2013), or which use contextual cues to modulate recommendations (Liang et al., 2018). This is somehow a reductionist vision of what makes personal taste, as it assumes that it can be summarized. It can also be said that it is an operational

definition that basically reverts the problem of providing a comprehensive definition: in a recommendation setting, taste is what can be leveraged to make relevant recommendations. It is also interesting to notice that it is not consistent with the relational approach that is used in sociology, where taste and distaste have traditionally been represented as a set of preferences that distinguish one social group from another – social groups being constituted on the basis of the economic, educational and cultural capital of individuals (measured through variables such as their occupation, their parents' occupations, or the highest degree they obtained). In the end, practitioners of both fields share the common objective to capture what distinguish people when it comes to their musical preferences. People engineering recommender systems are more interested in building systems able to predict items that people will like, while sociologists of taste are interested in finding what are the variables that best explain social differences in taste and distaste. Both are interested in building a system able to summarize and predict an individual's musical preferences.

Getting back to the tools required to study high-level dynamics of music listening in societies, it would be extremely useful to be able to capture some kind of "fingerprint" of an individual's musical preferences. From a computational perspective, a good fingerprint should possess different desirable properties. It should be expressive, and provide a good summary of the diversity of the music appreciated by the user. It should also be concise, i.e. be composed of as much information as necessary but not more. Most of all, it should be able to serve as a fingerprint, i.e. a signature able to identify a user among others. These properties may prove to be difficult to achieve simultaneously via a single fingerprinting procedure, and in the remainder of this paper we will investigate this question experimentally.

More precisely, we are interested in formalizing and comparing different views of taste, and in order to do so we will formalize these views in a fingerprinting problem, that is, an information summarization problem that we will study by considering two distinct sets of constraints. The first set is designed to capture a user's identity, in the sense of its identifiability among others. Identifiability through music is also a topic of interest for privacy purposes: with explicit preference data being ubiquitous on the open internet, measuring to what extent individuals can be uniquely identified through their portfolio of content preferences is important. We will try to answer the following questions:

RQ1: To what extent are users identifiable through their online activity data (favorite items and streaming history)?

RQ2: What information (content and size) is needed to identify people?

We wish to answer these questions by assigning users a so-called fingerprint – a small set of items that allows us to identify users in a unique way.

The second set of constraints is expressed as a representativeness problem, i.e. finding the essence of one's preferences. We will adopt a data-driven approach, and propose one formalization of what a taste fingerprinting procedure could be, similar to a classic recommendation setup, and evaluated through a prediction task. We will then confront the two sets of constraints, in order to answer the following question:

RQ3: Are the items that make one's preferences unique representative of these preferences?

In our experimental setup, we will use a dataset containing the explicit preferences (e.g. artists and songs that have been deliberately *liked* by users, by clicking on a heart-shaped icon) of about 1M users of a music streaming platform, as well as liked and streamed artists for another 50K users.

The remainder of the paper is organized as follows: in the next section we provide an overview of the previous work in social science and recommender systems related to the measure of the notion of "musical taste". Section 3 presents the data, while sections 4 and 5 present the experiments we conducted and the results we obtained for the fingerprinting problem with the two different sets of constraints. Section 6 concludes the paper.

2. RELATED WORK

In order to measure and quantify musical taste, we need to understand all the aspects that this term can describe. In this section, we make an overview of characteristics necessary to study musical taste through three axes. First, we dive through existing ways of collecting data. Then, we discuss different representations of music. Finally, we overview two diverging views of musical taste found in the literature – as an attribute of distinction among others, or as a set of characteristics of our preferences.

2.1 MUSIC PREFERENCE DATA COLLECTION
In sociology and psychology, collecting declarative data about musical preferences and consumption habits through surveys and interviews is common. Interacting directly with the respondents is advantageous for several reasons. The use of a Likert scale for instance allows to have a deeper understanding of how much respondents do or do not like certain music (Peterson, 1992; Bryson, 1996). Information about context of music consumption can be collected (DeNora, 2000), as well as sociodemographics, that can then be crossed with declared music preferences (Bourdieu, 1984; Peterson,

1992; Bryson, 1996; Coulangeon, 2017; Lahire, 2008). However, the sample of surveyed individuals is usually limited, and the results can be biased as such surveys are often run either in a specific country, or on a specific social group, like students for example (Delsing et al., 2008; Brown, 2012; Langmeyer et al., 2012). Additionally, the respondents may find it difficult to realistically assess what music they like to listen to and in what proportion. Flegal et al. (2019) show that some people struggle to estimate their own weight, and we can imagine that there might be a gap between declared preferences and the music that respondents actually listen to. For instance, it is possible that people tend to overstate listening to some more socially appealing music genres, and neglect to mention the less socially accepted music they like.

On the other hand, recommender systems mostly rely on observable data, often collected as traces of activity in online platforms. The huge amount of collected data should allow a good understanding of the users' listening practices, and even though the context or sociodemographics are not explicitly collected, the data could be used to deduce some implicit information. For example, Way et al. (2019) estimate the relocation of certain users by analysing the changes in their IP address. However, the collected traces are often ambiguous and considered as implicit markers of preference (or negative markers, in the case of skipped songs for example) (Oard and Kim, 1998; Majumdar et al., 2009).

A way to have a complete understanding of people's preferences would be to cross observable and declared data. This idea has been recently proposed (Cura et al., 2022) in the form of "augmented interviews" leveraging digital traces to inform and assist social science researchers conducting interviews.

2.2 MUSIC REPRESENTATION
In order to quantify musical taste, one must first be able to segment the musical space itself. For this, music preferences can be assessed either directly using music items, like artists or songs (Bourdieu, 1984), or through the mediation of aggregated categories. In surveys, for the sake of brevity, preferences are often collected via set of music genres (Peterson, 1992; Bryson, 1996; Coulangeon, 2017). Even though representing music through genres may seem obvious, it is important to keep in mind that no universal genre taxonomy exists, thus using genres to depict people's musical taste can create bias (Sordo et al., 2008). Music can also be classified by so called "mood", that can be identified either through audio features (Soleymani et al., 2015; Delbouys et al., 2018) or through declared data (Rentfrow and Gosling, 2003). Bogdanov et al. (2013) used audio features in order to depict people's musical taste.

2.3 DISTINCTION AND ESSENCE

In the literature, taste is often defined as a set of traits that distinguishes us from others and marks our individuality. In sociology, musical taste has long been studied in association with social class belonging. Bourdieu (1984), Peterson (1992), Bryson (1996) and Coulangeon (2017) show the connection between musical preferences and social class – people present their taste as a mark of belonging to their "in-group" while differentiating themselves from an "out-group". Similar conclusions have been found in psychological studies, like Hargreaves et al. (2006), who studied adolescents and how they use music to build their social identity. Later, Lahire (2008) studies intra-individual behavioral variations and emphasizes that most people have preferences that are not typical for their social group, and thus taste is an individual characteristic. The need people have for distinctiveness or "uniqueness" in order to self-identify has been studied in psychology as well (Fromkin and Snyder, 1980).

The notion of distinctive identity is also reminiscent of that of digital identifiability, that is, to what extent people's behavior (and digital traces of it) can be used to uniquely identify individuals. For example, De Montjoye et al. (2013) use mobility data and show that four spatio-temporal points are enough to uniquely identify 95% of individuals. Narayanan and Shmatikov (2008) use the Netflix Prize dataset to de-anonimize users through the movies they have watched on the platform. They show that 5–10 movie ratings are enough to identify most users. These studies present an extreme form of distinction, where each individual is literally identified in a unique way among all others. However, no such experiment has been run on music streaming data.

An alternate definition of musical taste would be a set of factors that characterize listening behavior of an individual. This is typically the definition implicitly adopted as the core principle for designing recommender systems, where the goal is to understand the essence of the user's preferences in order to suggest them similar music. Two main approaches exist in recommender systems. In collaborative filtering, the idea is to assign users a descriptive vector, or embedding, based on similarities between other users. The same process is applied to determine the similarity between items. This can be done through matrix factorization (Koren et al., 2009) based on either implicit feedback, like streaming activity logs, or explicit feedback such as users' collections of favorite items and playlisting of songs. Content-based recommendation, on the other hand, tries to define the items' features that a user will respond positively to. These features can be represented by various tags (Pazzani and Billsus, 2007) that can be automatically computed based

on audio features in the case of music (Cano et al., 2005; Van den Oord et al., 2013; Schedl et al., 2015) or social tags that can furnished by music providers or collected from the Web (Eck et al., 2007).

As concluding remarks, one may point out that the literature is rich with attempts to characterize musical taste, but they seem to be hard to reconcile, as they diverge on several key aspects. The first one is quantization of the musical space, the second being the data collected and the analysis methods. But most importantly there are conflicting hypothesis on the very nature of an individual's musical taste. While social sciences emphasize the importance it bears in the construct of one's self-identity, the emerging field of recommender systems assumes a form of homogeneity, even predictability of one's taste.

This raises a series of open questions: to what extent is it possible to identify people based on their musical preferences? Assuming there are distinctive traits in one's musical consumption, are these truly reflective of their global behavior?

3. DATASET

3.1 OVERVIEW

For this study, we work with data obtained from the music streaming service Deezer,[1] that currently counts about 16M active subscribing users worldwide and has a catalog of 90M tracks. First, we collected explicit feedback data (i.e. "likes") from 1M randomly selected users, who have been active during October 2022. Let us call this data sample D_L. Users can explicitly "like" songs, albums, and artists which then appear in their "favorites" collection. As of the date of the data collection, among these 1M users 87.1% of them had explicitly liked at least one artist, and 88.9% had liked at least one song. All together the users had liked 586 512 artists and 10 822 633 unique songs.

3.2 DISTRIBUTION OF MUSICAL ITEMS BY RECEIVED LIKES

The distribution of these items according to the number of unique users who like them follows a heavy-tailed distribution (Figure 1, top). For artists, the median value is equal to one — which means that at least half of them have been liked by only one user — while the average is around 38. The most popular artist has been liked by 86 877 users. We can thus see a huge disparity between the artists, with a few extremely popular artists that attract most of the users, and many artists that are almost unknown. The songs follow a similar popularity distribution, with a median of 1, an average around 18, and a maximum of 75 453 likes.
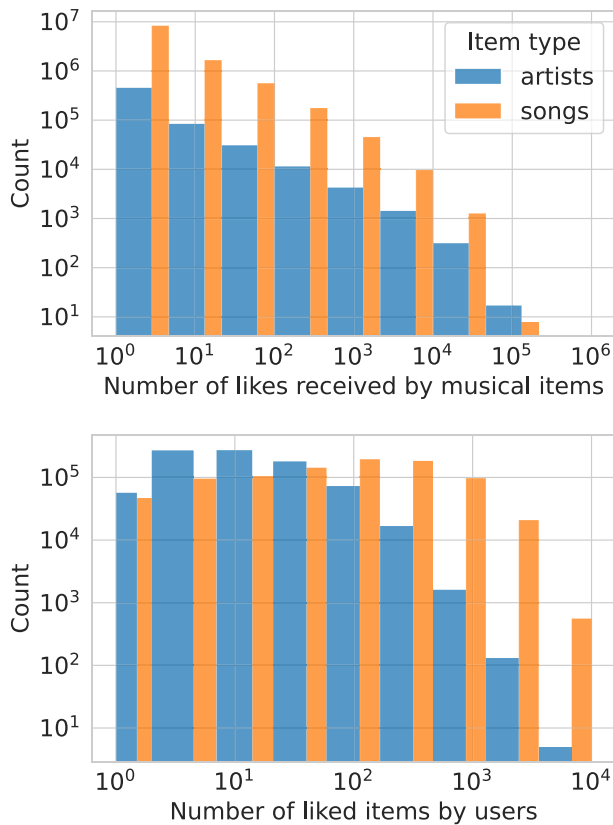
Figure 1 Heavy-tailed empirical distributions in the D_L data sample. Top: Distribution of artists' and songs' number of "fans" (i.e. users who coined these artists/songs as "liked"). A large proportion of items is liked by only a few users, while some items are very popular (hundreds of thousands of fans). Bottom: The distribution of the number of given likes per user follows here again a heavy-tailed distribution, with some users liking ten thousand more items than other users. The proportion of users liking many items drops faster for artists than for songs.

3.3 DISTRIBUTION OF USERS ACCORDING TO THE SIZE OF THEIR FAVORITES' COLLECTION

The distribution of users according to the number of artists they have liked similarly follows a heavy-tailed distribution (Figure 1, bottom). Half of the users have liked 10 artists or less, with an average of 26 liked artists per user. Some outliers exist, such as one user who has liked 7 497 artists. Users tend to like songs more than artists, with 215 favorite songs by user on average. The user experience on the platform contributes to this gap between explicitly liking artists and songs: indeed, the like button can be easily hit on a song while the user is listening to it, while liking an artist requires the user to specifically go to the artist's page.

3.4 ITEM POPULARITY METRIC

In our experiments, we will need to consider items' popularity, and we found it would be easier to represent it with a discrete variable. We decided to split items into popularity bins, from the least to the most popular, in a way such that in each bin the sum of likes received by all items is the same. We arbitrarily fixed
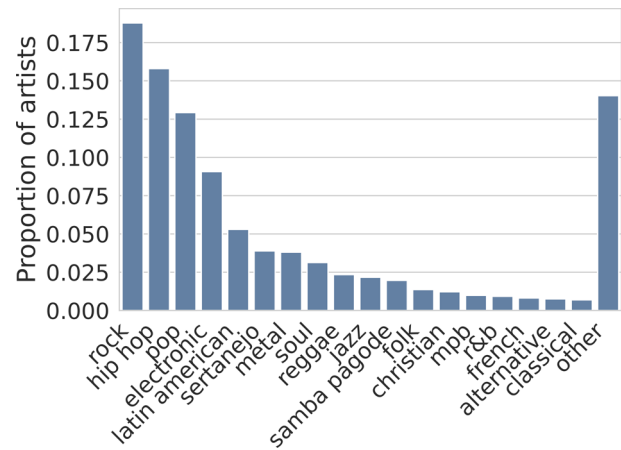
Figure 2 Proportion of D_L users' favorite artists in each music genre.

Bin	Number of artists	Number of likes
0	116	19283 – 86877
1	308	8534 – 19283
2	676	3690 – 8534
3	1865	1253 – 3690
4	7925	196 – 1253
5	575622	1 – 196

Table 1 D_L**'s artists split in 6 popularity bins.** The sum of likes for all artists is constant in each bin.

the number of bins to 6. Table 1 shows the distribution of the number of artists in each bin, as well as the maximum and minimum number of fans for artists in the bin.

3.5 GENRE TAGS

Internally, Deezer uses a taxonomy of 33 main genres to classify music, and attributes one main genre tag to most musical items in the catalog. These tags are mostly provided by music labels and recording providers, but can also be manually annotated by human editors. According to main genre tags, rock, hip-hop, pop and electronic music are the most popular music genres among the users in our dataset (Figure 2).

3.6 STREAMING DATA

In RQ3, we want to compare the users' fingerprints calculated from their favorite items with those calculated on their streaming activity. To do so, we also use a separate data sample, D_S, containing 1 year of streaming logs from April 1st 2022 to March 31st 2023 (D_{S_year}) and favorite artists (D_{S_favart}) for 60K active users. We made sure that all users were active during the entire year, in order to make comparable sub-samples for a day (D_{S_day}), week (D_{S_week}), and month (D_{S_month}) with the same users in each subset.

3.7 OPENING THE DATASET

Unless a user configured otherwise, the artists and songs that they have clicked as "liked" are publicly visible on the website using the user's ID, and can be retrieved thanks to the streaming service API.[2] We personally did not use the Deezer API and got anonymized data directly from Deezer, containing both private and public users. In section 4, we show that some users can be identified in a unique way through their favorite items. Sharing the dataset could thus raise some serious privacy concerns and we have decided not to do it in this form. Further work on means to effectively anonymize this data is required. For example, Cormode et al. (2008) obtained promising results for anonymization of sparse bipartite graphs, which is exactly the structure of our data, and it would be interesting to consider how such anonymization methods would impact our experimental results.

4. DISTINCTIVE MUSICAL TASTE FINGERPRINTS

4.1 PROBLEM DEFINITION

Previous work in the cultural sociology literature (Lahire, 2008) has focused on musical taste uniqueness and individuality. Adopting this standpoint, we wonder if it is possible to find for each user a subset of their liked or streamed items, a *fingerprint*, that could be assigned to them only – that is, that would make them unique in the crowd. This raises several questions, that include: how many items need to be selected for each user to discriminate him/her from all the others? Are certain music genres more discriminative than others?

4.1.1 Problem formulation

Let $V(u)$ be the set of liked or streamed items of user u. We look for a method to derive for each u a fingerprint, that is a subset $F(u) \subset V(u)$ which meets the following conditions:

- Non-inclusiveness: $\forall u' \neq u, F(u) \not\subset V(u')$. A fingerprint of one user can not be included in the favorite items of another user. This means that if a user's fingerprint is composed of artists a and b, this user is the only one in the dataset to like both artists a and b. Therefore, it means $F(u)$ can be used to uniquely identify u.
- Minimal size: if for one user several fingerprints validate the previous constraint, the smallest one should be chosen.

4.1.2 Problem complexity

We are planning to perform a polynomial reduction from SET COVER to FINGERPRINT.

Let us revisit the SET COVER decision problem, which is defined as follows: Given a finite universe U, a collection S of subsets of U, and a positive integer k, the problem is to determine whether there exists a sub-collection S' of S such that the union of the sets in S' covers the entire universe U, and the size of S' is at most k. It is important to note that SET COVER is known to be NP-complete.

Now, we introduce the decision problem called FINGERPRINT associated to our problem: Given $V_1, ..., V_n$, respectively the set of liked items of n individuals $u_1, ..., u_n$, and an integer k, we want to ascertain whether it is possible for the size of a fingerprint of u_1 to be less than or equal to k.

Now, let us describe a polynomial reduction from SET COVER to FINGERPRINT. To do this, let us represent the collection $S_1, ...S_n$ in SET COVER as a matrix M_S, where each row corresponds to the indicator vector of S_i. Essentially, SET COVER is about determining if it is possible to select at most k rows of M_S in a way that ensures each column contains at least one "1".

Now, let us also reformulate FINGERPRINT in matrix form. For $2 \leq i \leq n$, the (i-1)-th column of the matrix M_F represents the indicator vector of V_i, limited to the elements in V_1. In other words, the matrix M_F has $|V_1|$ rows corresponding to items in V_1. The concept of non-inclusiveness translates into ensuring that there is at least one "0" in each column of M_F. FINGERPRINT aims to find out if it is possible to select fewer than k rows of M_F while maintaining this property.

It is worth noting that if we interchange the "0" and "1" in the matrix M_S and define the $V_1, ..., V_n$ in such a way that the matrix $M_F = M_S$, solving the SET COVER instance can be achieved by solving the corresponding FINGERPRINT instance. As a result, FINGERPRINT is also NP-hard.

So, as of now (and possibly indefinitely), there is no polynomial algorithm available to resolve the fingerprinting problem. First, we propose a simple baseline, by randomly selecting items. This method matches the first constraint of non-inclusiveness, however it does not guarantee a minimal size of the fingerprints. Considering the broad-tail distribution of the number of likes received by items, scaling up the dataset by adding users increases the risk for two users to like the same items, meaning that, for each user, the size and content of its fingerprint totally depend on the total number of users in the dataset and the items they have liked. Therefore, we propose a greedy algorithm that will calculate the fingerprints globally, taking into account all other users, while minimizing their sizes locally, for each user.

4.2 METHODS

As already mentioned, the constraint of finding fingerprints of minimal size makes the problem hard to solve, and no method exists to do it in a reasonable time. Therefore, we first propose a baseline method that matches only the constraints of non-inclusiveness, and then present an approximate method to minimize the

fingerprints' sizes. We compute fingerprints based on favorite songs and artists on the 1M-user dataset, as well as favorite artists and streamed artists on a day, week, month and year time period on the 50K-user dataset. In the case of streams, we consider any user-artist interaction only once, no matter the number of times the user has streamed the artist.

4.2.1 Baseline: random selection

This first method, that we name F_{uniq_rand} builds fingerprints following our two constraints: uniqueness and non-inclusiveness. Following the same idea as De Montjoye et al. (2013), for a user u, random items from $V(u)$ are sampled and added to the fingerprint $F(u)$, as long as there exists at least one other user u' such that $F(u) \subseteq V(u')$ and $|F(u)| < |V(u)|$.

4.2.2 Minimizing fingerprints' size

The random sampling method is simple, but it likely creates fingerprints that are larger than necessary. In order to minimize the sizes of the fingerprints, we propose a greedy approach. Let $G(U, I; L)$ be the user-item bipartite graph, where U is the set of vertices representing the users, I is the set of vertices representing the items, and L are the edges linking users and items: there is an edge $(u, i) \in L$ if the user u has liked the item i. For a vertex u in U, $V(u)$ are the vertices in I that are connected with u by an edge. For each item i, let $W(i)$ be the set of users connected to i, and $d(i) = |W(i)|$ its degree.

For a user u, we first compute the weights of each item in $V(u)$, or, in other words the number of users that have liked each item in $V(u)$. Then, the item i_{min} with the smallest weight is selected and appended to $F(u)$. Then all the users that have not liked i_{min} are removed from the graph, as well as the item i_{min}, and the weights of the remaining items in $V(u)$ are recalculated. The steps are repeated while there are other users than u remaining and $|F(u)| < |V(u)|$. The full algorithm, called $F_{uniq\_minsize}$, is given in Algorithm 1.

We assume that, depending on the size of the dataset, the number of uniquely identifiable users will not be the same, and the same goes for the average fingerprint size. As the complexity of our algorithm is $O(n*m)$, the computation time will be strongly impacted by the number of users in the dataset, as well as the number of musical items they have liked, which makes it complicated to run on huge datasets, like the whole population of a streaming platform for example. In order to estimate how the number of identifiable users and their fingerprint sizes evolve with the dataset size and the two algorithms, F_{uniq_rand} and $F_{uniq\_minsize}$, we run both algorithms on subsets of 10^n users of D_L, with n going from 3 to 6, and for each n we repeat the procedure on $10^6/n$ different random subsets.

Algorithm 1 $F_{uniq\_minsize}(u)$

Input: u - user
Output: *fingerprint* - list of items

$neighbors\_users \leftarrow U$
$I_u \leftarrow V(u)$
for i in I_u **do**
    $item\_users(i) \leftarrow W(i)$
end for
$fingerprint \leftarrow \emptyset$
while $|neighbors\_users| > 1$ **do**
    if $I_u \neq \emptyset$ **then**
        $\hat{i} \leftarrow \arg\min_{i \in I_u} |item\_users(i)|$
        $fingerprint \leftarrow fingerprint \cup \{\hat{i}\}$
        $I_u \leftarrow I_u - \{\hat{i}\}$
        $neighbors\_users \leftarrow \{u' \in neighbors\_users$
                                    $if\ u' \in items\_users(\hat{i})\}$
        for i in I_u **do**
            $item\_users(i) \leftarrow W(i) \cap neighbors\_users$
        end for
    else
        return $\emptyset$
   **end if**
end while
return *fingerprint*

Also, we want to see the impact of the streaming period on those metrics, so we separately run $F_{uniq\_minsize}$ on D_{S_day}, D_{S_week}, D_{S_month} and D_{S_year}, and, additionally, D_{S_favart}.

4.3 RESULTS

4.3.1 Users' identifiability

To answer **RQ1**, we took interest in the number of users who are identifiable through their online activity. In the following sections, we will denote D_{L_uniq} the subset of D_L that contains uniquely identifiable users. As expected, songs seem to be more discriminative than artists: in D_L, 60% of users can be identified by their favorite artists, and 90% by their favorite songs.

However, users differ according to the number of items they have liked: the fewer favorite items users have, the harder they will be to identify (Figure 3). For instance, only 15% of the users with 5 favorite artists or less can be identified, while users who have liked more than 25 artists can be identified more than 95% of the time. The more items a user has liked, the more they become a so-called "power-user", i.e. a user whose collection of items fully contains all the favorite items of other users who have smaller collections (Figure 4). In a dataset of 1M users, a user who has liked one thousand or more artists covers, on average, the favorite artists of more than 1% of all the users. Overall, users with at least one hundred favorite artists cover the likes of 41% of the users from the dataset, and users with more than one thousand favorite artists cover the likes of 32% of the users (Figure 5). However, "power-users" of different ranges mostly cover the same users. For instance, 93%
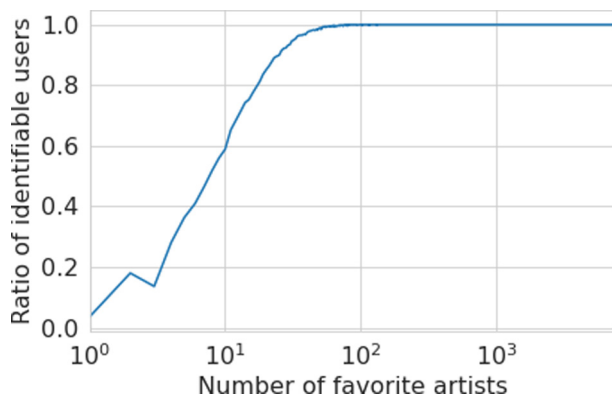
Figure 3 **Share of identifiable users in D_L depending on the number of items they have liked.** For example, among users with 10 favorite artists and more, about 60% can be identified.
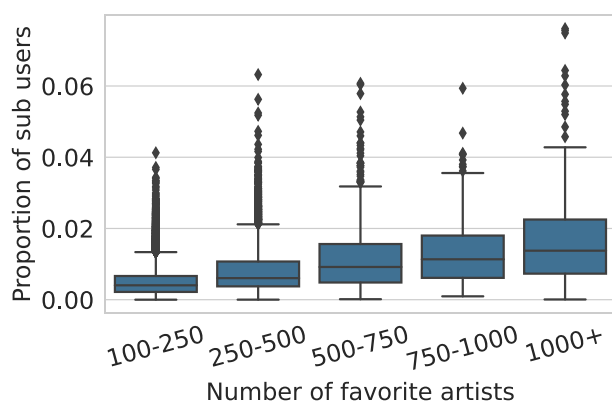

Figure 4 **Distributions of how many users (in proportion of D_L) have all their favorite artists included in those of a "power-user", for various ranges of "power-user" collection size.** For example, the likes of 1% of users are fully included on average in those of a user with 750–1000 favorite artists.
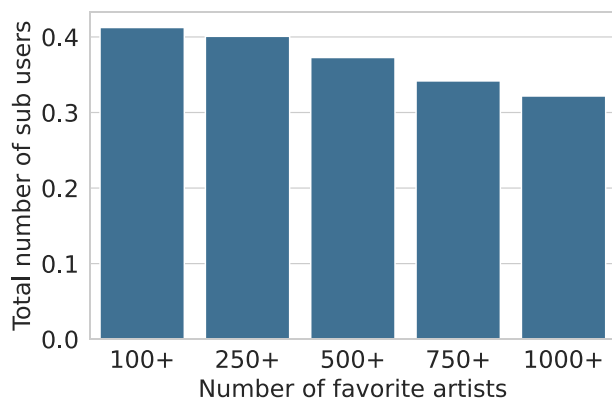

Figure 5 **Proportion of users (from D_L) whose favorite artists are included in the favorite artists of "power-users".** For example, 40% of users are included in users with more than 250 favorite artists.

of the users covered by users with 1000+ liked artists are also covered by users with 100–1000 liked artists, and 86% of the users covered by users with 1000+ liked artists are also covered by users with 100–250 liked artists. Therefore, the size of the dataset is a much more important factor for identifiability of the dataset than so-called "power-users".
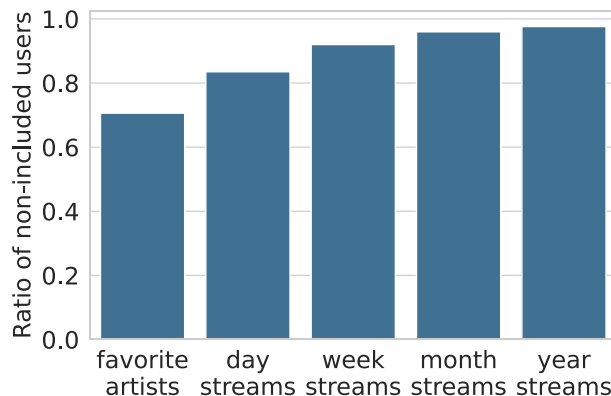
Figure 6 **Ratio of users (from D_S) identifiable through their liked and streamed artists, for different time periods.** For example, 97% of the users are identifiable via their yearly streamed artists.
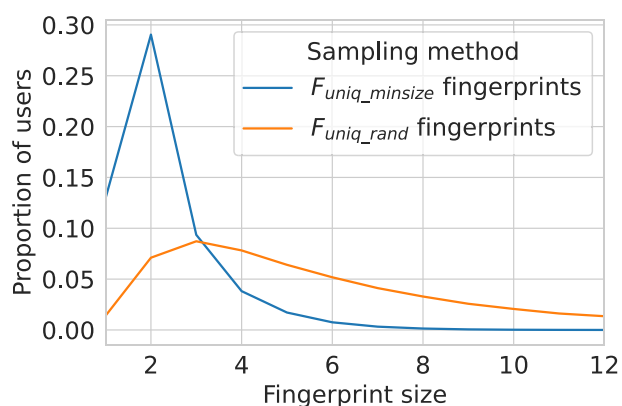
Figure 7 **Distributions of fingerprint sizes,** computed with F_{uniq_rand} and $F_{uniq\_minsize}$ based on users' favorite artists (D_L).

Additionally, we computed $F_{uniq\_minsize}$ on D_S. Expectedly, streams allow a much higher identifiability than likes, as users like much fewer artists and songs than they stream (Figure 6). Extending the time period for retaining stream logs strongly increases identifiability: one month of stream logs is enough to identify 95% of the users.

4.3.2 Fingerprint size

To answer **RQ2**, we first looked at the size of the assigned fingerprints. In D_L, we find unique fingerprints of an average size of 6.7 artists and 3.6 songs by drawing random items (Figure 7). For songs, the maximum size fingerprint is huge (176 songs to discriminate one user). Indeed, the dataset contains a few users with huge collections of liked items, up to almost 10^5 favorite songs. The favorite items of such users are most likely to cover a lot of other users' collections, which is why we would need this many items to discriminate them from others. However, considering the average and the median fingerprint size, which is 3 (for songs), we can assume that such a high fingerprint size is more of an exception than a rule.

With $F_{uniq\_minsize}$, we find unique fingerprints of an average size of 2.3 artists and 1.4 songs. Among 1M users, 45% of them are identifiable with only one song.

Table 2 shows that the average size of fingerprints based on songs increases only slightly with the size of the dataset. It can thus be assumed that even though the number of identifiable users will decrease in a larger dataset (Table 2), the average size of unique minimum size fingerprints based on songs will remain around 1.5.
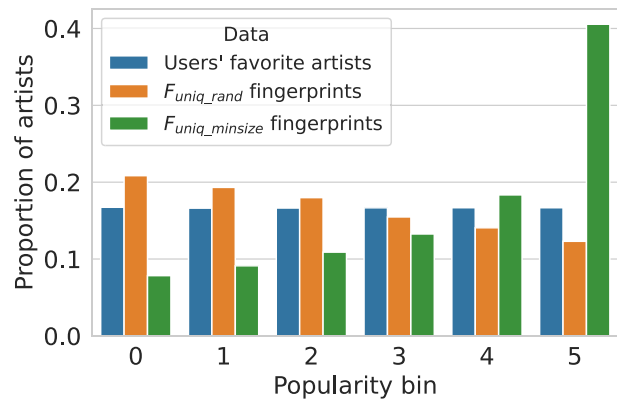
The fingerprints' size based on favorite artists D_{S_favart} (average 1.9, median 2) is comparable to one day of streams for the same users D_{S_day} (average 1.8, median 2), and slightly decreases with larger time periods (average 1.4, median 1 for a year of streams D_{S_year}).

4.3.3 Composition of the fingerprints

Another metric of interest to answer **RQ2** is the fingerprints' content. First, we compare the artists found in the fingerprints based on likes and streams, respectively D_{S_favart} and D_{S_year}. To this extent, we divide, for each user, the number of common artists by the total number of unique artists in both fingerprints. The found average ratio is around 1%, which means that there is no redundancy between the two kinds of fingerprints. Therefore, in a situation where anonymized streaming logs are shared, crossing this data with open access likes data should not lead to deanonymization, at least with the $F_{uniq\_minsize}$ method.

To have a deeper understanding of what kind of music is more discriminative, we compare the popularity and genre distributions of the fingerprint items with the users' favorite items in general (on D_L). Unsurprisingly, the popularity of an artist or a song is an important indicator of whether or not it might be included in one's fingerprint (Figure 8): the less popular the item, the more discriminative it is. As for the genres, the most popular ones, such as hip-hop, pop, rock and electronic music, seem to be underrepresented, while other, less popular genres, are overrepresented in the fingerprints (Figure 9).

Figure 8 Distribution of popularity among the artists in the fingerprints. We compare the distribution of popularity among users' favorite artists, F_{uniq_rand} fingerprints and $F_{uniq\_minsize}$ fingerprints (D_L).

Artists							
Sampling method	Number of users	Unique users (%)	Min $F(u)$ size	Max $F(u)$ size	Median $F(u)$ size	Mean $F(u)$ size	Standard deviation
F_{uniq_rand}	1000	87.3	1	13	2	2.4	1.4
	10000	77.5	1	33	3	3.5	2.3
	100000	67.7	1	58	4	4.9	3.6
	871248	58.1	1	137	5	6.7	5.3
$F_{uniq\_minsize}$	1000	87.3	1	4	1	1.3	0.5
	10000	77.5	1	7	1	1.6	0.7
	100000	67.7	1	10	2	1.9	1.0
	871248	58.1	1	14	2	2.3	1.2
Songs							
F_{uniq_rand}	1000	96.8	1	8	1.9	1.7	0.8
	10000	94.4	1	33	2	2.2	1.2
	100000	92	1	98	3	2.9	1.7
	889017	89.9	1	176	3	3.6	2.4
$F_{uniq\_minsize}$	1000	96.8	1	2	1	1.0	0.1
	10000	94.4	1	5	1	1.1	0.3
	100000	92	1	8	1	1.3	0.5
	889017	89.9	1	194	1	1.4	1.1

Table 2 Distributions of fingerprint sizes, computed with F_{uniq_rand} and $F_{uniq\_minsize}$ based on favorite artists and songs, for different numbers of users in the dataset.
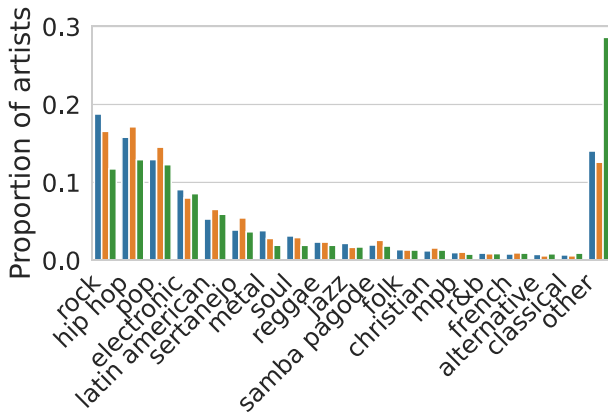
Figure 9 Distribution of genres among the artists in the fingerprints. We compare the distribution of genres among users' favorite artists, F_{uniq_rand} fingerprints and $F_{uniq\_minsize}$ fingerprints (D_t).

If we consider musical taste through individuality and uniqueness, as Lahire (2008) did, we are then able to create fingerprints of musical taste. However, is individuality on its own a sufficient definition of taste, and do these unique fingerprints capture the essence of the users' preferences?

5. REPRESENTATIVE MUSICAL TASTE FINGERPRINTS

The method we describe in the previous section can be used to distinguish users and to capture what makes their musical taste unique. In the process, it seems to have selected elements that do not necessarily reflect the overall distribution of their preferences.

In the previous section, we use the term fingerprint; in this section, we will keep using this concept, by analogy to the previous section, even though we are not looking to identify users anymore. Here, we consider a representative fingerprint as a set of items that summarize a user's preferences. We use items, and not embeddings or other latent variables, as we want our fingerprint to be easily interpretable, and again, as a mirror with the previous section.

We propose to measure the representativeness of a fingerprint by means of a prediction task: i.e. given the subset of items selected, can we reconstruct the full set of a user's liked items? We then present a fingerprinting method that allows us to build a representative fingerprint according to two defined evaluation methods. Finally, we compare it with the unique fingerprints computed in Section 4.

5.1 PROBLEM DEFINITION

We formulate the problem in a way similar to recommendation: a subset $F(u)$ is considered as representative of $V(u)$ if there exists a method $F^\star$ such that $\forall u \in U, V(u) \approx F^\star(F(u))$, . In other words, we consider

that a fingerprint is representative if a method that can recover the initial set of items from it exists. Building an $F^\star$ function is a ubiquitous task in recommender system research, where the problem is very similarly defined.

We chose to define $F^\star$ as a simple prediction function based on the nearest neighbor algorithm, computing the proximity between the artists using matrix factorisation (Koren et al., 2009). For favorite items, we start by building a sparse artist-user matrix M, where $M[u,i]=1$ if the user u has liked the artist i. For streams, $M[u,i]=1$ if the user u has streamed the artist i at least once during the given time span. We then compute a singular value decomposition (SVD), and use the first 128 dimensions of the SVD as our artists' embeddings. The artists' nearest neighbours are then computed based on the Euclidean distances between their embeddings.

Let $N(i)$ be a list of i's nearest neighbors ordered from the closest to the furthest. Let w_i be a weight associated to each item i in a fingerprint $F(u)$. This weight represents the number of items we need to recover from u. If all items in $F(u)$ are equally important, then we want to recover the same number of items from each item in $F(u)$: $\forall i \in F(U), w_i = (|V(u)| - |F(u)|)/|F(u)|$. For a user u, $F^\star$ returns a set of predicted items $P(u)$ by simply taking, for each item i in $F(u)$, the w_i closest neighbors of i from the list of i's 150 most similar artists.

5.2 EVALUATION PROXY

The representativeness score of a fingerprint is calculated based on how close the predicted items are to the user's favorite items. We propose two methods to compare $P(u)$ and $V(u)$:

- Item-wise. This evaluation is the most strict. The predicted items $P(u)$ are compared exactly to the actual user's favorite items (except the ones included in the fingerprint). The prediction accuracy for a user u is thus equal to $|P(u) \cap (V(u)|F(u))|/|P(u)|$. This metric is widely used in recommender systems for offline evaluation tasks, where ground truth user-item interactions are available.
- Genre-wise. Here, we compare if the predicted items follow similar distributions in terms of genre as the items from the user's actual favorite items. The prediction score is thus simply the L_1 distance between the distribution of genres in $P(u)$ and the one in $V(u)$.

Other metrics could also be used, based on the mainstreamness of the artists for example.

5.3 EXPERIMENTS
5.3.1 A method to sample representative fingerprints

We propose a simple method, that we name $F_{rep\_kmedoid}$, to compute fingerprints that would be representative of the users' preferences.

		F_{rep_rand}		$F_{rep\_kmedoid}$	
Evaluation	Number of favourite artists	Mean accuracy	Standard deviation	Mean accuracy	Standard deviation
Item-wise	<25	0.05	0.11	0.08	0.13
	25–50	0.14	0.12	0.25	0.13
	50–75	0.16	0.12	0.28	0.12
	75–100	0.18	0.11	0.30	0.12
	100–150	0.21	0.12	0.32	0.12
	>150	0.26	0.12	0.37	0.10
Genre-wise	<25	0.38	0.31	0.40	0.28
	25–50	0.65	0.14	0.73	0.09
	50–75	0.70	0.13	0.78	0.08
	75–100	0.71	0.12	0.81	0.07
	100–150	0.77	0.10	0.83	0.08
	>150	0.88	0.12	0.97	0.05

Table 3 Item-wise and genre-wise prediction accuracy with $F_{rep\_kmedoid}$ fingerprints and randomly sampled fingerprints of the same sizes on D_{S_favart}.

Considering that each artist is represented with an embedding of size 128 (computed in Section 5.1), the favorite artists of user u, $V(u)$, are split into k clusters using the k-medoids algorithm. The medoids are then used as representative artists of each cluster to build the user's fingerprint, and the weight $w(i)$ associated to each artist i in the fingerprint is the size of the related cluster. We assume that the diversity of music genres in the users' favorite artists varies from one user to another, thus the optimal k may not be the same for different users. In order to determine the optimal k for each user, we computed fingerprints with k going from 1 to 15% of $|V(u)|$ (as we consider a fingerprint as concise information about the users' preferences, we set maximum k limit to 20), then run the prediction task $F^\star$ on the obtained fingerprint. For each user, we retain the optimal k value that gave the highest prediction score with an item-to-item evaluation.

As a baseline, we use a method F_{rep_rand}, which consists in randomly sampling k items in $V(u)$ for a user u, with u's optimal k value for $F_{rep\_kmedoid}$. Table 3 shows that the prediction scores for $F_{rep\_kmedoid}$ fingerprints on liked items are indeed higher than with F_{rep_rand}, both with item-to-item and genre-wise evaluation, and the score is higher for users with larger music collections. A better prediction accuracy is achieved with streaming data (Table 4) – for yearly streaming logs, we can restore almost 40% of the exact items through the fingerprints. Reaching an accuracy of 1 with an item-wise evaluation is not feasible within such a vast item

	Accuracy		Optimal k	
Data sample	Mean	Standard deviation	Mean	Standard deviation
Favorite artists	0.09	0.12	2.66	3.26
Day streams	0.07	0.11	1.86	1.67
Week streams	0.13	0.11	5.03	4.42
Month streams	0.26	0.13	8.82	5.70
Year streams	0.35	0.12	9.73	6.23

Table 4 Prediction accuracy and optimal k with an item-to-item evaluation for $F_{rep\_kmedoid}$ on favorite artists and streamed artists for different time periods (D_s).

space, and this level of precision is also uncommon in real-world recommendation systems. Based on the positive dynamics of the prediction accuracy on larger datasets, in the following, we will consider $F_{rep\_kmedoid}$ as a method that aims to capture the essence of users' musical taste.

An interesting thing to notice is the optimal k size in different datasets: a smaller average fingerprint size is observed with favorite artists and single-day streams. The average size then grows with larger streaming time spans, and so does the standard deviation (Table 4). The average size can be easily connected to the amount of data to recover. Complementarily, the growing standard deviation can be explained by the heterogeneity of the users: on a one year span, some users will listen to a large variety of different genres, and some will stick to only a few, which is why the ideal fingerprint size might be very different from one user to another.

5.3.2 Uniqueness vs essence

To answer **RQ3**, we now want to confront our two sampling methods, $F_{uniq\_minsize}$ and $F_{rep\_kmedoid}$.

First, we run both sampling methods on uniquely identifiable users D_{L_uniq}, then run the prediction on both obtained fingerprints: the prediction accuracy from the unique fingerprints is lower (3% with item-to-item evaluation, 42% genre-wise) than the representative fingerprints (10% with item-to-item evaluation, 50% genre-wise) (Figure 10), meaning that the most discriminative items in the users' libraries are not representative of their overall preferences.

Second, we found that only 279 435 users remain identifiable from D_L's representative fingerprints, comparing to 507 037 in D_L overall. Thus, extracting the essence of one's musical library most likely leads to a loss of the information that makes them unique.

Trying to quantify the essence and the uniqueness of one's musical taste seems to represent two diverging goals, which require distinct computation methods.
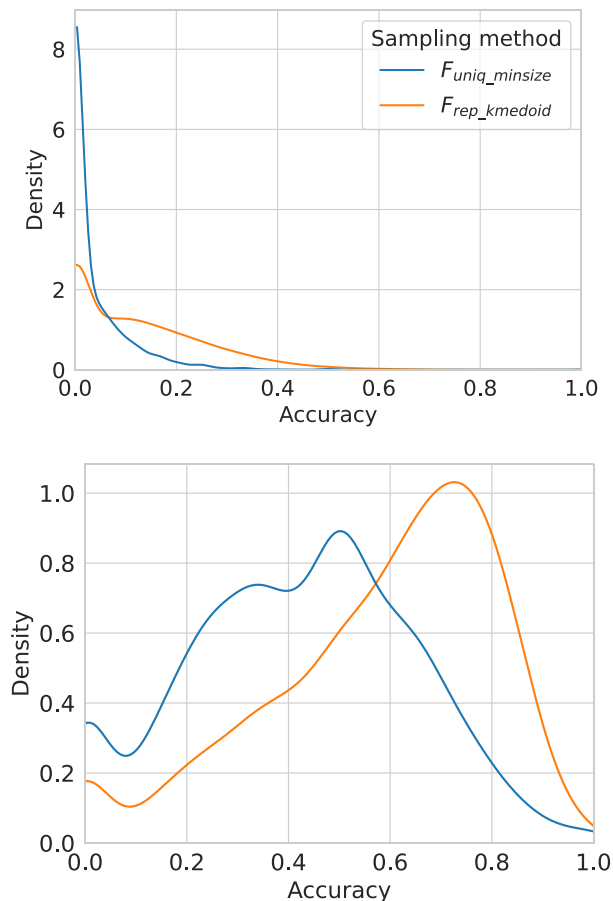
Figure 10 Item-wise (top) and genre-wise (bottom) prediction accuracy with $F_{uniq\_minsize}$ fingerprints and $F_{rep\_kmedoid}$ fingerprints, performed on D_{L_uniq}.

6. DISCUSSION

6.1 CONCLUSION

Building on a large set of literature, we emphasize how preference elicitation encompasses several conflicting definitions. We propose to make two of them explicit, stressing constraints of uniqueness (respectively representativeness) as optimization goals adapted to distinction (respectively characterization) of an individual's taste fingerprint. We show that these different constraints lead to diverging solutions which in turn suggests that scientific work addressing musical taste should probably reflect on their exact objectives and make their understanding of the term explicit.

We run our experiments using data from a major streaming platform, containing both explicitly liked content and streaming logs. In a first section of experiments, we show that in a sample of 1M active users, 90% can be identified by their favorite songs, and one or two songs is enough to identify 45% of the users. On another sample of 50K users, we also show that streaming logs are even more identifying, especially if collected for a long period of time – up to 97% of the users are identifiable via the artists they streamed for a year (**RQ1**).

However, the artists allowing to identify users are not the same when it comes to what they have liked or streamed. Also the more identifying items are expectedly the less popular ones, and by consequence, those from less popular genres (**RQ2**).

In a second section of experiments we propose a method to depict users' preferences by creating representative subsets of users' favorite items that we call fingerprints. This method can further be used in situations when concise information about the users' preferences is needed: in recommendation systems, or scientific work that uses the concept of musical taste.

We show that the best items for identifying users are not the most representative of their preferences: using a prediction task, we can recover an average of 10% of the users' favorite artists from the representative fingerprints against 3% from the unique fingerprints. Complementarily, only 279 435 users remain identifiable based on their representative fingerprints, against 507 037 in the initial set. It thus seems that the essence and uniqueness of musical taste are opposite concepts (**RQ3**).

6.2 LIMITATIONS

The experiments proposed in this work are nonetheless limited by the nature of the data used to conduct them. As we have emphasized, observable data are handy to collect at scale, but arguably they are non-perfect proxies of an individual's true preferences. In particular, the information of explicit *distaste* is missing, though it appears to be a highly relevant indicator. An intuitive approach would be to leverage implicit feedback such as skips, but these are even noisier signals.

A more promising approach would be to build a richer, multi-modal dataset, containing both declared and observed data for a sufficient number of individuals. This will be the focus of our future work. Additionally, the evaluation of the fingerprinting methods could also be improved, in particular by means of an experiment involving the users themselves, for instance using an interface such as the one presented by Cura et al. (2022).

6.3 ADDRESSING PRIVACY ISSUES

Unlike streaming logs, information about users' likes is publicly accessible on the Deezer platform and most of their competitors, unless users specifically indicate their account as private. The fact that most users can be identified by their likes basically shows that a significant share of them are by default 1-anonymous Sweeney (2002), thus not anonymous. It reveals an important privacy issue – the usual practice of hashing the users' IDs does not seem to be enough to anonymize a dataset. It can be especially compromising to share personal data, such as geolocation for example, combined with information about the users' likes. Future work could be done to explore ways to aggregate or obfuscate such

data in order to ensure k-anonymity, while keeping its expressiveness at the same time.

In the music information systems used on platforms, which must remain expressive for users, there is no category for describing music that would be both more precise than music genres (which are ill-defined categories) and more aggregated than the precise catalog items consumed by users: tracks, artists, and albums. Consequently, for lack of a better alternative, our results suggest that publicly available information about individuals' music preferences should likely be aggregated at the level of music genres to strengthen anonymity (e.g. possibly defined as clusters of artists, whose size should be adjusted to ensure k-anonymity, with artist cluster sizes depending on k).

NOTES

1 www.deezer.com.
2 developers.deezer.com/api/user.

ACKNOWLEDGEMENTS

FUNDING INFORMATION

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

KM, MM, TL and OB designed the study. KM analyzed the data and ran the experiments. KM, MM and TL drafted the manuscript. All authors wrote and approved the final version of the manuscript.

AUTHOR AFFILIATIONS

Kristina Matrosova ⓘ orcid.org/0000-0002-1831-3705
Géographie-cités, CNRS, France; LIPN, USPN, France

Manuel Moussallam ⓘ orcid.org/0000-0003-0886-5423
Deezer Research, France

Thomas Louail ⓘ orcid.org/0000-0001-8563-6881
Géographie-cités, CNRS, France; PACTE, CNRS, Sciences Po Grenoble, France

Olivier Bodini ⓘ orcid.org/0000-0002-1867-667X
LIPN, USPN, France

REFERENCES

Bogdanov, D., Haro, M., Fuhrmann, F., Xambo, A., Gomez, E., and **Herrera, P.** (2013). Semantic audio content-based music recommendation and visualization based on user preference examples. *Information Processing & Management*, 49(1):13–33. DOI: https://doi.org/10.1016/j.ipm.2012.06.004

Bourdieu, P. (1984). *Distinction – A Social Critique of the Judgement of Taste*. Harvard University Press.

Brown, R. A. (2012). Music preferences and personality among Japanese university students. *International Journal of Psychology*, 47(4):259–268. DOI: https://doi.org/10.1080/00207594.2011.631544

Bryson, B. (1996). "Anything but heavy metal": Symbolic exclusion and musical dislikes. *American Sociological Review*, pages 884–899. DOI: https://doi.org/10.2307/2096459

Cano, P., Koppenberger, M., and **Wack, N.** (2005). Content-based music audio recommendation. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, pages 211–212. DOI: https://doi.org/10.1145/1101149.1101181

Cormode, G., Srivastava, D., Yu, T., and **Zhang, Q.** (2008). Anonymizing bipartite graph data using safe groupings. In *34th International Conference on Very Large Data Bases*, pages 833–844. DOI: https://doi.org/10.14778/1453856.1453947

Coulangeon, P. (2017). Cultural openness as an emerging form of cultural capital in contemporary France. *Cultural Sociology*, 11(2):145–164. DOI: https://doi.org/10.1177/1749975516680518

Cura, R., Beaumont, A., Beuscart, J.-S., Coavoux, S., de Fozieres, N. L., Bigot, B. L., Renisio, Y., Moussallam, M., and **Louail, T.** (2022). Uplifting interviews in social science with individual data visualization: The case of music listening. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–9. DOI: https://doi.org/10.1145/3491101.3503553

De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and **Blondel, V. D.** (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3(1):1–5. DOI: https://doi.org/10.1038/srep01376

Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J., and **Moussallam, M.** (2018). Music mood detection based on audio and lyrics with deep neural net. *arXiv preprint* arXiv:1809.07276.

Delsing, M. J., Ter Bogt, T. F., Engels, R. C., and **Meeus, W. H.** (2008). Adolescents' music preferences and personality

characteristics. *European Journal of Personality*, 22(2):109–130. DOI: https://doi.org/10.1002/per.665

DeNora, T. (2000). *Music in Everyday Life*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511489433

Eck, D., Lamere, P., Bertin-Mahieux, T., and **Green, S.** (2007). Automatic generation of social tags for music recommendation. *Advances in Neural Information Processing Systems*, 20.

Ferwerda, B. and **Schedl, M.** (2014). Enhancing music recommender systems with personality information and emotional states: A proposal. In *Posters, Demos, Late-Breaking Results and Workshop Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization*.

Flegal, K. M., Ogden, C. L., Fryar, C., Afful, J., Klein, R., and **Huang, D. T.** (2019). Comparisons of self-reported and measured height and weight, BMI, and obesity prevalence from national surveys: 1999–2016. *Obesity*, 27(10):1711–1719. DOI: https://doi.org/10.1002/oby.22591

Fromkin, H. L. and **Snyder, C. R.** (1980). The search for uniqueness and valuation of scarcity. In *Social Exchange*, pages 57–75. Springer. DOI: https://doi.org/10.1007/978-1-4613-3087-5_3

George, D., Stickle, K., Rachid, F., and **Wopnford, A.** (2007). The association between types of music enjoyed and cognitive, behavioral, and personality factors of those who listen. *Psychomusicology: A Journal of Research in Music Cognition*, 19(2):32. DOI: https://doi.org/10.1037/h0094035

Hargreaves, D. J., North, A. C., and **Tarrant, M.** (2006). *Musical preference and taste in childhood and adolescence*. Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780198530329.003.0007

Koren, Y., Bell, R., and **Volinsky, C.** (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37. DOI: https://doi.org/10.1109/MC.2009.263

Lahire, B. (2008). The individual and the mixing of genres: Cultural dissonance and self-distinction. *Poetics*, 36(2–3):166–188. DOI: https://doi.org/10.1016/j.poetic.2008.02.001

Langmeyer, A., Guglhor-Rudan, A., and **Tarnai, C.** (2012). What do music preferences reveal about personality? A cross-cultural replication using selfratings and ratings of music samples. *Journal of Individual Differences*, 33(2):119. DOI: https://doi.org/10.1027/1614-0001/a000082

Laplante, A. (2014). Improving music recommender systems: What can we learn from research on music tastes? In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 451–456.

Liang, D., Krishnan, R. G., Hoffman, M. D., and **Jebara, T.** (2018). Variational autoencoders for collaborative filtering. In *Proceedings of the World Wide Web Conference*, pages 689–698. DOI: https://doi.org/10.1145/3178876.3186150

Majumdar, A., Kumar, A., and **Manohar, S.** (2009). Music recommendations based on implicit feedback and social

circles: The Last FM data set. https://cseweb.ucsd.edu/classes/wi15/cse255-a/reports/fa15/007.pdf.

Narayanan, A. and **Shmatikov, V.** (2008). Robust deanonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, pages 111–125. DOI: https://doi.org/10.1109/SP.2008.33

North, A. C. (2010). Individual differences in musical taste. *The American Journal of Psychology*, 123(2):199–208. DOI: https://doi.org/10.5406/amerjpsyc.123.2.0199

Oard, D. W. and **Kim, J.** (1998). Implicit feedback for recommender systems. In *Proceedings of the AAAI Workshop on Recommender Systems*, volume 83, pages 81–83. Madison, WI.

Pazzani, M. J. and **Billsus, D.** (2007). Contentbased recommendation systems. In Brusilovsky, P., Kobsa, A., and Nejdl,W., editors, *The Adaptive Web*, pages 325–341. Springer. DOI: https://doi.org/10.1007/978-3-540-72079-9_10

Peterson, R. A. (1992). Understanding audience segmentation: From elite and mass to omnivore and univore. *Poetics*, 21(4):243–258. DOI: https://doi.org/10.1016/0304-422X(92)90008-Q

Rentfrow, P. J. and **Gosling, S. D.** (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, 84(6):1236. DOI: https://doi.org/10.1037/0022-3514.84.6.1236

Schedl, M., Knees, P., McFee, B., Bogdanov, D., and **Kaminskas, M.** (2015). Music recommender systems. In *Recommender Systems Handbook*, pages 453–492. Springer. DOI: https://doi.org/10.1007/978-1-4899-7637-6_13

Soleymani, M., Aljanaki, A., Wiering, F., and **Veltkamp, R. C.** (2015). Content-based music recommendation using underlying music preference structure. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. DOI: https://doi.org/10.1109/ICME.2015.7177504

Sordo, M., Celma, O., Blech, M., and **Guaus, E.** (2008). The quest for musical genres: Do the experts and the wisdom of crowds agree? In *Proceedings of the International Conference on Music Information Retrieval*, pages 255–260.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570. DOI: https://doi.org/10.1142/S0218488502001648

Uitdenbogerd, A. and **Schyndel, R.** (2002). A review of factors affecting music recommender success. In *Proceedings of the 3rd International Conference on Music Information Retrieval*, pages 204–208.

Van den Oord, A., Dieleman, S., and **Schrauwen, B.** (2013). Deep content-based music recommendation. *Advances in Neural Information Processing Systems*, 26.

Vargas, S. and **Castells, P.** (2013). Exploiting the diversity of user preferences for recommendation. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pages 129–136.

Way, S. F., Gil, S., Anderson, I., and **Clauset, A.** (2019). Environmental changes and the dynamics of musical identity. In *Proceedings of the International AAAI* *Conference on Web and Social Media,* volume 13, pages 527–536. DOI: https://doi.org/10.1609/icwsm. v13i01.3250

]u[ᙚ

Chapter 7

Measuring the influence of recommendation on music listening

How we model music taste, shape the musical space, and choose algorithms with specific parameters all significantly impact the outcomes of recommendations. In turn, these recommendations can influence the users' preferences and listening habits, creating a feedback loop between the system and the user. After exploring various approaches to modeling user preferences, we began to question how the structure of the data itself, combined with the algorithm used, might shape the recommendations, particularly concerning fairness issues.

Around this time, in 2022, Lesota et al. (2022) published a study exploring how different algorithms affected the recommendation of local music on the Last.fm platform. According to their experiments, ItemKNN, an item-based CF algorithm, would tend to promote more local music, while NeuMF, a neural network-based MF model, would lean towards recommending more American content at the expense of local music. We found this finding interesting, but we wondered why such bias was observed, an aspect that the authors did not explore in their paper. We decided to take the subject of local music as an example to understand where bias in music recommendation can come from.

To extend their work, we turned to streaming data from Deezer, which allowed us to work with a broader and more diverse user base than the LFM-2b dataset used by Lesota et al. (2022). Given the contrasting performance of ItemKNN and NeuMF on the same data, we hypothesized that the bias might arise from the interactions between the structure of the data and the specific characteristics of each algorithm.

Initially, we attempted to model Deezer data using SBMs to detect structural patterns in user behavior. Our first approach was to split users from each country into distinct categories (or blocks) based on the proportion of local music they listened to. We then used a DCBM, which aimed to better reproduce the distribution of local music preferences within these blocks. We were hoping

that this approach would capture the underlying patterns of user preferences and apply them to artist networks.

Next, we computed artist graphs based on both the real interaction matrix and the one generated by the SBM. However, the generated graphs failed to preserve genre and musical relationships between artists, which were present in the original data. Moreover, the local music recommendation biases that we observed in the real Deezer data, were not reproduced in the generated data.

This led us to question whether all music datasets share similar structural patterns. We decided to compare the impact of recommendation algorithms on both the Deezer and Last.fm datasets. Upon investigation, we found that the biases present in the Deezer dataset were markedly different from those reported in the Last.fm dataset by Lesota et al. (2022).

To explore these discrepancies further, we conducted a comparative analysis of the LFM-2b and Deezer datasets, and uncovered key differences in data structure, user demographics, and local music consumption patterns. Ultimately, we discovered a significant lack of country tags in both datasets. This raised important questions about the feasibility of measuring algorithmic biases when crucial metadata is lacking.

This work culminated in the paper entitled ***"Do Recommender Systems Promote Local Music? A Reproducibility Study Using Music Streaming Data"***, published in the *Reproducibility track* of the proceedings of the *18th ACM Conference on Recommender Systems* in 2024.

The paper was co-authored by Lilian Marey (Télécom Paris, Deezer Research), Guillaume Salha-Galvan (Deezer Research), Thomas Louail (CNRS, Géographie-cités), Olivier Bodini (Université Sorbonne Paris Nord), and Manuel Moussallam (Deezer Research). The study was primarily designed by myself and Manuel Moussallam, with early conceptual input from Thomas Louail and Olivier Bodini, and later contributions from Guillaume Salha-Galvan. The experiments were conducted by myself and Lilian Marey. I wrote the first draft of the paper, Guillaume Salha-Galvan revised and refined the final version.

This study highlights the challenges of measuring the influence of RSs and identifies critical aspects that need consideration in this task, such as the representativeness of the population, the quality of metadata, and the specific parameters used in recommendation algorithms.

Do Recommender Systems Promote Local Music?
A Reproducibility Study Using Music Streaming Data

Kristina Matrosova[*]
CNRS, Geographie-Cités
France
LIPN, USPN
France

Lilian Marey
LTCI, Télécom Paris
France
Deezer Research
France

Guillaume Salha-Galvan
Deezer Research
France

Thomas Louail
CNRS, Geographie-Cités
France
PACTE, CNRS, Sciences Po Grenoble
France

Olivier Bodini
LIPN, USPN
France

Manuel Moussallam
Deezer Research
France

ABSTRACT

This paper examines the influence of recommender systems on local music representation, discussing prior findings from an empirical study on the LFM-2b public dataset [1]. This prior study argued that different recommender systems exhibit algorithmic biases shifting music consumption either towards or against local content. However, LFM-2b users do not reflect the diverse audience of music streaming services. To assess the robustness of this study's conclusions, we conduct a comparative analysis using proprietary listening data from a global music streaming service, which we publicly release alongside this paper. We observe significant differences in local music consumption patterns between our dataset and LFM-2b, suggesting that caution should be exercised when drawing conclusions on local music based solely on LFM-2b. Moreover, we show that the algorithmic biases exhibited in the original work vary in our dataset, and that several unexplored model parameters can significantly influence these biases and affect the study's conclusion on both datasets. Finally, we discuss the complexity of accurately labeling local music, emphasizing the risk of misleading conclusions due to unreliable, biased, or incomplete labels. To encourage further research and ensure reproducibility, we have publicly shared our dataset and code.

CCS CONCEPTS

• **Information systems** → *Recommender systems*; *Personalization*.

KEYWORDS

Music Recommendation, Fairness, Algorithmic Bias, Local Music.

---

[1]Previously available at http://www.cp.jku.at/datasets/LFM-2b/, the LFM-2b dataset has recently been taken down due to license issues.

1 INTRODUCTION

Recommender systems are essential for music streaming services like Apple Music, Deezer, and Spotify [9, 26, 38, 39]. They help mitigate information overload problems by showcasing the most relevant content for each user, within large catalogs of millions of songs, albums, and artists [7, 19, 23, 32, 39]. They also assist users in discovering new music they might like on these services [6, 8, 19]. With the rise of streaming as the predominant form of music consumption [25, 34], there has been, however, a noticeable increase in debates about the responsibilities of these systems. Concerns are also growing about their ability to promote a fair and diverse musical landscape and the various biases they might introduce or amplify when recommending music [2, 13, 14, 15, 17, 18, 29, 30, 40].

In particular, Lesota et al. [31] recently argued that some music recommender systems might intensify the predominance of US music consumption in other countries. Specifically, in an empirical study focused on the LFM-2b dataset of listening actions on Last.fm [37], these authors investigated the extent to which standard recommender systems favor US-produced content over *local* music from the country of origin of each user. Their findings suggest that NeuMF [24], a neural network-based collaborative filtering algorithm, recommends lower proportions of local music than what users from each country actually listen to. In other words, NeuMF exhibits an *algorithmic bias* [15, 22] against local music on LFM-2b. On the contrary, the more classical ItemKNN [12] method yields more calibrated recommendations, and even fosters the consumption of local music in most countries under consideration.

This study undoubtedly raised essential issues regarding the uneven impact of recommender systems on local music consumption. Nonetheless, as big as it might be, the LFM-2b dataset used for evaluation is of a particular nature. Last.fm users tend to be active on the Internet and social media, and are not evenly distributed across countries. Therefore, they might not represent the full spectrum of

music streaming service users. Moreover, as detailed in Section 2, the study did not analyze how important parameters related to model training impact these biases. For these reasons, it remains unclear whether the conclusions of Lesota et al. [31] would hold in other experimental settings and using a different dataset.

In this paper, we address this question by conducting a comprehensive comparative study using proprietary listening data from the global music streaming service Deezer. Our study concentrates on France, Germany, and Brazil – three countries where Deezer is one of the leading market players. Our contributions are as follows:

- Firstly, we show that the Deezer and LFM-2b datasets present different local music consumption patterns. This discrepancy suggests caution when drawing conclusions about local music representation based solely on one dataset like LFM-2b.
- Secondly, we demonstrate that NeuMF and ItemKNN exhibit different algorithmic biases towards local music on Deezer compared to LFM-2b when following the evaluation setup of Lesota et al. [31]. Importantly, we also uncover several factors that significantly influence these biases – including their magnitude but also their direction – thereby affecting this study's overall conclusions on both datasets. These factors include the number of tracks each model recommends, their training variability, and whether they were trained on data from individual countries or the entire dataset.
- Thirdly, we explain that accurately labeling local music is a complex endeavor, and that the proportion of local music consumed and recommended can vary significantly depending on the source of the labels, their level of completeness, and the various biases introduced by human annotators. Consequently, we recommend prioritizing the development of comprehensive, transparent, and reliable local data labeling in future research. We argue that this foundational step is crucial for studies aiming to understand local music biases, as results based on unreliable labels may be misleading.
- Lastly, along with this paper, we publicly release our Deezer dataset as well as the source code of our experiments. This release aims to ensure full reproducibility of our results and to facilitate future studies on local music recommendation.

The remainder of this paper is organized as follows. In Section 2, we introduce the problem more formally and review the related work in more detail. In Section 3, we introduce our Deezer dataset and compare it to LFM-2b in terms of local music consumption. We report and discuss results from our empirical study on local music recommendation and biases in Section 4, and conclude in Section 5.

2 PRELIMINARIES

We begin this section by formally presenting the problem under consideration, before reviewing the related work.

2.1 Problem Formulation

2.1.1 Notation. In this paper, we consider a set $\mathcal{V}$ of music tracks available in the catalog of a music streaming service, and a set $\mathcal{U}$ of $M \in \mathbb{N}^*$ users on this same service. We denote by $N_{\text{listened}}(u)$ the number of streams performed by each user u over a predefined time period. Moreover, we denote by $N_{\text{local}}(u) \in \{0, \ldots, N_{\text{listened}}(u)\}$ the number of these streams that are of music tracks from the

country of origin of u according to some data labeling[2]. We refer to $N_{\text{local}}(u)$ as the number of *local* streams of u. Using this formalism, the proportion of local music listened to by u is:

$$\text{L}(u) = \frac{N_{\text{local}}(u)}{N_{\text{listened}}(u)} \in [0, 1]. \tag{1}$$

Additionally, we consider a music recommender system:

$$\text{MRS}_K : \mathcal{U} \to \mathcal{V}. \tag{2}$$

MRS_K recommends[3] K music tracks from $\mathcal{V}$ to each user of the music streaming service, for some fixed value $K < N$. The number of local music tracks recommended to the user u by MRS_K among these K tracks is $N_{\text{local,MRS}_K}(u) \in \{0, \ldots, K\}$. The proportion of local music tracks recommended to u by MRS_K is:

$$\text{L}_{\text{MRS}_K}(u) = \frac{N_{\text{local,MRS}_K}(u)}{K} \in [0, 1]. \tag{3}$$

2.1.2 Objective. Our main goal in this paper is to investigate the impact of MRS_K on local music representation. In line with Lesota et al. [31], our main indicator of interest will be the *algorithmic bias* of MRS_K in favor or against local music, defined as follows:

$$\text{Bias}_{\text{MRS}_K} = \frac{1}{M} \sum_{u \in \mathcal{U}} \left(\text{L}_{\text{MRS}_K}(u) - \text{L}(u) \right), \tag{4}$$

with $\text{Bias}_{\text{MRS}_K} \in [-1, 1]$. In essence, a positive bias (respectively, a negative bias) indicates that, on average, MRS_K recommends more local music (resp., less local music) than what users of the music streaming service organically listen to. The remainder of this paper will analyze this value for different datasets, recommender systems, and settings. We will aim to uncover the various factors that might influence the intensity or even the direction of this bias.

2.2 The study of Lesota et al. [31] on LFM-2b

Analyzing local music algorithmic biases was one of the key objectives of Lesota et al. [31], with a particular emphasis on the predominance of US music consumption in other countries.

2.2.1 Context. Over the past decades, US music has dominated the global music industry, with its cultural influence spreading worldwide. Trends from the US have been widely adopted even in local music productions [20], and the ratio of US music on radio charts has been rapidly increasing[4] since the 1960s [1]. The consequences of this dominance are mixed [1, 10, 11]. On the one hand, it can potentially stimulate local cultural development through the adaptation to global trends and reinforcement of local identity, a process known as *glocalization* [1, 10]. On the other hand, it is also sometimes perceived as a threat, termed *cultural imperialism*, which could lead to the decline of local cultures [11, 33, 42]. While music has become more centralized with the rise of music streaming services [5, 31], at the time of Lesota et al.'s study [31], limited research

---

[2]We note that associating music tracks with specific countries can be an ambiguous task, and that different data labeling rules may yield inconsistent results. The importance of the labeling source will be pointed out and further discussed throughout this paper.
[3]At this stage, we do not formulate assumptions regarding the specific data or paradigm (e.g., a collaborative filtering or content-based approach [7]) used to recommend tracks.
[4]However, this growth slowed down from the 1990s due to several factors, including the emergence of CDs, which made music production more accessible worldwide, content localization by MTV, and the introduction of laws in countries like France, imposing local music quotas on radio station programming [1].

Do Recommender Systems Promote Local Music? A Reproducibility Study Using Music Streaming Data

RecSys '24, October 14–18, 2024, Bari, Italy

had focused specifically on the impact of these services and their recommender systems on local music consumption.

2.2.2 Results. In 2022, Lesota et al. [31] published results from their empirical study conducted on a subset[5] of the LFM-2b public dataset, which includes listening events from users of the Last.fm music website [37]. This study explored the prominence of US cultural imperialism in online music consumption, revealing that while the US maintains a strong position among Last.fm users, its influence varies significantly across countries. The authors also observed varying glocalization patterns depending on countries. The final part of their study, which our reproducibility paper focuses on, investigated whether recommender systems can increase existing predominances and, overall, shift music consumption towards specific countries at the expense of local content. The authors answered this question positively, explaining that the influence is uneven and algorithm-dependent. Their experiments suggest that NeuMF [24], a neural network-based collaborative filtering algorithm, recommends lower proportions of local music than what Last.fm users in each country organically listen to. In other words, NeuMF exhibits a negative $Bias_{MRS_K}$ local bias, as computed in Equation (4). Conversely, the more traditional ItemKNN [12] method tends to promote the consumption of local music in most countries, i.e., it is associated with a positive $BiasMRS_K$ local music bias.

2.2.3 Limitations of Lesota et al. [31] and Motivations of our Work. This study undoubtedly raises important issues regarding the uneven impact of recommender systems on local music representation. It positions itself within a growing body of scientific research focusing on the fairness of music recommender systems and their biases, not only regarding local music but also other aspects, including gender and music genres [14, 15, 17, 18, 30, 40].

Nonetheless, we believe this study also suffers from limitations that motivate our work. From an algorithmic perspective, the authors investigated biases using a single number of recommended tracks, $K = 10$. The effect of varying K, i.e., allowing their systems to recommend different numbers of tracks, on the magnitude or even the direction of biases is unclear. Additionally, systems like the neural network-based NeuMF include randomness in training [24], yet the robustness of the study's conclusions to this randomness remains unverified. Thirdly, their systems were trained on all users, but using data solely from users of each specific country might alter the findings. Bauer and Schedl [3, 4], for instance, suggest distinguishing between country-specific and global mainstreamness to ensure a realistic representation of musical preferences in different countries and promote less biased recommendations.

Beyond these algorithmic considerations, replicating this study with a different dataset is also worthwhile. Indeed, Last.fm users tend to be active on the internet and social media and are not evenly distributed across countries [37]. Therefore, they might not reflect the diverse audience of music streaming services. Furthermore, Lesota et al. [31] relied on MusicBrainz, an open music encyclopedia [16], to associate artists with country labels. However, MusicBrainz labels may not only be imprecise but are also

missing for some LFM-2b tracks, which were simply excluded by Lesota et al. [31]. Zanger et al. [44] suggest that this exclusion could lead to measurement errors due to label biases. Indeed, less popular artists or those from locations or genres unrepresented among MusicBrainz human annotators might lack more labels[6], leading to a distorted representation of local music in some countries. These factors raise the question of whether the findings of Lesota et al. [31] would remain valid with actual listening data from a music streaming service, or by using alternative labeling sources. The remainder of this paper will aim to clarify these aspects.

3 COMPARING LFM-2B TO DEEZER DATA

In this section, we present our dataset of listening events from Deezer and subsequently provide a comparative analysis with LFM-2b.

3.1 The Deezer Dataset

3.1.1 Overview. We examine a proprietary dataset from the global music streaming service Deezer, comprising the listening history of 30 000 randomly selected users on this platform in March 2019. The dataset features an equal distribution of users from the three countries of our study – France, Germany, and Brazil – with 10 000 users from each. It includes approximately 4 million streams across over 565 000 distinct music tracks. To maintain consistency with Lesota et al. [31], we did not filter listening events by streaming context. As a result, the dataset includes both organic streams and recommendations, which we later discuss in Section 4.

3.1.2 Local Music Labeling. The country of each user is determined based on their IP address. Determining an artist's country may sometimes be ambiguous, for example in the case of artists born in a country but who gained fame in another one. To capture this complexity, we consider three different country labels in our work:

- Our dataset includes the main *country of activity* and *country of origin* of each artist, as provided by Deezer when available, and compiled by this service from public and private sources.
- Moreover, we added the publicly accessible country labels from the MusicBrainz open music encyclopedia [16] when available. We recall that these labels were the ones used by Lesota et al. [31] in their original study on LFM-2b.

3.2 Descriptive Analysis

We now provide a descriptive analysis of our Deezer dataset. We compare it to the subset of LFM-2b[7] of Lesota et al. [31] and additionally restrict this subset to the three countries of interest in this study. The sample contains 254 users from France, 805 users from Germany, and 1064 users from Brazil, for a total of over 3 million listening events in 2018 and 2019, on around 100 000 music tracks.

3.2.1 LFM-2b vs Deezer. Overall, we observe that Deezer and LFM-2b exhibit quite different patterns of local music consumption.

First of all, by plotting the proportion of local streams in both datasets, using MusicBrainz labels and considering only labeled

---

[5]The entire LFM-2b dataset includes approximately 2 billion listening events over 15 years from about 120 000 users. Lesota et al. [31] analyzed a subset of 14 million interactions from 2018 and 2019, involving 13 000 users in 20 countries selected for having at least 100 users and artists who had collectively created at least 1 000 tracks.

[6]As an illustration, Jul, a French rapper, has no music genre annotation in MusicBrainz, despite being one of the most-streamed artists on Deezer in France during 2018 and 2019.

[7]We use data kindly provided by the authors via private email communications.
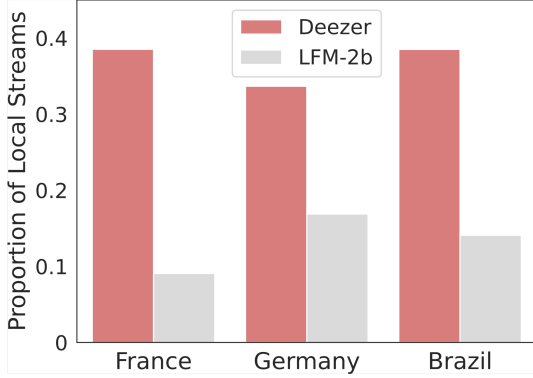
Figure 1: Proportion of local streams by country, according to the LFM-2b and Deezer datasets. All values are computed using MusicBrainz labels, by considering labeled tracks only.

Table 1: Top 10 most streamed music tracks by French users, in the LFM-2b (top) and Deezer (bottom) datasets. All reported country labels originate from MusicBrainz.

Dataset	Artist/Band	Title	Country Label	Singing Language	Release Year	Music Genre
LFM-2b	Portishead	Glory Box	GB	EN	1994	Trip Hop
	Radiohead	Karma Police	GB	EN	1997	Alt. Rock
	The Verve	Bitter Sweet Sym.	GB	EN	1997	Alt. Rock
	Franz Ferdinand	Take Me Out	GB	EN	2004	Indie Rock
	a-ha	Take On Me	NL	EN	1985	Synth Pop
	Angèle	Balance ton quoi	BE	**FR**	2018	Pop
	The xx	Intro	GB	EN	2009	Trip Hop
	4 Non Blondes	What's Up?	US	EN	1993	Alt. Rock
	Metronomy	The Look	GB	EN	2010	Indie Rock
	Wax Tailor	Que Sera	**FR**	EN	2004	Trip Hop
Deezer	Ninho	Goutte d'eau	**FR**	**FR**	2019	Rap
	Angèle	Tout oublier	BE	**FR**	2018	Pop
	Lady Gaga	Shallow	US	EN	2018	Folk Pop
	Lomepal	Trop beau	**FR**	**FR**	2018	Rap/Pop
	David Guetta	Say My Name	**FR**	EN	2018	EDM
	Ariana Grande	7 rings	US	EN	2019	Pop
	Alonzo	Assurance vie	**FR**	**FR**	2019	Rap
	DJ Snake	Taki Taki	**FR**	EN	2018	EDM
	Kaaris	Gun salute	**FR**	**FR**	2019	Rap
	Booba	PGP	**FR**	**FR**	2019	Rap

tracks in both cases (Figure 1), we observe a significantly lower rate of local music in the LFM-2b dataset. For Brazil, for instance, LFM-2b exhibits 2.5 times fewer local streams compared to Deezer. Moreover, we report in the first two columns of Figure 2 histograms of the percentage of local streams per user, here again using MusicBrainz labels. We note that the distributions are different between the two datasets. In the Deezer dataset, across all three countries, users exhibit varying patterns: some do not listen to local music at all, while others listen to local music only, with a large spectrum of behaviors in between. Conversely, the LFM-2b dataset shows a stark contrast, with few users listening to a majority of local music.
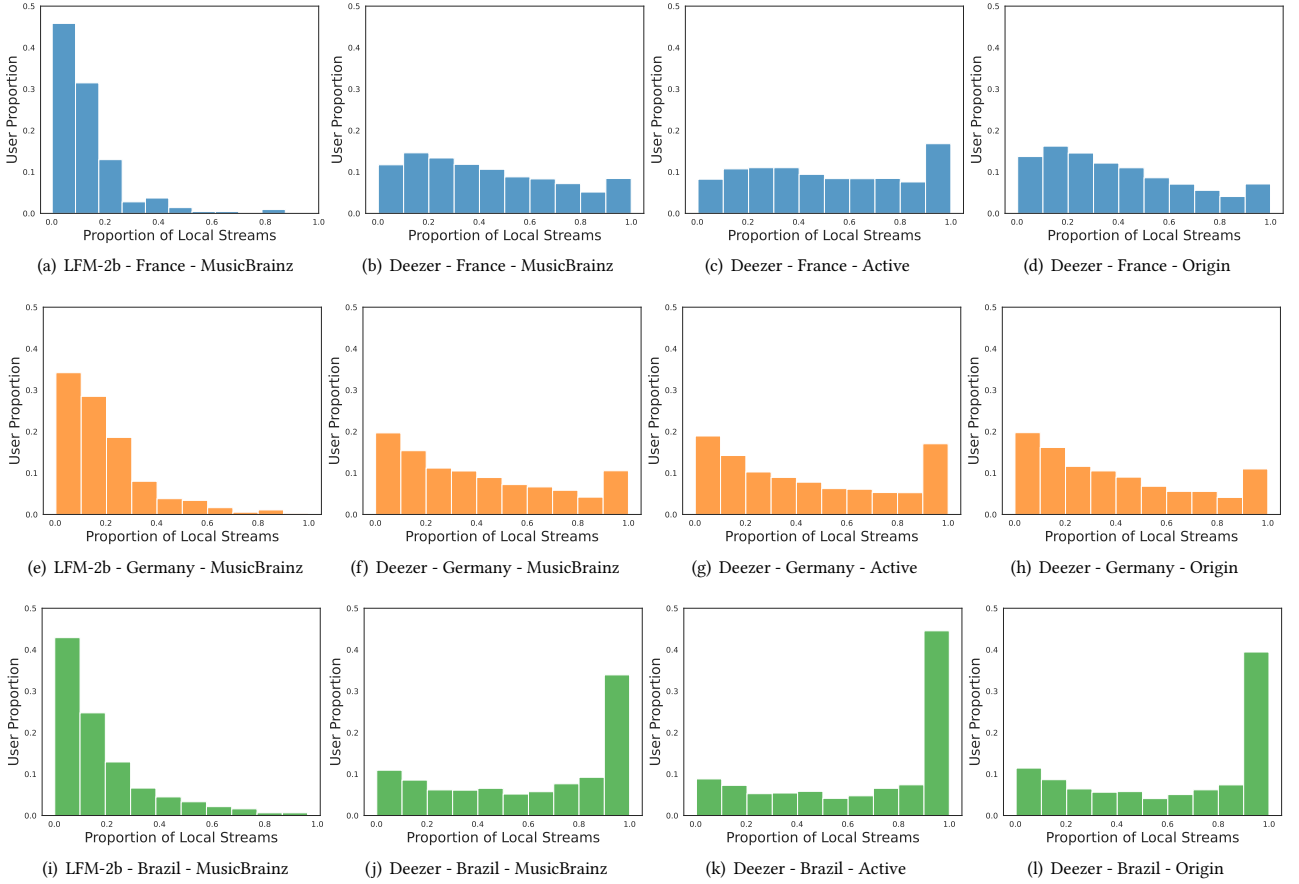
To go further, we present in Table 1 the top 10 most streamed music tracks in France, in both datasets. The Deezer dataset not only contains a higher proportion of French music (both in terms of artists and lyrics), but also predominantly features recent releases, from the year prior to, or the same year as the streams. The prevalent genres include pop, rap, and electronic music. On

Table 2: Percentages of (i) labeled streams, (ii) local streams (among the labeled streams) and (iii) local streams (among all streams) in the Deezer dataset, by country and label source. A labeled stream corresponds to a stream of a music track associated with a country label. A local stream corresponds to a stream where the user and the artist have the same country label.

Country	Label Source	Labeled Streams	Local Streams Among Labeled	Local Streams Among All
France	Deezer - Activity	76 %	50 %	38 %
	Deezer - Origin	75 %	34 %	26 %
	MusicBrainz	76 %	38 %	29 %
Germany	Deezer - Activity	60 %	40 %	24 %
	Deezer - Origin	62 %	30 %	18 %
	MusicBrainz	69 %	33 %	23 %
Brazil	Deezer - Activity	41 %	48 %	19 %
	Deezer - Origin	36 %	37 %	13 %
	MusicBrainz	38 %	38 %	14 %

the contrary, the top tracks in the LFM-2b dataset predominantly consist of older releases (dating back one or more decades) with English lyrics and genres such as indie, alternative rock, or trip hop, which are more niche. These differences can be attributed to several characteristics of the Last.fm website, upon which the LFM-2b dataset is built [37]. Firstly, Last.fm caters primarily to music enthusiasts who have a strong inclination towards collecting and organizing their music libraries, potentially resulting in a preference for less mainstream music genres. Secondly, this website's users are predominantly English-speaking persons, which introduces a population bias. While the Deezer dataset may appear to be more reflective of realistic music consumption patterns, it is important to acknowledge the possibility of similar biases existing within it, as well as in data from other streaming services. Hence, it is crucial to proceed with caution when asserting the presence of cultural patterns based on such data. Using multiple data sources for cross-validation becomes imperative to ensure the reliability and accuracy of conclusions.

3.2.2 Impact of Label Sources. Table 2 presents the proportions of labeled streams and local streams (among the labeled ones, and among all streams) in the Deezer dataset, according to the three label sources, i.e., Deezer's country of origin and country of activity, as well as MusicBrainz labels. We observe that none of the labeling sources provides complete coverage. Across the three countries considered, between 64% and 24% of the streams remain unlabeled. Streams in different countries exhibit varying levels of label coverage. For example, the artist's country is identified in 75-76% of streams by French users, depending on the label source, while for streams from Brazil, this coverage drops to only 36-41%. Label coverage varies by country depending on the source. For instance, Deezer's activity labels provide the highest coverage for streams from Brazil, but they offer the least coverage for streams from Germany. Furthermore, the proportions of local consumption strongly vary depending on the label source. For instance, considering only labeled streams, only 38% of French users' streams consist of French tracks according to MusicBrainz labels, whereas Deezer's

Do Recommender Systems Promote Local Music? A Reproducibility Study Using Music Streaming Data

RecSys '24, October 14–18, 2024, Bari, Italy

Figure 2: Histograms of the proportion of local streams per user (considering labeled tracks only). Results are split by dataset (i.e., LFM-2b or Deezer), country (i.e., France, Germany, or Brazil), and labeling source (i.e., MusicBrainz labels, Deezer's country of activity, or Deezer's country of origin).

activity labels indicate 50%. This difference is notable given that both sources have identical label coverage rates for French streams. Due to incomplete labeling, there's a significant difference between the proportions of local streams among labeled streams versus all streams. Calculating local streams based solely on labeled data can suggest higher local consumption than what is actually observed across all streams. Moreover, these values aren't proportional; for instance, when considering only labeled streams, France and Brazil vie for the title of the largest local consumer (depending on the label source). However, when all streams are taken into account, Brazil exhibits the lowest local consumption by a substantial margin.

In summary, obtaining a complete and universally unquestionable labeling of local music proves to be a challenging task. Overall, we advocate against simply filtering out unlabeled tracks, as done in the reference study [31]. Indeed, such an approach may result in removing a majority of the streams from the study, potentially undermining the validity of the study's conclusion. As outlined in Section 2.2.3, this filtering operation can also introduce label

biases [44], when annotators exhibit preferences for specific countries, music genres, or languages during the annotation process. The extent to which these discrepancies between labels result in inconsistencies regarding the measurement of local music algorithmic biases will be analyzed in Section 4.

4 EXPERIMENTAL ANALYSIS OF LOCAL MUSIC RECOMMENDATION AND BIASES

In this section, we now present our empirical analysis of local music recommendation and biases on the LFM-2b and Deezer datasets. We start by describing the experimental setting. Then, we report and discuss our findings. Notably, we compare them with the main conclusions of the original study of Lesota et al. [31].

4.1 Experimental Setting

4.1.1 Models. We examine the same two collaborative filtering [28] recommender systems analyzed in the reference study:

- NeuMF [24] is a deep learning-based recommender system that integrates traditional matrix factorization (MF) [28] with

neural networks [21]. NeuMF first learns *embedding* vector representations of users and music tracks in the same vector space where proximity reflects similarity, by factorizing a user-track interaction matrix. These embeddings are then processed by a deep neural network architecture trained to predict the most relevant tracks to recommend to each user.

- ItemKNN [12] is a more traditional recommender system based on the nearest neighbor approach [28]. For each user, it assigns similarity scores to tracks by evaluating how similar these tracks are to those the user has previously listened to. This similarity is determined in a collaborative filtering fashion, by analyzing the interactions of other users. When recommending a set of K tracks, ItemKNN selects the top K neighbors with the highest similarity scores. Unlike NeuMF, ItemKNN operates directly on the user-track interaction matrix and does not learn embedding representations.

Lesota et al. [31] trained their ItemKNN and NeuMF models using the entire LFM-2b dataset, including users from all countries. In contrast, our study not only trains these models on the complete LFM-2b and Deezer datasets but also considers country-specific variants. We developed these variants by using only listening data from users within the same country – a realistic setting for a global music streaming service. For instance, to recommend music to a Deezer French user, we would employ the ItemKNN or NeuMF variant trained on Deezer's listening data from French users.

4.1.2 Task and Implementation Details. For both datasets, we train all models on a top 10 track recommendation task, evaluated using mean reciprocal rank (MRR@10) scores [45] computed on a validation set of 10% randomly selected users, masked during the training phase. Afterwards, we compute the local biases $\text{Bias}_{\text{MRS}_K}$, as defined in Equation (4), for each model in each country, averaged across all users in that country[8]. While in the reference study [31], the authors only reported results for a single number of recommended tracks ($K = 10$), here we consider the more general case of a varying K, with K ranging from 10 to 100 with a step of five tracks.

As in the reference study, we use the implementation of ItemKNN and NeuMF available in RecBole [46], a Python library based on PyTorch [36] that aims to provide a unified framework for developing and reproducing recommendation algorithms. For ItemKNN, we retrieve nearest neighbors to recommend by using cosine similarities computed from the user-track train interaction matrix, with a null value for the shrink parameter [35]. We train all NeuMF models for a maximum of 300 epochs using the Adam optimizer [27], with a learning rate of 0.001, batch sizes of 512 items, a dropout rate of 0.1 [41], and minimizing a binary cross-entropy loss [24]. All NeuMF models learn embedding vectors of dimension 64. For interested readers, we provide exhaustive information on each layer of every neural network in our public GitHub repository (see Section 4.3).
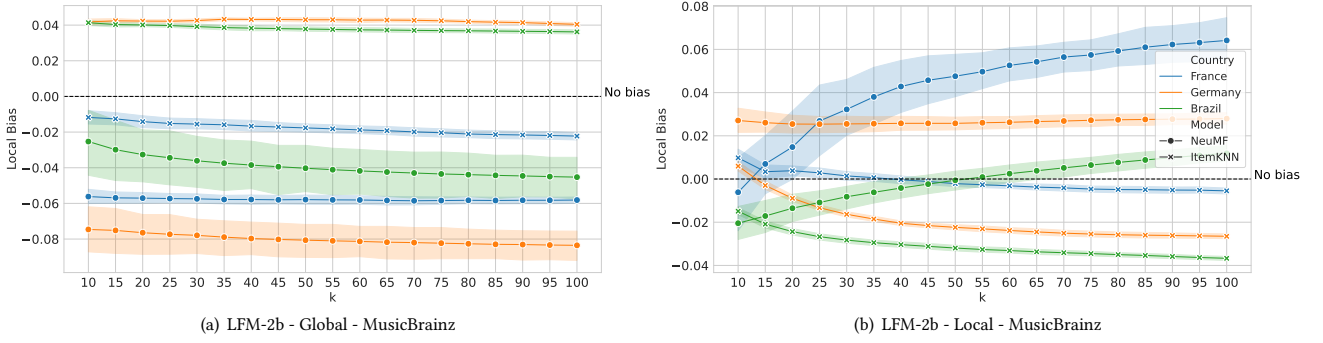
4.2 Results and Discussion

4.2.1 Results on LFM-2b. We begin our analysis with the results obtained on the LFM-2b public dataset. We report in Figure 3 the

---

[8]We note that one might alternatively compute biases using only users from a test set. However, reporting biases for all users not only aligns with the evaluation protocol of Lesota et al. [31], but also reflects the practical goal of music streaming services, which would typically aim to address local biases across their entire user base.

local music algorithmic biases of ItemKNN and NeuMF on LFM-2b with MusicBrainz labels, averaged over 20 model runs with standard deviations to assess variability in the training process. Results are split by training variant, i.e., global or country-specific. In particular, Figure 3(a) reports results for the global ItemKNN and NeuMF variants trained on users from all countries, which matches the specific setting of Lesota et al. [31] (with $K = 10$ only in their study). Overall, we reproduce results comparable to those of the original study in this specific setting. NeuMF recommends lower proportions of local music than what users from France, Germany, and Brazil listen to, unveiling negative algorithmic biases. In contrast, ItemKNN tends to foster the consumption of local music in Brazil and Germany, while displaying a negative but relatively small bias in France. Our results are consistent when modifying the number K of recommended tracks.

However, Figure 3(b) reveals that the results change drastically when training ItemKNN and NeuMF in a country-specific fashion, i.e., using data from LFM-2b but selecting users of a single country only, instead of all users. For instance, for $K = 10$, NeuMF now shows a positive bias in Germany, while ItemKNN shows a negative bias in Brazil. Interestingly, increasing the number of recommended tracks K can also reverse the bias direction. For example, NeuMF exhibits a negative bias for $K \in \{10, \ldots, 45\}$ but a positive bias for $K \in \{50, \ldots, 100\}$. This observation highlights the importance of testing different values of K to draw robust conclusions about the potential biases of each recommender system against local music. Lastly, Figure 3(b) underlines the importance of accounting for variability in the training process, particularly for NeuMF, which shows large ±1 standard deviation intervals (this variability primarily stems from randomness in the initialization of neural weights, dropout components, and the use of different training splits for each model run). As an illustration, in France and for $K = 10$, NeuMF's interval overlaps with the "No bias" horizontal dotted line. This emphasizes that NeuMF has shown both positive and negative biases in our experiments, depending on each training instance.

4.2.2 Results on Deezer. We now compare these results with those obtained using the proprietary Deezer dataset, which contains streams from users of the music streaming service Deezer. Figure 4 presents the local music algorithmic biases of ItemKNN and NeuMF on this dataset. Once again, all biases are averaged over 20 model runs with standard deviations. Results are categorized by training variant (i.e., global or country-specific) and label source (i.e., MusicBrainz labels, Deezer's country of activity, or Deezer's country of origin). We begin our discussion with an inspection of Figure 4(a), which displays results for the global ItemKNN and NeuMF variants using MusicBrainz labels. This setting is consistent with the one used in the reference study [31] and our earlier Figure 3(a), but applied to the Deezer dataset instead of LFM-2b. We observe that the algorithmic biases exhibited by ItemKNN and NeuMF on LFM-2B vary significantly on Deezer. At $K = 10$, for instance, all models are associated with a positive average bias value, contrary to previous results from Figure 3(a) and Lesota et al. [31] on LFM-2b. We also notice that biases tend to be of higher magnitude on Deezer users. These apparent discrepancies tend to reinforce our discussion from Section 3, highlighting the need for caution when drawing conclusions based solely on one dataset like

(a) LFM-2b - Global - MusicBrainz

(b) LFM-2b - Local - MusicBrainz

Figure 3: Local music algorithmic biases of ItemKNN and NeuMF on LFM-2b users in France, Germany, and Brazil, computed for numbers of recommended tracks K varying from 10 to 100 with a step of 5 tracks. Results are split by training variant ("Global" models are trained using listening data from users of all countries, while "Local" models are trained using only listening data from users of the same country). All values are averaged over 20 model runs and reported with ± 1 standard deviation intervals. Values above (respectively, under) the "No bias" 0-level horizontal dotted line indicate that the model exhibits a positive (resp., a negative) algorithmic bias towards local music.

LFM-2b. While this public dataset serves as a convenient starting point for researchers, the observed biases may not be consistent across different datasets with varying listening patterns.
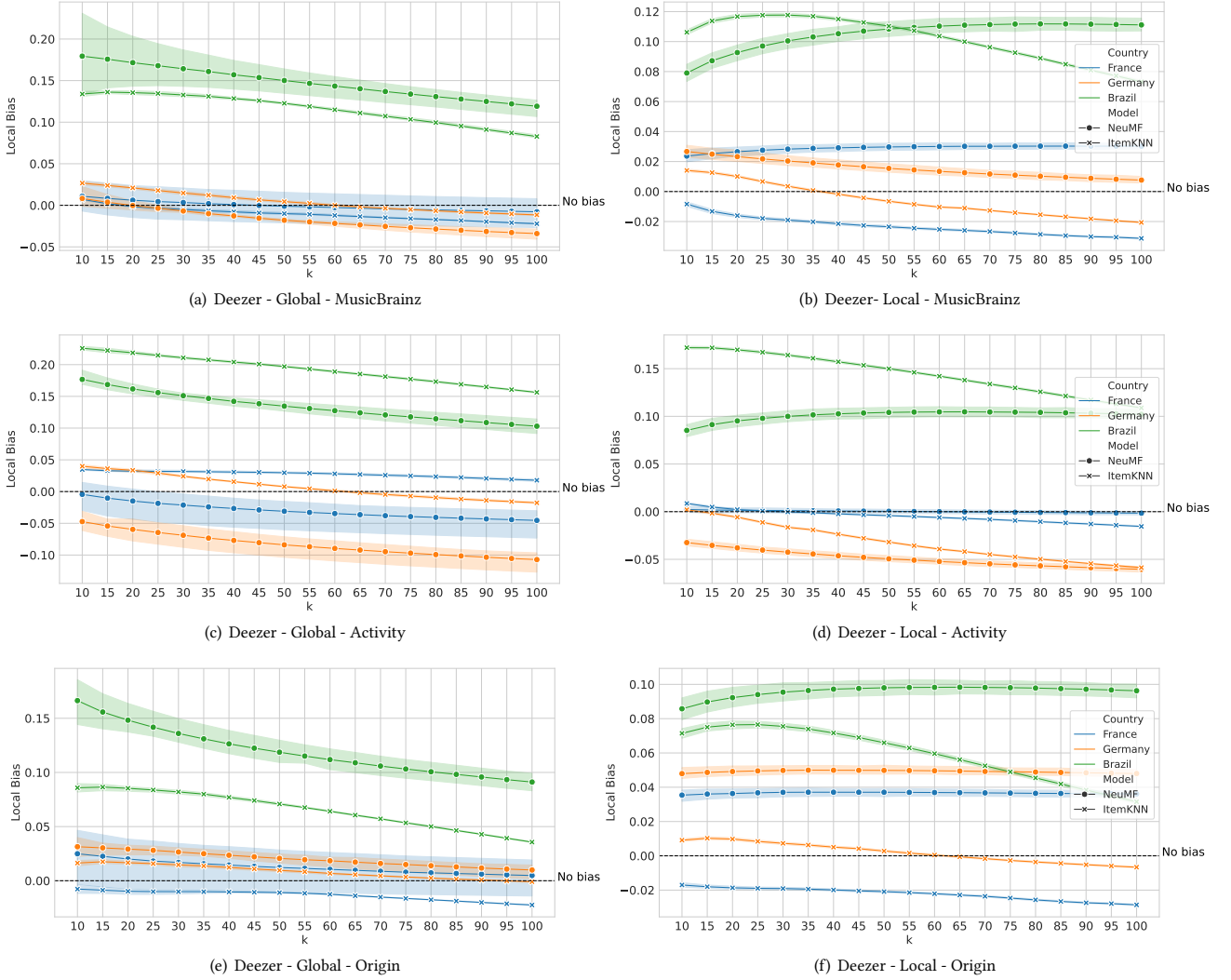
Figure 4(b) reports comparable experiments using our local variants of ItemKNN and NeuMF with MusicBrainz labels. Below, Figure 4(c) and Figure 4(d) present results for global and local models, respectively, but using Deezer's country of activity labels instead of MusicBrainz labels. Finally, Figure 4(e) and Figure 4(f) show results for global and local models, respectively, but using Deezer's country of origin. Overall, these figures confirm our previous insights from Section 4.2.1. Training models with data coming from one country only can significantly alter local music biases in both magnitude and direction. Additionally, changing the number of recommended music tracks K and using different user splits or weight initializations can also affect these biases. While these algorithmic factors were not examined in the original work [31], our experiments reveal that they can change the global picture and the study's conclusion. Properly accounting for these factors is, therefore, crucial to ensure a robust and reliable analysis of local music algorithmic biases.

Figure 4 also highlights that changing the label sources can also substantially affect conclusions. Section 3 had already uncovered that the proportions of local music consumed by users may strongly depend on the label source. Figure 4 further demonstrates that this variation leads to inconsistencies in the measurement of local music algorithmic biases. For instance, the global and local NeuMF models are consistently associated with a negative bias in Germany when relying upon the (country of) "Activity" label (Figure 4(c) and Figure 4(d)), but on the opposite they are associated with a positive bias when using the (country of) "Origin" label (Figure 4(e) and Figure 4(f)). While changing the label source drastically impacts the results, we acknowledge that, nonetheless, some findings remain robust to these changes. For example, Brazil is consistently associated with the highest positive biases across almost all settings in Figure 4. We hypothesize that this consistent behavior may be due to the higher number of Brazilian Deezer users who listen exclusively

or almost exclusively to local music, according to all three label sources (see Figure 2). This aspect might be reflected in our models, although further analysis would be required for confirmation.

4.2.3 Limitations and Future Work. In concluding our discussion, we acknowledge some limitations of our experimental analysis, which also offer opportunities for future research. Firstly, as explained in Section 3.1.1, our Deezer dataset includes both organic streams and recommended streams, to maintain consistency with the original study. However, focusing solely on organic streams for model training could be worthwhile. As Lesota et al. [31] have also pointed out, incorporating recommended tracks may distort the model's insights about user preferences. Examining the impact of this adjustment on biases towards local music would be an interesting avenue for further investigation. Secondly, Villermet et al. [43] showed that users are very different when it comes to their use of the different features offered by streaming platforms – including algorithmic recommendation – and that only a minority of them primarily relies on algorithmic recommendation to select the music they listened to on streaming platforms. Thus, reproducing the experiment by considering only users who interact with music recommender systems to a certain degree could provide more relevant findings. Thirdly, while we used a specific definition of bias, its perception might actually be subjective, and individual user interviews may be useful to gain additional insights.

At first glance, one might also want to analyze the local music algorithmic biases of numerous other recommender systems, beyond ItemKNN and NeuMF. However, we believe that a more crucial preliminary step for future work will be to improve the accuracy of local music labels. Our study underscores the challenge of accurately labeling local music, and demonstrates how variations in label sources can substantially affect conclusions regarding local music representation and recommendation. Conducting extensive bias analyses on numerous recommender systems with the current state of labeling – where no single label source covers more than

(a) Deezer - Global - MusicBrainz

(b) Deezer- Local - MusicBrainz

(c) Deezer - Global - Activity

(d) Deezer - Local - Activity

(e) Deezer - Global - Origin

(f) Deezer - Local - Origin

Figure 4: Local music algorithmic biases of ItemKNN and NeuMF on Deezer users in France, Germany, and Brazil, computed for numbers of recommended tracks K varying from 10 to 100 with a step of 5 tracks. Results are split by training variant ("Global" models are trained against listening data from users of all countries, while "Local" models are trained using only listening data from users of the same country), and by label source (i.e., MusicBrainz labels, Deezer's country of activity, or Deezer's country of origin). All values are averaged over 20 model runs and reported with ± 1 standard deviation intervals. Values above (resp. under) the "No bias" 0-level dotted line indicate that the model exhibits a positive (resp. negative) algorithmic bias.

80% of streams, and where labels reflect biases from human annotators [44] – could prove fruitless. Indeed, results obtained from misleading labels could render such analyses unreliable. Consequently, we recommend prioritizing the development of comprehensive and reliable local data labeling in future research. We believe that cross-referencing assumptions across multiple labels remains one of the most reliable practices towards achieving this goal.

4.3 Open-Source Code and Data Release

Along with this paper we release two important resources. Firstly, an anonymized version of our Deezer proprietary dataset, containing 4 million listening events from 30 000 Deezer users in France, Germany, and Brazil, along with all three local music labels from our work[9]. The release of this industrial dataset aims to foster future research activities on music recommender systems and local music consumption analysis. In addition, we are open-sourcing the entire Python source code of our experimental analysis, to ensure the

---

[9]Dataset available at https://zenodo.org/records/13309698.

reproducibility of our results. All materials are publicly available on GitHub [10].

5 CONCLUSION

In conclusion, although LFM-2b [37] is publicly available and serves as a convenient starting point for researchers inspecting music recommender systems, caution should be exercised when drawing conclusions about local music consumption based solely on this dataset. Our paper has emphasized significant differences in local music consumption patterns between LFM-2b and a proprietary dataset comprising the listening history of French, German, and Brazilian users of the music streaming service Deezer.

By replicating Lesota et al. [31]'s investigation of algorithmic biases in recommender systems for local music, we have also demonstrated that the two collaborative filtering models they analyzed, NeuMF [24] and ItemKNN [12], display varying biases on Deezer compared to LFM-2b. Moreover, we have identified several factors related to model training that had not been examined in this previous work and can significantly influence these biases, thereby modifying the study's overall conclusions.

Importantly, we have also explained that the proportion of local music consumed and recommended can vary significantly depending on the label source under consideration, its level of completeness, and biases introduced by human annotators [44]. While obtaining complete and universally accepted local music labels proves to be challenging, we have nonetheless recommended to prioritize the research in this direction. We have argued that this foundational labeling step is crucial for studies aiming to understand local music biases, as results based on unreliable labels may be misleading.

Overall, our work highlights the importance of using multiple model settings and data sources for cross-validation, and to ensure robust conclusions regarding the biases of music recommender systems – not only for local music but also potentially for other aspects such as gender and music genres. As a consequence, we have decided to publicly release our Deezer dataset along with this paper, including listening logs and labels from all three sources used in our experiments. We hope that this release of industrial resources will foster further research. As discussed in Section 4.2.3, our results come with certain limitations that open up interesting avenues for future analyses. Investigating these future directions would undoubtedly contribute to better measuring and enhancing the fairness of recommender systems on music streaming services.

FUNDING INFORMATION

REFERENCES

[1] Peter Achterberg, Johan Heilbron, Dick Houtman, and Stef Aupers. 2011. A Cultural Globalization of Popular Music? American, Dutch, French, and German Popular Music Charts (1965 to 2006). *American Behavioral Scientist* 55, 5 (2011), 589–608.
[2] Christine Bauer, Marta Kholodylo, and Christine Strauss. 2017. Music Recommender Systems Challenges and Opportunities for Non-Superstar Artists. In *Proceedings of the 30th Bled eConference*. 21–32.
[3] Christine Bauer and Markus Schedl. 2018. On the Importance of Considering Country-Specific Aspects on the Online-Market: an Example of Music Recommendation Considering Country-Specific Mainstream. In *Proceedings of the 51st Hawaii International Conference on System Sciences*. 3647–3656.
[4] Christine Bauer and Markus Schedl. 2019. Global and Country-Specific Mainstreaminess Measures: Definitions, Analysis, and Usage for Improving Personalized Music Recommendation Systems. *PloS One* 14, 6 (2019), e0217389.
[5] Pablo Bello and David Garcia. 2021. Cultural Divergence in Popular Music: the Increasing Diversity of Music Consumption on Spotify across Countries. *Humanities and Social Sciences Communications* 8, 1 (2021), 1–8.
[6] Walid Bendada, Théo Bontempelli, Mathieu Morlon, Benjamin Chapus, Thibault Cador, Thomas Bouabça, and Guillaume Salha-Galvan. 2023. Track Mix Generation on Music Streaming Services using Transformers. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 112–115.
[7] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender Systems Survey. *Knowledge-Based Systems* 46 (2013), 109–132.
[8] Léa Briand, Théo Bontempelli, Walid Bendada, Mathieu Morlon, François Rigaud, Benjamin Chapus, Thomas Bouabça, and Guillaume Salha-Galvan. 2024. Let's Get It Started: Fostering the Discoverability of New Releases on Deezer. In *Proceedings of the 46th European Conference on Information Retrieval*. Springer, 286–291.
[9] Léa Briand, Guillaume Salha-Galvan, Walid Bendada, Mathieu Morlon, and Viet-Anh Tran. 2021. A Semi-Personalized System for User Cold Start Recommendation on Music Streaming Apps. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2601–2609.
[10] Manuel Castells. 2011. *The Power of Identity*. John Wiley & Sons.
[11] Diana Crane. 2014. Cultural Globalization and the Dominance of the American Film Industry: Cultural Policies, National Film Industries, and Transnational Film. *International Journal of Cultural Policy* 20, 4 (2014), 365–382.
[12] Mukund Deshpande and George Karypis. 2004. Item-Based Top-N Recommendation Algorithms. *ACM Transactions on Information Systems (TOIS)* 22, 1 (2004), 143–177.
[13] Karlijn Dinnissen. 2022. Improving Fairness and Transparency for Artists in Music Recommender Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3498–3498.
[14] Karlijn Dinnissen and Christine Bauer. 2022. Fairness in Music Recommender Systems: A Stakeholder-Centered Mini Review. *Frontiers in Big Data* 5 (2022), 913608.
[15] Bora Edizel, Francesco Bonchi, Sara Hajian, André Panisson, and Tamir Tassa. 2020. FaiRecSys: Mitigating Algorithmic Bias in Recommender Systems. *International Journal of Data Science and Analytics* 9 (2020), 197–213.
[16] MusicBrainz: The Open Music Encyclopedia. 2024. https://musicbrainz.org/.
[17] Andres Ferraro. 2019. Music Cold-Start and Long-Tail Recommendation: Bias in Deep Representations. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 586–590.
[18] Andres Ferraro, Dmitry Bogdanov, Xavier Serra, and Jason Yoon. 2019. Artist and Style Exposure Bias in Collaborative Filtering Based Music Recommendations. In *ISMIR 2019 Workshop on Designing Human-Centric MIR Systems*.
[19] Andres Ferraro, Peter Knees, Tao Ye, Massimo Quadrana, and Fabien Gouyon. 2023. MuRS: Music Recommender Systems Workshop. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1227–1230.
[20] Michael Fuhr. 2015. *Globalization and Popular Music in South Korea: Sounding Out K-Pop*. Routledge.
[21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT.
[22] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2125–2126.
[23] Casper Hansen, Christian Hansen, Lucas Maystre, Rishabh Mehrotra, Brian Brost, Federico Tomasi, and Mounia Lalmas. 2020. Contextual and Sequential User Embeddings for Large-Scale Music Recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 53–62.
[24] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web*. 173–182.
[25] R Scott Hiller and Jason M Walter. 2017. The Rise of Streaming Music and Implications for Music Production. *Review of Network Economics* 16, 4 (2017), 351–385.
[26] Kurt Jacobson, Vidhya Murali, Edward Newett, Brian Whitman, and Romain Yon. 2016. Music Personalization at Spotify. *Proceedings of the 10th ACM Conference on Recommender Systems*, 373–373.
[27] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.
[28] Yehuda Koren and Robert Bell. 2015. Advances in Collaborative Filtering. *Recommender Systems Handbook* (2015), 77–118.
[29] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The Unfairness of Popularity Bias in Music Recommendation: A Reproducibility Study. In *Proceedings*

of the 42nd European Conference on Information Retrieval. Springer, 35–42.

[30] Oleg Lesota, Alessandro Melchiorre, Navid Rekabsaz, Stefan Brandl, Dominik Kowald, Elisabeth Lex, and Markus Schedl. 2021. Analyzing Item Popularity Bias of Music Recommender Systems: are Different Genders Equally Affected?. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 601–606.

[31] Oleg Lesota, Emilia Parada-Cabaleiro, Stefan Brandl, Elisabeth Lex, Navid Rekabsaz, and Markus Schedl. 2022. Traces of Globalization in Online Music Consumption Patterns and Results of Recommendation Algorithms. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference*. 291–297.

[32] Yang Li, Kangbo Liu, Ranjan Satapathy, Suhang Wang, and Erik Cambria. 2024. Recent Developments in Recommender Systems: A Survey. *IEEE Computational Intelligence Magazine* 19, 2 (2024), 78–95.

[33] David Morley. 2006. *Globalisation and Cultural Imperialism Reconsidered*. Routledge, London and New York.

[34] International Federation of the Phonographic Industry. 2023. *Engaging with Music*. IFPI Technical Report.

[35] ItemKNN Page on RecBole Documentation. 2024. https://recbole.io/docs/user_guide/model/general/itemknn.html.

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems* 32 (2019).

[37] Markus Schedl, Stefan Brandl, Oleg Lesota, Emilia Parada-Cabaleiro, David Penz, and Navid Rekabsaz. 2022. LFM-2b: A Dataset of Enriched Music Listening Events for Recommender Systems Research and Fairness Analysis. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*. 337–341.

[38] Markus Schedl, Peter Knees, Brian McFee, and Dmitry Bogdanov. 2021. Music Recommendation Systems: Techniques, Use Cases, and Challenges. In *Recommender Systems Handbook*. Springer, 927–971.

[39] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. 2018. Current Challenges and Visions in Music Recommender Systems Research. *International Journal of Multimedia Information Retrieval* 7, 2 (2018), 95–116.

[40] Dougal Shakespeare, Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. 2020. Exploring Artist Gender Bias in Music Recommendation. In *RecSys 2020 Workshop on the Impact of Recommender Systems*.

[41] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.

[42] John Tomlinson. 2001. *Cultural Imperialism: A Critical Introduction*. A&C Black.

[43] Quentin Villermet, Jérémie Poiroux, Manuel Moussallam, Thomas Louail, and Camille Roth. 2021. Follow the Guides: Disentangling Human and Algorithmic Curation in Online Music Consumption. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 380–389.

[44] Michael Zanger-Tishler, Julian Nyarko, and Sharad Goel. 2024. Risk Scores, Label Bias, and Everything but the Kitchen Sink. *Science Advances* 10, 13 (2024), eadi8411.

[45] Eva Zangerle and Christine Bauer. 2022. Evaluating Recommender Systems: Survey and Framework. *ACM Computing Surveys* 55, 8 (2022), 1–38.

[46] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, et al. 2021. RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*. 4653–4664.

Conclusion

Summary

In this thesis, we have explored some challenges related to the modeling and influencing of musical preferences through the lens of music streaming platforms. Our investigation spanned from understanding the nature of music streaming data to developing novel methods for quantifying musical taste and assessing the impact of recommendation algorithms. The work presented across the chapters of this thesis collectively addresses the complexity of these tasks and contributes to the broader understanding of how music preferences are shaped and can be analyzed in the digital age.

We began by examining the unique characteristics of music streaming data. This chapter focused on the inherent challenges of working with such data, including issues like missing information, the long-tail distribution of music consumption, and the complexities in interpreting user behavior. Understanding these challenges is crucial as they directly impact the design and effectiveness of computational methods used in subsequent analyses.

Moving further, we took interest in the interdisciplinary approaches to understanding musical taste. We examined sociological, psychological, culturological theories and computational models, highlighting the complexities in quantifying such a subjective and culturally nuanced concept as musical taste. This chapter set the stage for my own work by underscoring the limitations of current methods in capturing the richness of individual musical preferences.

Additionally, we reviewed existing methodologies for categorizing and representing music within computational frameworks. We explored how music is labeled and the different techniques used to measure similarity between music items. These representations are foundational for any subsequent analysis of musical preferences, as they shape how music is understood and processed by algorithms.

We then went through a technical overview of MRSs, and discussed the core challenges these systems face. These include the cold start problem, context awareness, balancing familiarity with discovery, and ensuring fairness towards different stakeholders. This review underscored the importance of how musical

data and preferences are modeled, as these factors significantly influence the performance and biases of RSs. This chapter also set the stage for my own work by highlighting the limitations and biases that can arise in these systems, particularly in relation to how they influence and reflect musical preferences.

Finally, we went through my original contributions to the field. First, we introduced a novel approach to modeling musical taste through the concept of 'musical taste fingerprint' — a method to quantify and capture an individual's musical preferences using data from a music streaming platform. This research tackled two contrasting perspectives on musical taste: one that emphasizes uniqueness, where a fingerprint is defined by the items that uniquely identify an individual, and another that focuses on representativeness, summarizing the breadth of a user's preferences in a way that can be leveraged for recommendation. Through extensive experimentation, we demonstrated that these two perspectives often lead to conflicting solutions, underscoring the complexity of accurately modeling musical taste. The findings also revealed significant privacy implications, as users could be uniquely identified by only a small set of preferences, raising important questions about data privacy on streaming platforms. This work highlighted the importance of explicitly acknowledging the objectives when defining and computing musical taste in computational settings, providing a foundational method for future work in both personalized recommendations and privacy-preserving systems.

The second major contribution of this thesis involved exploring the influence of RSs on local music consumption, particularly examining how algorithmic biases can shift music recommendations either towards or against local content. By reproducing the findings of a prior study using both public and proprietary datasets, it became clear that the impact of these biases varies significantly depending on the dataset, model parameters, and labeling accuracy. The analysis revealed discrepancies between different datasets and their associated sets of users, emphasizing the importance of using multiple data sources to ensure robust conclusions. Additionally, the complexity of accurately labeling music was underscored, as variations in labeling methods can lead to different interpretations and outcomes. This work emphasizes the need for careful consideration of algorithmic choices in RSs to support a fair and culturally diverse music landscape.

Future work

In this thesis, we have addressed several important questions, but many others remain unresolved. While our contributions have provided valuable insights, there is still room for improvement, and they have also raised new questions that warrant further investigation. Additionally, throughout this work, we have touched upon numerous topics in the literature that, while relevant, could not

be fully explored within the limited time frame of a PhD thesis. They however represent potential areas for future exploration.

Toward an ideal musical taste fingerprint

In our study on musical taste fingerprints, we highlighted how different definitions can lead to diverging solutions. While this analysis is insightful, it raises a fundamental question: what constitutes the 'ideal' fingerprint? Does such an ideal exist, or is it context-dependent? A potential approach to answering this would involve directly engaging users. For instance, users could be asked to rate generated fingerprints based on their relevance, or create their own fingerprints, which we could then analyze to understand the underlying patterns or preferences. This user-driven approach could provide valuable insights into the characteristics that make a fingerprint effective and pave the way for computational methods to replicate these characteristics.

Additionally, it would be beneficial to evaluate the performance of these fingerprints in real-world recommendation scenarios. For example, could these fingerprints help users transition seamlessly from one platform to another? Currently, most streaming platforms represent users as vectors within their own distinct recommendation space, built from interaction histories and preferences. As each platform develops its own approach to modeling user preferences, this creates a barrier for users who wish to transfer their music profiles between services. While some platforms now allow users to import favorite tracks or artists from another service, this rarely provides the depth of personalization needed for high-quality recommendations. Typically, users must rebuild their listening history on a new platform to receive accurate suggestions. Introducing a universal musical taste fingerprint, independent of platform-specific algorithms, could help resolve this issue. Such a fingerprint would be a standardized representation of a user's musical preferences that could be transferred between platforms, allowing for a seamless experience across services. This could not only improve the user experience by providing better recommendations immediately upon switching platforms but also enhance interoperability in the streaming ecosystem.

Ensuring privacy in open-access datasets

While data anonymization was not the primary focus of this thesis, it emerged as a critical area for further exploration, especially given the growing demand within the research community for publicly available datasets.

Most datasets today originate from private streaming services, which face significant challenges when it comes to data sharing. These companies are often particularly concerned about sharing data because of the financial stakes involved; the data may reveal proprietary insights or compromise competitive

advantages. Moreover, they have a responsibility to protect their users' privacy, which adds another layer of complexity to public research. To address these concerns, future research should investigate advanced anonymization techniques that balance the need for data utility with the protection of personal information.

Differential privacy (DP), for instance, could be a valuable tool in this context. By adding some noise to the data, or partially removing some users or user-item interactions, DP allows for the sharing of useful aggregated information without revealing specific user details, thereby protecting both the company's proprietary knowledge and user privacy. An even more secure variation, local differential privacy (LDP), introduces noise directly on each user's data locally before it is transferred to the server, offering stronger privacy guarantees by ensuring that even the data collectors cannot reconstruct the original data accurately. This makes LDP particularly useful for scenarios where highly sensitive information is involved.

However, even when data is aggregated or pseudonymized (where identifiable details are replaced with fake identifiers) it may not fully guarantee privacy. For example, membership inference attacks - a type of privacy breach where an attacker attempts to determine whether a specific individual's data is included in a dataset - can re-identify individuals within an anonymized dataset. Troncoso et al. (2020) demonstrated that aggregate location data is vulnerable to such attacks because patterns in the data, such as frequent visits to certain locations or unique movement behaviors, can still be linked back to specific individuals.

Another promising avenue is the use of generated data or synthetic models that can replicate the characteristics of actual data without exposing sensitive information. By training models on real data and then using these models to generate synthetic datasets, researchers could have access to valuable insights while ensuring that the original data remains secure. This approach could help bridge the gap between the need for data sharing in public research and the privacy and commercial concerns of streaming services. Developing such privacy-preserving methods is essential not only for fostering collaboration and innovation within the field but also for ensuring that public research can access the data necessary to drive advancements without compromising commercial interests or user trust.

Addressing the lack of labels in music catalogs

One of the key challenges uncovered in this thesis is the significant lack of reliable spatial information associated with artists (such as country/region/city of origin, city of "residence", history of tour locations, etc.) for artists, which poses a fundamental problem for understanding the global music landscape. Without accurate country information, it is difficult to answer simple questions, such as estimate how many artists from different origins are represented in streaming

platforms' catalogs. This, in turn, means we cannot accurately assess the number of streams that artists from different countries receive. Consequently, we are unable to study the importance of local music in various countries, making it difficult to understand and support the cultural diversity and significance of local music scenes globally. This labeling issue also directly impacts the fairness of music recommendations. Without reliable data on artists' origins, it becomes impossible to measure whether local artists are being fairly represented in recommendations, both within their home countries and internationally. The lack of accurate spatial information hinders our ability to analyze and improve RSs to ensure they promote local music appropriately, or more generally are able to add geographical ingredients in content recommendation. Associated features could include the recommendation of local bands to a user when traveling, or spatial operators in search for musical content.

There have been some efforts towards addressing these issues. For example, the open music encyclopedia MusicBrainz offers a large-scale, community-driven database of artist metadata, including artist origin. Wikidata[1] provides a more general-purpose, structured knowledge base that includes detailed geographical data on artists and can be cross-referenced with other datasets. Discogs[2] contributes by providing an extensive user-curated database of discographies, focusing on release and label information that can help trace music distribution and cultural impact. Freesound[3] emphasizes audio content and annotations, offering metadata related to sound samples that could help in music classification and analysis. Despite these varied approaches, significant gaps remain, particularly for lesser-known artists.

To address these challenges, future work should focus on two main strategies:

1. Crossing different information sources: A multi-source approach could help mitigate the gaps in data that currently prevent us from understanding local music consumption and recommendation fairness. By gathering data from a variety of sources, we can cross-check and enrich the information available, improving the accuracy of artist labeling.

2. Collective effort in labeling: Unfortunately, the artists who are least likely to have accurate labels are often those who are less popular or come from more 'niche' countries. This means that even with additional datasets, certain artists will likely continue to be underrepresented. To combat this, a collective effort is needed to label as many artists as possible. Techniques such as web scraping, lyrics language detection, and other automated methods could be employed to generate country labels where they are missing. By systematically improving the coverage of artist country

---

[1]wikidata.org

[2]discogs.com

[3]freesound.org

labels, we can better estimate the representation and consumption of local music, ultimately leading to more equitable RSs.

In summary, enhancing the accuracy of artist labels is crucial for both understanding local music consumption patterns and ensuring fairness in music recommendations. By addressing these data gaps, we can enable more comprehensive research and more equitable outcomes in music streaming in general.

Addressing diverse users' expectations in music recommendations

All users may not have the same expectations from algorithmic recommendations: some might prefer to stay within their comfort zone, using algorithmically generated playlists to listen to familiar songs and artists, with minimal interest in discovery. Others might seek primarily discovery, not limited to similar music to their usual preferences — similarly to the experience of exploring random vinyls in a record store. It is likely that most users fall somewhere along a spectrum between these two extremes.

Given this possible diversity in expectations, it seems unlikely that a single recommendation algorithm could satisfy all users equally. For instance, algorithms that favor familiar content might not appeal to users who are looking for novelty, and vice versa. This mismatch could potentially lead to dissatisfaction among users whose expectations are not met, which could be perceived as a form of unfairness.

Studies like Celma (2010) and Schedl and Hauger (2015) explore the fairness of MRS toward users with different music consumption patterns, often assuming users' expectations based solely on streaming behavior, such as preferences for mainstream versus niche content.

Yet, it's uncertain if we can draw direct parallels like these. For example, a user with narrow, mainstream tastes might still be open to more diverse recommendations, but streaming history alone may not reveal this. Mehrotra et al. (2019) address a similar challenge by jointly leveraging both user intent and interaction signals. Their study demonstrates that relying solely on behavioral data — such as streaming logs — without accounting for user intent can lead to inaccurate predictions of user satisfaction, emphasizing the need for more nuanced, mixed-method approaches.

Colleagues from the RECORDS project have collected and analyzed rich interview data that could provide deeper insights into user expectations beyond what streaming logs reveal. While incorporating this interview data was beyond the scope of my PhD due to time and resource constraints, I believe it holds great potential for future work. In general, I hope to explore personalizing and diversifying recommendations further in future research.

Exploring diffusion of music through time and space

Although this thesis did not directly address the topic, the propagation of music through time and space is a subject of personal interest to me. During my PhD, I co-supervised two internships that explored related areas.

Predicting a song's popularity evolution. In the first internship, conducted by Basile Leretaille, we aimed to understand whether it is possible to predict and model the evolution of a song's popularity over time. On an individual level, research suggests that interest in a song often follows an inverted U-shaped curve: as we repeatedly listen to a song, our enjoyment typically increases until it reaches a peak, after which we become saturated and our interest gradually declines Berlyne (1973); Sguerra et al. (2022). We wondered if a similar pattern could be observed on a collective level.

We were initially inspired by models from epidemiology, which have been widely used to study the spread of information and behaviors across populations Pastor-Satorras and Vespignani (2001). To avoid black-box solutions like neural networks, which, despite their effectiveness, often lack transparency in how they make predictions, we decided not to pursue well-established neural network-based approaches such as Andreas et al. (2020). Instead, we turned to compartmental models, which allow us to divide populations into different categories and simulate how songs "spread" through user interactions. Additionally, we explored graph-based propagation models that account for the structure of the listener network (Kempe et al., 2003), as well as cellular automata models, inspired by systems like Conway's Game of Life (Gardner, 1970), which offer transparent mechanisms for simulating local interactions between users and songs.

However, we observed, through both techniques, that there was no universal pattern for collective consumption like there seems to exist on an individual level. While some songs do experience a gradual rise and fall in popularity, many follow entirely different trajectories. For instance, a song released years ago may suddenly gain popularity 'overnight' due to external factors, such as being featured by a popular influencer on social media or appearing in a movie [4]. This type of sudden surge in popularity is often driven by external circumstances rather than the music itself, making it difficult to predict.

The decline in popularity, on the other hand, might be influenced by factors like the song's complexity or catchiness. However, demonstrating this link would require a deep dive into the audio characteristics of the music itself, which we did not have the time or resources to explore in this study. This aspect also connects to the broader question of the 'hit formula' — to date, research has not identified any specific musical pattern or audio trait that guarantees a track's

---

[4]RailsConf 2016 - Closing Keynote: Paul Lamere

success Raza and Nanath (2020); Zangerle et al. (2019). Understanding these dynamics would likely require extensive research, combining large-scale data analysis with detailed audio feature studies. Moreover, some songs experience multiple revivals, becoming popular again at different points in time, often due to changing cultural contexts or trends. Others follow more predictable cycles, such as summer hits or Christmas songs that consistently resurface during certain seasons. When we began this internship, we were perhaps overly optimistic, not fully appreciating the complexity of these patterns. In hindsight, I realize that understanding a song's popularity evolution could be a PhD subject in itself, given the many layers and variables involved.

Spatial dynamics of new releases. In another internship, conducted by Alvin Opler, we explored the geographical patterns in the spatial propagation of music, that is the propagation of the listening of a new musical piece from the moment it is made available on a global streaming platform, therefore available at the same time in all the countries where this platform operates. Our focus was on several artists and songs, both French and global, across different levels of popularity, to determine if there were differences in the speed and geographical trajectory of their diffusion.

To better understand these diffusion patterns, we used the Family and Vicsek (1985) scaling method, a model from statistical physics, to observe how roughness or unevenness in the spread of new music evolves over time and distance. This approach helped us quantify the spatial spread of a song's popularity, analyzing the simultaneous growth of its reach in both local and distant regions.

For globally famous artists, as expected, we found that when they launch an album or single, their music tends to become popular simultaneously across various regions. This is due to their established notoriety and the power of the internet, which ensures that fans around the world are aware of the release at the same time. A similar pattern was observed in France for very famous French artists. However, we hypothesized that the situation would be different for emerging French artists, particularly those whose musical identity, as expressed by their style or lyrics, is closely tied to their place of origin, such as rap from Marseille or Breton music. We expected that these artists would first gain popularity in their local region, with their music gradually spreading to the rest of the country, similarly to an epidemic caused by a virus. This hypothesis was inspired by older studies, like Ford (1971)'s 1970 qualitative study on the diffusion of rock music in the US, over a much larger timescale of several decades.

To test our hypothesis, we attempted to map the listening of the successive releases of recently emerging artists, from the beginning of their careers. However, we did not find strong confirmation of our expectations. This preliminary study suggests that, in the current digital age, music diffusion may rely less on

local, physical, human-to-human interactions and more on global virtual channels. Music seems to spread primarily through digital platforms rather than following the traditional patterns of geographic diffusion. Despite these findings, it is important to explore this subject further to draw more definitive conclusions. It's possible that our methodology had limitations, or that the scale of our study was too narrow. For example, examining music diffusion on a larger scale, such as between countries, might reveal patterns that are not evident within a relatively small country like France.

That being said, we still observed some discrepancies in music preferences across different geographic areas, with clustering by geographic proximity often correlating with genre. This indicates that there are still local specificities in music preferences, suggesting that geography does play a role in shaping musical tastes. There is definitely room for further exploration in this area, particularly in understanding how geographic location influences music preferences and vice versa.

Even though these two internship projects did not necessarily lead conclusive results or match our initial expectations, they laid important groundwork for these questions. These foundational efforts have opened up exciting possibilities for future research, which would be valuable to explore further.

Bibliography

P. Achterberg, J. Heilbron, D. Houtman, and S. Aupers. A cultural globalization of popular music? american, dutch, french, and german popular music charts (1965 to 2006). *American behavioral scientist*, 55(5):589–608, 2011.

G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

D. Afchar. *Interpretable Music Recommender Systems*. PhD thesis, Sorbonne Université, 2023.

D. Afchar, R. Hennequin, and V. Guigue. Of spiky svds and music recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 926–932, 2023.

N. Aizenberg, Y. Koren, and O. Somekh. Build your own music recommender by modeling internet radio streams. In *Proceedings of the 21st international conference on World Wide Web*, pages 1–10, 2012.

R. Alaei, N. O. Rule, and G. MacDonald. Individuals' favorite songs' lyrics reflect their attachment style. *Personal Relationships*, 29(4):778–794, 2022.

A. Anderson, L. Maystre, I. Anderson, R. Mehrotra, and M. Lalmas. Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of The Web Conference*, pages 2155–2165, 2020.

I. Anderson, S. Gil, C. Gibson, S. Wolf, W. Shapiro, O. Semerci, and D. M. Greenberg. "Just the Way You Are": Linking music listening on Spotify and personality. *Social Psychological and Personality Science*, 12(4):561–572, 2021.

A. Andreas, C. X. Mavromoustakis, G. Mastorakis, S. Mumtaz, J. M. Batalla, and E. Pallis. Modified machine learning techique for curve fitting on regression models for covid-19 projections. In *2020 IEEE 25th international workshop on computer aided modeling and design of communication links and networks (CAMAD)*, pages 1–6. IEEE, 2020.

M. Anglada-Tort, H. Lee, A. E. Krause, and A. C. North. Here comes the sun: music features of popular songs reflect prevailing weather conditions. *Royal Society Open Science*, 10(5):221443, 2023.

C. Baccigalupo, E. Plaza, and J. Donaldson. Uncovering affinity of artists to multiple genres from social behaviour data. In *ISMIR*, pages 275–280, 2008.

H. Bahuleyan. Music genre classification using machine learning techniques. *arXiv preprint arXiv:1804.01149*, 2018.

L. Baltrunas, M. Kaminskas, B. Ludwig, O. Moling, F. Ricci, A. Aydin, K.-H. Lüke, and R. Schwaiger. Incarmusic: Context-aware music recommendations in a car. In *E-Commerce and Web Technologies: 12th International Conference, EC-Web 2011, Toulouse, France, August 30-September 1, 2011. Proceedings 12*, pages 89–100. Springer, 2011.

B. K. Baniya, J. Lee, and Z.-N. Li. Audio feature reduction and analysis for automatic music genre classification. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 457–462. IEEE, 2014.

F. S. Barrett, K. J. Grimm, R. W. Robins, T. Wildschut, C. Sedikides, and P. Janata. Music-evoked nostalgia: affect, memory, and personality. *Emotion*, 10(3):390, 2010.

C. Bauer and M. Schedl. On the importance of considering country-specific aspects on the online-market: an example of music recommendation considering country-specific mainstream. 2018.

C. Bauer and M. Schedl. Global and country-specific mainstreaminess measures: Definitions, analysis, and usage for improving personalized music recommendation systems. *PloS one*, 14(6):e0217389, 2019.

P. Bello and D. Garcia. Cultural divergence in popular music: the increasing diversity of music consumption on spotify across countries. *Humanities and Social Sciences Communications*, 8(1):1–8, 2021.

W. Bendada, T. Bontempelli, M. Morlon, B. Chapus, T. Cador, T. Bouabça, and G. Salha-Galvan. Track mix generation on music streaming services using transformers. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 112–115, 2023.

D. E. Berlyne. Aesthetics and psychobiology. *Journal of Aesthetics and Art Criticism*, 31(4), 1973.

T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proceedings of the 12th international society for music information retrieval conference (ISMIR)*, volume 11, pages 591–596, 2011.

A. S. Bhat, V. Amith, N. S. Prasad, and D. M. Mohan. An efficient classification algorithm for music mood detection in western and hindi music using audio feature extraction. In *2014 fifth international conference on signal and image processing*, pages 359–364. IEEE, 2014.

A. Bonneville-Roussy, P. J. Rentfrow, M. K. Xu, and J. Potter. Music through the ages: Trends in musical engagement and preferences from adolescence through middle adulthood. *Journal of personality and social psychology*, 105 (4):703, 2013.

T. Bontempelli, B. Chapus, F. Rigaud, M. Morlon, M. Lorant, and G. Salha-Galvan. Flow moods: Recommending music by moods on deezer. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 452–455, 2022.

P. Bourdieu. *Distinction – A Social Critique of the Judgement of Taste*. Harvard University Press, 1984.

R. A. Brown. Music preferences and personality among Japanese university students. *International Journal of Psychology*, 47(4):259–268, 2012.

B. Bryson. "Anything but heavy metal": Symbolic exclusion and musical dislikes. *American Sociological Review*, pages 884–899, 1996.

R. Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.

R. Burke. Hybrid web recommender systems. In *The adaptive web*, pages 377–408. Springer, 2007.

Z. Cai, L. Fu, and W. Li. Research and analysis of music development based on k-means and pca algorithm. In *Journal of Physics: Conference Series*, volume 2083, page 032044. IOP Publishing, 2021.

C. L. Caldwell-Harris. Emotionality differences between a native and foreign language: Theoretical implications. *Frontiers in psychology*, 5:93402, 2014.

P. Cano and M. Koppenberger. The emergence of complex network patterns in music artist networks. In *Proceedings of the 5th international symposium on music information retrieval (ISMIR 2004)*, pages 466–469, 2004.

J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, 1998.

G. Carney. Music geography. *Journal of Cultural Geography*, 18(1):1–10, 1998.

G. O. Carney. From down home to uptown: the diffusion of country-music radio stations in the united states. *Journal of Geography*, 76(3):104–110, 1977.

M. Castelluccio. The music genome project. *Strategic Finance*, pages 57–59, 2006.

Z. Cataltepe, Y. Yaslan, and A. Sonmez. Music genre classification using midi and audio features. *EURASIP Journal on Advances in Signal Processing*, 2007:1–8, 2007.

O. Celma. Music recommendation. In *Music recommendation and discovery: The long tail, long fail, and long play in the digital music space*, pages 43–85. Springer, 2010.

Ò. Celma and P. Cano. From hits to niches? or how popular artists can bias music recommendation and discovery. In *Proceedings of the 2nd KDD work-*

shop on large-scale recommender systems and the netflix prize competition, pages 1–8, 2008.

Ò. Celma Herrada et al. *Music recommendation and discovery in the long tail*. Universitat Pompeu Fabra, 2009.

T. Chamorro-Premuzic and A. Furnham. Personality and music: Can traits explain how people use music in everyday life? *British journal of psychology*, 98(2):175–185, 2007.

W. Chen, P. Ren, F. Cai, F. Sun, and M. de Rijke. Improving end-to-end sequential recommendations with intent-aware diversification. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 175–184, 2020.

Z. Cheng and J. Shen. Just-for-me: an adaptive personalization system for location-aware social music recommendation. In *Proceedings of international conference on multimedia retrieval*, pages 185–192, 2014.

K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics, 2014.

K. Choi, G. Fazekas, and M. Sandler. Automatic tagging using deep convolutional neural networks. *arXiv preprint arXiv:1606.00298*, 2016.

R. Cilibrasi, P. Vitányi, and R. De Wolf. Algorithmic clustering of music. In *Proceedings of the Fourth International Conference onWeb Delivering of Music, 2004. EDELMUSIC 2004.*, pages 110–117. IEEE, 2004.

W. W. Cohen and W. Fan. Web-collaborative filtering: Recommending music by crawling the web. *Computer Networks*, 33(1-6):685–698, 2000.

P. Coulangeon. Social stratification of musical tastes: questioning the cultural legitimacy model. *Revue française de sociologie*, 46(5):123–154, 2005.

A. S. Cowen, X. Fang, D. Sauter, and D. Keltner. What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures. *Proceedings of the National Academy of Sciences*, 117(4):1924–1934, 2020.

F. Darke and L. Below Blomkvist. Categorization of songs using spectral clustering, 2021.

H. Datta, G. Knox, and B. J. Bronnenberg. Changing their tune: How consumers' adoption of online streaming affects music consumption and discovery. *Marketing Science*, 37(1):5–21, 2018.

M. De Gemmis, P. Lops, C. Musto, F. Narducci, and G. Semeraro. Semantics-

aware content-based recommender systems. *Recommender systems handbook*, pages 119–159, 2015.

Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, 3(1): 1–5, 2013.

M. De Witte, A. Spruit, S. van Hooren, X. Moonen, and G.-J. Stams. Effects of music interventions on stress-related outcomes: a systematic review and two meta-analyses. *Health psychology review*, 14(2):294–324, 2020.

C. DeHart. Metal by numbers: Revisiting the uneven distribution of heavy metal music. *Metal Music Studies*, 4(3):559–571, 2018.

R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam. Music mood detection based on audio and lyrics with deep neural net. *arXiv preprint arXiv:1809.07276*, 2018.

M. J. Delsing, T. F. Ter Bogt, R. C. Engels, and W. H. Meeus. Adolescents' music preferences and personality characteristics. *European Journal of Personality*, 22(2):109–130, 2008.

C. Desrosiers and G. Karypis. A comprehensive survey of neighborhood-based recommendation methods. In *Recommender Systems Handbook*, pages 107–144. Springer, 2011.

K. Dinnissen and C. Bauer. Fairness in music recommender systems: A stakeholder-centered mini review. *Frontiers in big Data*, 5:913608, 2022.

K. Dinnissen and C. Bauer. Amplifying artists' voices: Item provider perspectives on influence and fairness of music streaming platforms. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 238–249, 2023.

M. D. Ekstrand, M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, and M. S. Pera. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Conference on fairness, accountability and transparency*, pages 172–186. PMLR, 2018.

M. Elahi, F. Ricci, and N. Rubens. A survey of active learning in collaborative filtering recommender systems. *Computer Science Review*, 20:29–50, 2016.

E. V. Epure, G. Salha, M. Moussallam, and R. Hennequin. Modeling the music genre perception across language-bound cultures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4765–4779, 2020.

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231. AAAI Press, 1996.

F. Fabbri. What kind of music? *Popular music*, 2:131–143, 1982.

F. Family and T. Vicsek. Scaling of the active zone in the eden process on percolation networks and the ballistic deposition model. *Journal of Physics A: Mathematical and General*, 18(2):L75–L81, 1985.

S. Feld. A sweet lullaby for world music. *Public culture*, 12(1):145–172, 2000.

A. Ferraro. Music cold-start and long-tail recommendation: bias in deep representations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 586–590, 2019.

A. Ferraro, X. Serra, and C. Bauer. Break the loop: Gender imbalance in music recommenders. In *Proceedings of the 2021 conference on human information interaction and retrieval*, pages 249–254, 2021a.

A. Ferraro, X. Serra, and C. Bauer. What is fair? exploring the artists' perspective on the fairness of music streaming platforms. In *IFIP conference on human-computer interaction*, pages 562–584. Springer, 2021b.

B. Ferwerda, M. Schedl, and M. Tkalcic. Personality & emotional states: Understanding users' music listening needs. In *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 23rd Conference on User Modeling, Adaptation, and Personalization (UMAP 2015) Dublin, Ireland, June 29-July 3, 2015*, volume 1388. CEUR-WS. org, 2015.

L. Ford. Geographic factors in the origin, evolution, and diffusion of rock and roll music. *Journal of Geography*, 70(8):455–464, 1971.

M. Forman. *The'hood comes first: Race, space, and place in rap and hip-hop*. Wesleyan University Press, 2002.

S. Frith. *Performing rites: On the value of popular music*. Harvard University Press, 1996.

X. Gao, C. Gupta, and H. Li. Lyrics transcription and lyrics-to-audio alignment with music-informed acoustic models. *MIREX*, 2021.

J. Garcia-Gathright, B. St. Thomas, C. Hosey, Z. Nazari, and F. Diaz. Understanding and evaluating user satisfaction with music discovery. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 55–64, 2018.

M. Gardner. Mathematical games: The fantastic combinations of john conway's new solitaire game "life". *Scientific American*, 223(4):120–123, 1970.

N. Gasser. *Why you like it: The science and culture of musical taste*. Flatiron Books, 2019.

D. George, K. Stickle, F. Rachid, and A. Wopnford. The association between types of music enjoyed and cognitive, behavioral, and personality factors of those who listen. *Psychomusicology: A Journal of Research in Music Cognition*, 19(2):32, 2007.

R. Glott, P. Schmidt, and R. Ghosh. Wikipedia survey–overview of results.

United Nations University: Collaborative Creativity Group, 8:1158–1178, 2010.

E. Gómez, M. Haro, and P. Herrera. Music and geography: Content description of musical audio from different parts of the world. In *ISMIR*, pages 753–758, 2009.

Gracenote. Feel like listening? computing musical mood at gracenote. `https://www.gracenote.com/computing-musical-mood-at-gracenote/`, 2016. Accessed: 2016-08-09.

M. Haim, A. Graefe, and H.-B. Brosius. Burst of the filter bubble? effects of personalization on the diversity of google news. *Digital journalism*, 6(3): 330–343, 2018.

D. J. Hargreaves. The development of aesthetic reactions to music. *Psychology of Music*, 1982.

D. J. Hargreaves. The effects of repetition on liking for music. *Journal of research in Music Education*, 32(1):35–47, 1984.

D. J. Hargreaves, A. C. North, and M. Tarrant. *Musical preference and taste in childhood and adolescence.* Oxford University Press, 2006.

F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

R. He and J. McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th international conference on world wide web*, pages 507–517, 2016.

X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182, 2017.

D. Hebert and M. Rykowski. *Music glocalization: Heritage and innovation in a digital age.* Cambridge Scholars Publishing, 2018.

J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, pages 241–250, 2000.

B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*, 2016.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

M. B. Holbrook and R. M. Schindler. Some exploratory findings on the development of musical tastes. *Journal of consumer research*, 16(1):119–124, 1989.

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

J. L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30:179–185, 1965.

Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 263–272. IEEE, 2008.

E. J. Humphrey, J. P. Bello, and Y. LeCun. Feature learning and deep architectures: New directions for music informatics. *Journal of Intelligent Information Systems*, 41(3):461–481, 2013.

K. M. Ibrahim, J. Royo-Letelier, E. V. Epure, G. Peeters, and G. Richard. Audio-based auto-tagging with contextual tags for music. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16–20. IEEE, 2020.

K. Jakubowski, T. Eerola, B. Tillmann, F. Perrin, and L. Heine. A cross-sectional study of reminiscence bumps for music-related memories in adulthood. *Music & Science*, 3:2059204320965058, 2020.

J. Jin. Fast network community detection by score. 2013.

Y. Jin, N. Tintarev, and K. Verbert. Effects of personal characteristics on music recommender systems with different levels of controllability. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 13–21, 2018.

M. Kaminskas and F. Ricci. Contextual music information retrieval and recommendation: State of the art and challenges. *Computer Science Review*, 6(2-3):89–119, 2012.

M. Kaminskas, F. Ricci, and M. Schedl. Location-aware music recommendation using auto-tagging and hybrid matching. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 17–24, 2013.

B. Karrer and M. E. Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

I. Karydis, A. Nanopoulos, H.-H. Gabriel, and M. Spiliopoulou. Tag-aware spectral clustering of music items. In *ISMIR*, pages 159–164, 2009.

D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, 2003.

A. Khorsandi and S. Saarikallio. Music-related nostalgic experiences of young migrants. In *The 3rd International Conference on Music & Emotion, Jyväskylä, Finland, June 11-15, 2013*. University of Jyväskylä, Department of Music, 2013.

D. Kim, K.-s. Kim, K.-H. Park, J.-H. Lee, and K. M. Lee. A music recom-

mendation system with a dynamic k-means clustering algorithm. In *Sixth international conference on machine learning and applications (ICMLA 2007)*, pages 399–403. IEEE, 2007.

J. Kim, M. Won, C. C. Liem, and A. Hanjalic. Towards seed-free music playlist generation: Enhancing collaborative filtering with playlist title information. In *Proceedings of the ACM Recommender Systems Challenge 2018*, pages 1–6. 2018.

K. Kim, N. Askin, and J. A. Evans. Disrupted routines anticipate musical exploration. *Proceedings of the National Academy of Sciences*, 121(6): e2306549121, 2024.

P. Knees and M. Schedl. A survey of music similarity and recommendation from music context data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 10(1):1–21, 2013.

P. Knees, E. Pampalk, and G. Widmer. Artist classification with web-based data. In *ISMIR*, 2004.

P. Knees, M. Schedl, and M. Goto. Intelligent user interfaces for music discovery. *Trans. Int. Soc. Music. Inf. Retr.*, 3(1):165–179, 2020.

Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

D. Kowald, M. Schedl, and E. Lex. The unfairness of popularity bias in music recommendation: A reproducibility study. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 35–42. Springer, 2020.

A. E. Krause, A. C. North, and L. Y. Hewitt. Music-listening in everyday life: Devices and choice. *Psychology of music*, 43(2):155–170, 2015.

A. Kulesza, B. Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

U. Kużelewska and K. Wichowski. A modified clustering algorithm dbscan used in a collaborative filtering recommender system for music recommendation. In *Theory and Engineering of Complex Systems and Dependability: Proceedings of the Tenth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX, June 29–July 3 2015, Brunów, Poland*, pages 245–254. Springer, 2015.

H. Kwon, J. Han, and K. Han. Art (attractive recommendation tailor) how the diversity of product recommendations affects customer purchase preference in fashion industry? In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2573–2580, 2020.

B. Lahire. The individual and the mixing of genres: Cultural dissonance and self-distinction. *Poetics*, 36(2-3):166–188, 2008.

X. N. Lam, T. Vu, T. D. Le, and A. D. Duong. Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd international conference on Ubiquitous information management and communication*, pages 208–211, 2008.

A. Langmeyer, A. Guglhör-Rudan, and C. Tarnai. What is your music preference telling us about your personality? In *Poster presented at the 14th European Conference on Personality, Tartu, Estonia*, 2008.

A. Langmeyer, A. Guglhör-Rudan, and C. Tarnai. What do music preferences reveal about personality? A cross-cultural replication using self-ratings and ratings of music samples. *Journal of Individual Differences*, 33(2):119, 2012.

A. Laplante. Improving music recommender systems: What can we learn from research on music tastes? In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 451–456, 2014.

F. Le Bel. *Agglomerative clustering for audio classification using low-level descriptors*. PhD thesis, Ircam UMR STMS 9912, 2017.

A. LeBlanc and R. Cote. Effects of tempo and performing medium on children's music preference. *Journal of Research in Music Education*, 31(1):57–66, 1983.

A. LeBlanc, W. L. Sims, C. Siivola, and M. Obert. Music style preferences of different age listeners. *Journal of Research in Music Education*, 44(1):49–59, 1996.

D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

H. Lee, N. Jacoby, R. Hennequin, and M. Moussallam. Tracing the mechanisms of cultural diversity through 2.5 million individuals' music listening patterns. *Preprint at https://doi. org/10.31234/osf. io/73kyf*, 2024.

J. Lei and A. Rinaldo. Consistency of spectral clustering in stochastic block models. 2015.

O. Lesota, A. Melchiorre, N. Rekabsaz, S. Brandl, D. Kowald, E. Lex, and M. Schedl. Analyzing item popularity bias of music recommender systems: are different genders equally affected? In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 601–606, 2021.

O. Lesota, E. Parada-Cabaleiro, S. Brandl, E. Lex, N. Rekabsaz, and M. Schedl. Traces of globalization in online music consumption patterns and results of recommendation algorithms. In *Ismir 2022 Hybrid Conference*, 2022.

J. Li, B. Shao, T. Li, and M. Ogihara. Hierarchical co-clustering: a new way to organize the music data. *IEEE Transactions on Multimedia*, 14(2):471–481, 2011.

R. Z. Li, J. Urbano, and A. Hanjalic. Leave no user behind: Towards improving the utility of recommender systems for non-mainstream users. In *Proceedings*

of the 14th ACM International Conference on Web Search and Data Mining, pages 103–111, 2021.

D. Liang, R. G. Krishnan, M. D. Hoffman, and T. Jebara. Variational autoencoders for collaborative filtering. In *Proceedings of the World Wide Web Conference*, pages 689–698, 2018.

B. Lika, K. Kolomvatsos, and S. Hadjiefthymiades. Facing the cold start problem in recommender systems. *Expert systems with applications*, 41(4):2065–2073, 2014.

J. Liu, C. B. Hilton, E. Bergelson, and S. A. Mehr. Language experience predicts music processing in a half-million speakers of fifty-four languages. *Current Biology*, 33(10):1916–1925, 2023.

Q. Liu, A. Karatzoglou, L. Cai, and X. He. Integrating spatial and temporal contexts into a factorization model for poi recommendation. *International Journal of Geographical Information Science*, 32(3):1–23, 2017. doi: 10. 1080/13658816.2017.1400550.

P. Lops, M. de Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. In *Recommender Systems Handbook*, pages 73–105. Springer, 2011.

C. Louven. Hargreaves'"open-earedness": A critical discussion and new approach on the concept of musical tolerance and curiosity. *Musicae Scientiae*, 20(2):235–247, 2016.

F. Lu and N. Tintarev. A diversity adjusting strategy with personality for music recommendation. In *IntRS@ RecSys*, pages 7–14, 2018.

L. Lu, D. Liu, and H.-J. Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on audio, speech, and language processing*, 14(1):5–18, 2005.

L. Lü, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou. Recommender systems. *Physics reports*, 519(1):1–49, 2012.

J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

J. P. Mahedero, Á. MartÍnez, P. Cano, M. Koppenberger, and F. Gouyon. Natural language processing of lyrics. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 475–478, 2005.

B. Manolovitz and M. Ogihara. Repeated listens in the music discovery process. *Recommender Systems for Medicine and Music*, pages 119–134, 2021.

C. Martindale and K. Moore. Relationship of musical preference to collative, ecological, and psychophysical variables. *Music Perception*, 6(4):431–445, 1989.

C. Martindale, K. Moore, and J. Borkum. Aesthetic preference: Anomalous findings for berlyne's psychobiological theory. *The American Journal of Psychology*, pages 53–80, 1990.

K. Matrosova, L. Marey, G. Salha-Galvan, T. Louail, O. Bodini, and M. Moussallam. Do recommender systems promote local music? a reproducibility study using music streaming data. *arXiv preprint arXiv:2408.16430*, 2024a.

K. Matrosova, M. Moussallam, T. Louail, and O. Bodini. Depict or discern? fingerprinting musical taste from explicit preferences. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, 7(1):15–29, 2024b.

J. H. McDermott, A. F. Schultz, E. A. Undurraga, and R. A. Godoy. Indifference to dissonance in native amazonians reveals cultural variation in music perception. *Nature*, 535(7613):547–550, 2016.

B. McFee, L. Barrington, and G. Lanckriet. Learning content similarity for music recommendation. *IEEE transactions on audio, speech, and language processing*, 20(8):2207–2218, 2012.

C. McKay and I. Fujinaga. Automatic genre classification using large high-level musical feature sets. In *ISMIR*, volume 2004, pages 525–530. Citeseer, 2004.

R. Mehrotra, J. McInerney, H. Bouchard, M. Lalmas, and F. Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th acm international conference on information and knowledge management*, pages 2243–2251, 2018.

R. Mehrotra, M. Lalmas, D. Kenney, T. Lim-Meng, and G. Hashemian. Jointly leveraging intent and interaction signals to predict user satisfaction with slate recommendations. In *The World Wide Web Conference*, pages 1256–1267, 2019.

A. B. Melchiorre, E. Zangerle, and M. Schedl. Personality bias of music recommendation algorithms. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 533–538, 2020.

A. B. Melchiorre, N. Rekabsaz, E. Parada-Cabaleiro, S. Brandl, O. Lesota, and M. Schedl. Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management*, 58(5):102666, 2021.

C. Mellander, R. Florida, P. J. Rentfrow, and J. Potter. The geography of music preferences. *Journal of Cultural Economics*, 42:593–618, 2018.

M. Millecamp, N. N. Htun, Y. Jin, and K. Verbert. Controlling spotify recommendations: effects of personal characteristics on music recommender user interfaces. In *Proceedings of the 26th Conference on user modeling, adaptation and personalization*, pages 101–109, 2018.

D. Miranda and M. Claes. Personality traits, music preferences and depression

in adolescence. *International journal of adolescence and youth*, 14(3):277–298, 2008.

V. Mishra, K. Liew, E. V. Epure, R. Hennequin, and E. Aramaki. Are metal fans angrier than jazz fans? a genre-wise exploration of the emotional language of music listeners on reddit. In *Proceedings of the 2nd Workshop on NLP for Music and Spoken Audio (NLP4MusA)*, pages 32–36, 2021.

S. Mukhopadhyay, A. Kumar, D. Parashar, and M. Singh. Enhanced music recommendation systems: A comparative study of content-based filtering and k-means clustering approaches. *Revue d'Intelligence Artificielle*, 38(1), 2024.

F. Murtagh and P. Contreras. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, pages 111–125, 2008.

P. H. Nash. *Music regions and regional music*. Deccan Geographical Society, 1968.

P. H. Nash and G. O. Carney. The seven themes of music geography. *Canadian Geographer/Le Géographe canadien*, 40(1):69–74, 1996.

J.-F. Nault, S. Baumann, C. Childress, and C. M. Rawlings. The social positions of taste between and within music genres: From omnivore to snob. *European Journal of Cultural Studies*, 24(3):717–740, 2021.

G. Nave, J. Minxha, D. M. Greenberg, M. Kosinski, D. Stillwell, and J. Rentfrow. Musical preferences predict personality: Evidence from active listening and facebook likes. *Psychological science*, 29(7):1145–1158, 2018.

N. Neophytou, B. Mitra, and C. Stinson. Revisiting popularity and demographic biases in recommender evaluation and effectiveness. In *European Conference on Information Retrieval*, pages 641–654. Springer, 2022.

T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, pages 677–686, 2014.

X. Ning and G. Karypis. Slim: Sparse linear methods for top-n recommender systems. In *Proceedings of the 11th IEEE International Conference on Data Mining*, pages 497–506. IEEE, 2011.

W. Nordström and J. Håkansson. Finding clusters of similar artists-analysis of dbscan and k-means clustering. 2012.

A. North and D. Hargreaves. Subjective complexity, familiarity, and liking for popular music. psychomusicology: A journal of research in music cognition, 14 (1-2), 77–93, 1995.

A. C. North. Individual differences in musical taste. *The American Journal of Psychology*, 123(2):199–208, 2010.

A. C. North and D. J. Hargreaves. Situational influences on reported musical preference. *Psychomusicology: A Journal of Research in Music Cognition*, 15(1-2):30, 1996.

A. C. North and D. J. Hargreaves. Music and adolescent identity. *Music education research*, 1(1):75–92, 1999.

A. C. North, D. J. Hargreaves, and S. A. O'Neill. The importance of music to adolescents. *British journal of educational psychology*, 70(2):255–272, 2000.

J. Oduro-Frimpong. Glocalization trends: The case of hiplife music in contemporary ghana. *International journal of Communication*, 3:22, 2009.

I. Oh and G.-S. Park. The globalization of k-pop: Korea's place in the global music industry. *Korea Observer*, 44(3):389–409, 2013.

F. Pachet, D. Cazaly, et al. A taxonomy of musical genres. In *RIAO*, volume 2, pages 1238–1245. Citeseer, 2000.

F. Pachet, G. Westermann, and D. Laigre. Musical data mining for electronic music distribution. In *Proceedings First International Conference on WEB Delivering of Music. WEDELMUSIC 2001*, pages 101–106. IEEE, 2001.

E. Pariser. *The filter bubble: What the Internet is hiding from you*. penguin UK, 2011.

R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

N. Pavitha, D. Khanwelkar, H. More, N. Soni, J. Rajani, and C. Vaswani. Analysis of clustering algorithms for music recommendation. In *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, pages 1–6. IEEE, 2022.

M. J. Pazzani and D. Billsus. Content-based recommendation systems. In P. Brusilovsky, A. Kobsa, and W. Nejdl, editors, *The Adaptive Web*, pages 325–341. Springer, 2007.

C. S. Pereira, J. Teixeira, P. Figueiredo, J. Xavier, S. L. Castro, and E. Brattico. Music and emotions in the brain: familiarity matters. *PloS one*, 6(11):e27241, 2011.

D. Perrot. Scanning the dial: An exploration of factors in the identification of musical style. *Proc. of ICMPC 1999*, 1999.

R. A. Peterson. Understanding audience segmentation: From elite and mass to omnivore and univore. *Poetics*, 21(4):243–258, 1992.

C. E. Pimentel and E. D. O. P. Donnell. The relation between music preferenceand the big five personality traits. *Psicologia: ciência e profissão*, 28(4): 696–713, 2008.

E. Pistrick. *Performing nostalgia: Migration culture and creativity in south Albania*. Routledge, 2017.

N. A. Privandhani et al. Clustering pop songs based on spotify data using k-means and k-medoids algorithm. *Jurnal Mantik*, 6(2):1542–1550, 2022.

M. Prockup, A. F. Ehmann, F. Gouyon, E. M. Schmidt, O. Celma, and Y. E. Kim. Modeling genre with the music genome project: Comparing human-labeled attributes and audio features. In *ISMIR*, pages 31–37, 2015.

Y. Raimond and M. Sandler. Evaluation of the music ontology framework. In *Extended Semantic Web Conference*, pages 255–269. Springer, 2012.

L. Ramshaw and R. E. Tarjan. On minimum-cost assignments in unbalanced bipartite graphs. *HP Labs, Palo Alto, CA, USA, Tech. Rep. HPL-2012-40R1*, 20:14, 2012.

A. H. Raza and K. 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. IEEE, 2020.

S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 452–461. AUAI Press, 2009.

Y. Renisio, A. Beaumont, J.-S. Beuscart, S. Coavoux, P. Coulangeon, R. Cura, B. Le Bigot, M. Moussallam, C. Roth, and T. Louail. Integrating digital traces into mixed methods designs. 2024.

P. J. Rentfrow and S. D. Gosling. The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, 84(6):1236, 2003.

P. J. Rentfrow, L. R. Goldberg, and D. J. Levitin. The structure of musical preferences: a five-factor model. *Journal of personality and social psychology*, 100(6):1139, 2011.

P. J. Rentfrow, L. R. Goldberg, D. J. Stillwell, M. Kosinski, S. D. Gosling, and D. J. Levitin. The song remains the same: A replication and extension of the music model. *Music perception*, 30(2):161–185, 2012.

F. Ricci, L. Rokach, and B. Shapira. *Introduction to Recommender Systems Handbook*. Springer, 2011.

K. Robinson, D. Brown, and M. Schedl. User insights on diversity in music recommendation lists. In *ISMIR*, pages 446–453, 2020.

J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

S. Saarikallio. *Music as mood regulation in adolescence*. Number 67. University of Jyväskylä, 2007.

S. H. Saarikallio. Music in mood regulation: Initial scale development. *Musicae scientiae*, 12(2):291–309, 2008.

V. N. Salimpoor, D. H. Zald, R. J. Zatorre, A. Dagher, and A. R. McIntosh. Predictions and the brain: how musical sounds become rewarding. *Trends in cognitive sciences*, 19(2):86–91, 2015.

D. Sánchez-Moreno, A. B. G. González, M. D. M. Vicente, V. F. L. Batista, and M. N. M. García. A collaborative filtering method for music recommendation using playing coefficients for artists and users. *Expert Systems with Applications*, 66:234–244, 2016.

B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, pages 285–295. ACM, 2001.

J. B. Schafer, J. A. Konstan, and J. Riedl. Recommender systems in e-commerce. In *Proceedings of the 1st ACM conference on Electronic Commerce*, pages 158–166. ACM, 1999.

T. Schäfer and C. Mehlhorn. Can personality traits predict musical style preferences? a meta-analysis. *Personality and Individual Differences*, 116:265–273, 2017.

T. Schäfer and P. Sedlmeier. From the functions of music to music preference. *Psychology of Music*, 37(3):279–300, 2009.

M. Schedl. *Automatically extracting, analyzing, and visualizing information on music artists from the World Wide Web*. na, 2008.

M. Schedl and D. Hauger. Tailoring music recommendations to users by considering diversity, mainstreaminess, and novelty. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval*, pages 947–950, 2015.

M. Schedl, P. Knees, and G. Widmer. A web-based approach to assessing artist similarity using co-occurrences. In *Proceedings of the Fourth International Workshop on Content-Based Multimedia Indexing (CBMI'05)*. Citeseer, 2005.

M. Schedl, G. Breitschopf, and B. Ionescu. Mobile music genius: Reggae at the beach, metal on a friday night? In *Proceedings of international conference on multimedia retrieval*, pages 507–510, 2014.

M. Schedl, P. Knees, B. McFee, D. Bogdanov, and M. Kaminskas. Music recommender systems. In *Recommender Systems Handbook*, pages 453–492. Springer, 2015.

M. Schedl, H. Zamani, C.-W. Chen, Y. Deldjoo, and M. Elahi. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7:95–116, 2018.

A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and

metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260, 2002.

B. Sguerra, V.-A. Tran, and R. Hennequin. Discovery dynamics: Leveraging repeated exposure for user and music characterization. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 556–561, 2022.

D. Shakespeare, L. Porcaro, E. Gómez, and C. Castillo. Exploring artist gender bias in music recommendation. *arXiv preprint arXiv:2009.01715*, 2020.

E. Shakirova. Collaborative filtering for music recommender system. In *2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, pages 548–550. IEEE, 2017.

Y. Shavitt and U. Weinsberg. Song clustering using peer-to-peer co-occurrences. In *2009 11th IEEE International Symposium on Multimedia*, pages 471–476. IEEE, 2009.

M. Slaney and W. White. Similarity based on rating data. In *ISMIR*, pages 479–484, 2007.

J. A. Sloboda. Everyday uses of music listening: A preliminary study. *Music, mind and science*, pages 354–369, 1999.

S. L. Smith, M. Choueiti, K. Pieper, and H. Clark. Inclusion in the recording studio? gender and race/ethnicity of artists, songwriters & producers across 900 popular songs from 2012-2020, 2021. URL https://assets.uscannenberg.org/docs/inclusion-in-the-recording-studio.pdf.

M. Soleymani, A. Aljanaki, F. Wiering, and R. C. Veltkamp. Content-based music recommendation using underlying music preference structure. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2015.

M. Sordo and X. Serra. A survey of metadata and metadata standards for musical information retrieval. *Journal of New Music Research*, 43(2):184–203, 2014.

W. B. Swann Jr and P. J. Rentfrow. Blirtatiousness: Cognitive, behavioral, and physiological consequences of rapid responding. *Journal of Personality and Social Psychology*, 81(6):1160, 2001.

L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

P. Tagg. Analysing popular music: theory, method and practice. *Popular music*, 2:37–67, 1982.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

C. Troncoso, M. Isaakidis, G. Danezis, and H. Halpin. Knock knock, who's

there? membership inference on aggregate location data. In *NDSS*. The Network and Distributed System Security Symposium (NDSS), 2020.

D. R. Turnbull, S. McQuillan, V. Crabtree, J. Hunter, and S. Zhang. Exploring popularity bias in music recommendation models and commercial steaming services. *arXiv preprint arXiv:2208.09517*, 2022.

G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.

A. van den Oord, S. Dieleman, and B. Schrauwen. Deep content-based music recommendation. In *Advances in neural information processing systems*, pages 2643–2651, 2013.

Q. Villermet, J. Poiroux, M. Moussallam, T. Louail, and C. Roth. Follow the guides: disentangling human and algorithmic curation in online music consumption. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 380–389, 2021.

S. Volokhin and E. Agichtein. Understanding music listening intents during daily activities with implications for contextual music recommendation. In *Proceedings of the 2018 conference on human information interaction & retrieval*, pages 313–316, 2018.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17: 395–416, 2007.

H. Wallace. Clustering of musical genres. 2015.

X. Wang, D. Rosenblum, and Y. Wang. Context-aware mobile music recommendation for daily activities. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 99–108, 2012.

Y. Wang, W. Ma, M. Zhang, Y. Liu, and S. Ma. A survey on the fairness of recommender systems. *ACM Trans. Inf. Syst.*, 41(3), feb 2023. ISSN 1046-8188. doi: 10.1145/3547333. URL `https://doi.org/10.1145/3547333`.

M. K. Ward, J. K. Goodman, and J. R. Irwin. The same old song: The power of familiarity in music choice. *Marketing Letters*, 25:1–11, 2014.

A. B. Warriner, V. Kuperman, and M. Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45: 1191–1207, 2013.

J. Wasilewski and N. Hurley. Incorporating diversity in a learning to rank recommender system. In *The twenty-ninth international flairs conference*, 2016.

S. F. Way, S. Gil, I. Anderson, and A. Clauset. Environmental changes and the dynamics of musical identity. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 527–536, 2019.

R. Yves, A. Samer, S. Mark, and G. Frederick. The music ontology. In *Proceedings of the International Conference on Music Information Retrieval, ISMIR*, volume 2007, pages 417–422, 2007.

M. Zadel and I. Fujinaga. Web services for music information retrieval. In *ISMIR*, 2004.

E. Zangerle, M. Pichl, and M. Schedl. Culture-aware music recommendation. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 357–358, 2018.

E. Zangerle, M. Vötter, R. Huber, and Y.-H. Yang. Hit song prediction: Leveraging low-and high-level audio features. In *ISMIR*, pages 319–326, 2019.

E. Zangerle, M. Pichl, and M. Schedl. User models for culture-aware music recommendation: Fusing acoustic and cultural cues. *Trans. Int. Soc. Music. Inf. Retr.*, 3(1):1–16, 2020.

F. Zhang. Research on music classification technology based on deep learning. *Security and Communication Networks*, 2021:1–8, 2021.

S. Zhang, L. Yao, A. Sun, and Y. Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52 (1):1–38, 2019.

L. Zheng, V. Noroozi, and P. S. Yu. Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 425–434. ACM, 2017.

Y. Zheng, C. Gao, L. Chen, D. Jin, and Y. Li. Dgcn: Diversified recommendation with graph convolutional networks. In *Proceedings of the Web Conference 2021*, pages 401–412, 2021a.

Y. Zheng, C. Gao, X. Li, X. He, Y. Li, and D. Jin. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the Web Conference 2021*, pages 2980–2991, 2021b.

Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan. Large-scale parallel collaborative filtering for the netflix prize. In *Proceedings of the 4th international conference on Algorithmic Aspects in Information and Management*, pages 337–348. Springer, 2008.

M. Zhu, X. Yang, J. Xiao, X. Guo, B. Liu, X. Hu, and A. Zhou. A survey on cross-domain recommendation: Taxonomies, methods, and future directions. *arXiv preprint arXiv:2108.03357*, 2021.

R. L. Zweigenhaft. A do re mi encore: A closer look at the personality correlates of music preferences. *Journal of individual differences*, 29(1):45–55, 2008.

W. R. Zwick and W. F. Velicer. Comparison of five rules for determining the number of components to retain. *Psychological bulletin*, 99(3):432, 1986.