Music Recommender Systems:

Taking Into Account The Artists' Perspective

Andrés Ferraro

TESI DOCTORAL UPF / ANY 2021

THESIS SUPERVISOR
Dr. Xavier Serra Casals
Dept. of Information and Communication Technologies

Dissertation submitted to the Department of Information and Communication Technologies of Universitat Pompeu Fabra in partial fulfillment of the requirements for the degree of

DOCTOR PER LA UNIVERSITAT POMPEU FABRA

Music Technology Group (http://mtg.upf.edu), Department of Information and Communication Technologies (http://www.upf.edu/dtic), Universitat Pompeu Fabra (http://www.upf.edu), Barcelona, Spain.

The doctoral defense was held on ..... at the Universitat Pompeu Fabra and scored as .....

Dr. Xavier Serra Casals
(Thesis Supervisor)
Universitat Pompeu Fabra (UPF), Barcelona, Spain

Dr. Emilia Gómez
(Thesis Committee Member)
Universitat Pompeu Fabra (UPF), Barcelona, Spain

Dr. Cynthia Liem
(Thesis Committee Member)
Delft University of Technology, Delft, Netherlands

Dr. Òscar Celma
(Thesis Committee Member)
Spotify, New York, United States

This thesis has been carried out at the Music Technology Group (MTG) of Universitat Pompeu Fabra in Barcelona, Spain, from Oct. 2018 to Sep. 2021. It is supervised by Dr. Xavier Serra Casals. Work in Chapter 3 and 4 has been conducted in collaboration with Dr. Christine Bauer (Utrecht University, The Netherlands). Work in Chapter 5 has been conducted in close collaboration with Dr. Massimo Quadrana (Pandora-SiriusXM, USA) and Dr. Sergio Oramas (Pandora-SiriusXM, USA). Work in Chapter 6 has been conducted in close collaboration with Dr. Dietmar Jannach (University of Klagenfurt, Austria).

Work in Chapter 7 has been conducted in collaboration with Dr. Dmitry Bogdanov and also carried out in collaboration with the Kakao Corp. (Korea). A detailed list of collaborators include Y Kim, S Lee, B Kim, N Jo, S Lim, S Lim, J Jang, S Kim, XS Jay, H Jeon and J Yoon.

Work in Chapter 8 has been conducted in close collaboration with Dr. Xavier Favory (Universitat Pompeu Fabra, Spain), Dr. Kostas Drossos (Tampere University, Finland) and Dr. Dmitry Bogdanov (Universitat Pompeu Fabra, Spain).

Our work has been partially supported by the Kakao Corp. (Korea).

Acknowledgments

This work would not have been possible without the support of my supervisor Prof Xavier Serra, who was a great inspiration for this thesis. He showed me the importance of exploring and understanding the diverse music styles and traditions of the world.

I thank Dmitry Bogdanov for the guidance in part of my work and for always been there when I needed his help. I would also like to thank Christine Bauer for helping me during my PhD; for her patience and also for believing and sharing similar ideas.

I would like to thank all the music artists I interviewed for this thesis. These interviews are a fundamental part of the thesis. Having the opportunity to speak with the artists and know more about their context and worries was a very enriching experience for me.

Special thanks to all the researchers I collaborated with during the PhD including Diemar Jannach, Eduardo Fonseca, Konstantinos Drossos, Lorenzo Porcaro, Magdalena Fuentes, Massimo Quadrana, Miguel García, Minz Won, Sara Latifi, Sergio Oramas, Thomas Nuttall and Xavier Favory for their help during these years and also for the enriching and valuable discussions.

I thank Cristina Garrido, Sonia Espí, Lydia García for all the help during these years in administrative tasks.

I also want to acknowledge all the members of the MTG that were with me at different moments during the last 6 years for their important feedback, the useful discussions and also for making this journey more pleasant. Thanks to Alastair Porter, Albin Correya, Ángel Faraldo, António Ramires, Benno Weck, Ernic Guiso, Frederic Font, Furkan Yesiler, Jorge Marcos, Luis Joglar, Jordi Pons, Juan Gómez, Marius Miron, Miguel Perez, Olga Slizovskaia, Pablo Alonso, Pablo Zinemanas, Philip Tovstogan and Xavier Lizarraga. Thanks also to current and former members of Compmusic: Ajay, Alia, Georgi, Gopala, Jyoti, Rafael, Rong, Sankalp, Sertan, Swapnil and Vsevold.

I would like to thank my parents and siblings for believing in me and giving their support from a long distance. Last but not least, special thanks to Elisa for all the help during this time and for always giving me her human-centered point of view.

Abstract

Music streaming platforms nowadays play an important role in music consumption and have a big influence on the musical taste of the listeners. Machine learning-based recommender systems are a fundamental part of such streaming platforms defining what music people listen to and when. As for many applications of machine learning, there is an increasing debate in academia, industry and governments about the effects that recommender systems have in society and the ethical implications of such systems.

Bias in music recommender systems towards more popular items has been studied extensively in the past. This bias affects both artists and listeners since it reduces the possibility of a large catalog's portion of getting any exposure. Recently, in the recommender systems community, it was raised the importance of considering the multiple stakeholders of a system when generating the recommendations. However, most of the research in the music domain has taken into account the users' perspective only. This thesis goes beyond the problem of popularity bias, it tries to uncover other dimensions in which the music recommender systems can affect the artists and propose alternatives to mitigate such problems.

The contributions of this thesis are (i) identification of multiple aspects in which the current platforms and their recommender systems affect the music artists and concrete ways in which they could be more beneficial in the future, (ii) analysis of the algorithmic effect regarding gender imbalance in the recommendations and mitigation of such problem based on the output of artists' interview, (iii) analysis of the longitudinal effect of multiple state-of-the-art algorithms for session-based recommendations in users behavior negatively affecting the artists, (iv) publication of the first large-scale open dataset that contains audio and playlist information, (iv) novel contrastive learning approach proposed to combine multiple modalities (audio, genre and playlist information) beneficial for multiple tasks such as music recommendation, genre classification and automatic-tagging.

It is necessary to improve recommender systems through multidisciplinary research. Contributions like the ones presented in this thesis allow us to move a step forward in that direction, making streaming platforms more beneficial for both the artists and users.

Resum

Actualment les plataformes que ofereixen serveis de música en línia juguen un paper important en el consum de música i tenen una gran influència en les preferències musicals de les persones. Els sistemes de recomanació basats en aprenentatge automàtic són una part fonamental de les plataformes de música en línia, definint la música que escolten les persones en cada moment i lloc. A l'igual que en altres aplicacions d'aprenentatge automàtic, cada vegada és més discutit tant a nivell acadèmic, industrial i governamental els efectes que els sistemes de recomanació poden tenir en la societat i les implicacions ètiques d'aquests sistemes.

Els biaixos dels sistemes de recomanació musical cap als elements més populars han estat estudiats extensivament. Aquest biaix afecta tant a artistes com a usuaris ja que redueix la possibilitat d'aconseguir una mínima exposició a una gran proporció del catàleg musical. Recentment, a l'àrea de sistemes de recomanació, s'ha reconegut la importància de considerar els interessos de tots els grups de persones involucrats quan es generen recomanacions. No obstant això, la majoria de la recerca relacionada amb sistemes de recomanació en el domini de la música s'ha enfocat només en la perspectiva dels usuaris. Aquesta tesi no es limita als problemes de biaix de popularitat sinó que també intenta descobrir altres dimensions en què els sistemes de recomanació afecten als artistes musicals i proposa solucions per mitigar aquests problemes.

Les contribucions d'aquesta tesi són: (I) la identificació de múltiples aspectes en què les plataformes musicals i els seus sistemes de recomanació afecten els artistes i de quina manera podrien ser més beneficiosos en el futur; (II) l'anàlisi dels efectes d'algorismes de recomanació pel que fa a el balanç de gènere i una possible forma de mitigar aquests efectes basat en l'opinió dels artistes; (III) l'anàlisi de la influència a llarg termini en els usuaris generada per diversos algoritmes de recomanació basats en sessions que afecta negativament els artistes; (IV) definició d'un nou mètode basat en *contrastive learning* per combinar múltiples modalitats (àudio, gènere musical i informació de *playlists*) que aconsegueix millorar els resultats de diferents tasques com la recomnanació musical, la classificació de gènere musical i l'anotació automàtica de música.

Per millorar els sistemes de recomanació és important realitzar més recerca multidisciplinària. Contribucions com les presentades en aquesta tesi permeten moure'ns en aquesta direcció, fent possible que les plataformes de musica online siguin més beneficioses per als artistes i els usuaris.

Resumen

Actualmente las plataformas que ofrecen servicios de música online juegan un rol importante en el consumo de música y tienen una gran influencia en el gusto musical de las personas. Los sistemas de recomendación basados en aprendizaje automático son una parte fundamental de las plataformas de música online, definiendo la música que escuchan las personas en cada momento y lugar. Al igual que en otras aplicación de aprendizaje automático, cada vez es más discutido tanto a nivel académico, industrial y gubernamental los efectos que los sistemas de recomendación pueden tener en la sociedad y las implicancias éticas de dichos sistemas.

Los sesgos de los sistemas de recomendación musical hacia los elementos más populares ha sido estudiado extensivamente. Este sesgo afecta tanto a artistas como usuarios ya que reduce la posibilidad de lograr una mínima exposición a una gran proporción del catálogo musical. Recientemente, en el área de sistemas de recomendación, se ha reconocido la importancia de considerar los intereses de todos los grupos de personas involucrados en el sistema cuando se generan recomendaciones. Sin embargo, la mayoría de la investigación relacionada a sistemas de recomendación en el dominio de la música se ha enfocado solamente en la perspectiva de los usuarios. Esta tesis no se limita al problema del sesgo de popularidad sino que intenta descubrir otras dimensiones en las que los sistema de recomendación afectan a los artistas musicales y propone soluciones para mitigar dichos problemas.

Esta tesis realiza las siguientes contribuciones: (i) identificación de múltiples aspectos en los que las plataformas musicales y sus sistemas de recomendación afectan a los artistas y de qué manera podrían ser más beneficiosos en el futuro; (ii) análisis de los efectos de algoritmos de recomendación con respecto al balance de género e investigación de una posible forma de mitigar dichos efectos basado en la opinión de los artistas; (iii) análisis de la influencia a largo plazo en los usuarios generada por varios algoritmos de recomendación basados en sesiones afectando negativamente a los artistas; (iv) definición de un nuevo método basado en *contrastive learning* para combinar múltiples modalidades (audio, género y *playlists*) logrando un mejor desempeño en diferentes tareas como recomendación musical, clasificación de género y anotación de música.

Para mejorar los sistemas de recomendación es importante realizar más investigación multidisciplinar. Contribuciones como las presentadas en esta tesis permiten movernos en dicha dirección, haciendo posible que las plataformas de música online sean más beneficiosas para los artistas y los usuarios.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Streaming platforms are established as the most popular choice for listening to music (IFPI, 2019), with billions of users worldwide. The streaming platforms offer catalogs of music spanning tens of millions of songs and are increased every day with thousands of new songs. It is not possible for the listeners to know all the songs in the catalogs, thus, recommender systems play an important role inside such platforms to help listeners to decide what to choose. Therefore, these recommender systems have an increasing influence on what people consume and what gets more exposure, defining which song or artist gets promotion and which does not, in a way becoming the gatekeepers of the content. Given such an important responsibility, we have to carefully design the recommender systems considering the interest of the different groups of people that are affected by them.

Machine learning algorithms are commonly used in recommender systems. In multiple domains, machine learning algorithms show problems of unwanted discrimination and bias, gaining the attention of the media and raising concern from the general society because of the negative impact that the algo rithms could have on a large scale (Crawford, 2016; O'neil, 2016; Russell, 2019; Bostrom, 2017). In some cases, machine learning algorithms make visible some unwanted biases that are present in our society, since the algorithms are trained based on human decisions. This shows that even if the use of technology in a way brought up some issues, these issues can not be solved only by applying techniques from the field of computer science (Crawford & Calo, 2016). During the last years, there has been an increasing interest in this topic from the research community involving experts from multiple disciplines such as social science, philosophy, computer science and policy-oriented research.

In the research field of recommender systems, there was an increasing interest from the community to identify the effects that the algorithms could have and propose solutions. It was introduced the notion that recommender systems have multiple stakeholders who are affected by the system and have different

and sometimes contradictory goals or intentions. Making recommendations that consider the interests of the multiple stakeholders of a system started to gain more attention in the research community in the last years (Abdollahpouri et al., 2020a). However, most research on this topic is still focused on how the recommender systems perform from the consumers' perspective. There is less literature about how recommender systems affect the content producers or providers. Moreover, there is a research gap in considering the interests of the society at large when assessing the ethical impact of recommender systems (Milano et al., 2020).

In the field of music recommendation, the problem of popularity bias has been studied for many years (Celma, 2009). This bias affects both artists and listeners since it reduces the possibility of a large catalog's portion of getting any exposure (Holzapfel et al., 2018). Yet, not many works have studied music recommender systems from the artists' perspective, including research in the topic of popularity bias. There is a lack of understanding in general of how these systems affect the artists, which is key to define how the systems should work considering also the ethical implications (Born et al., 2021). For example, gender imbalance is a problem in the music industry that gained more attention in the last years (Smith et al., 2018; Aguiar et al., 2018). The music recommender systems could be increasing the imbalance if they reproduce the bias that is present in the users' consumption. However, the system could be used as a way to mitigate such a problem, giving more visibility to female artists and gender minorities. Therefore, this is a clear example where technology can be applied either positively or negatively and must be aligned with the goals of society.

This thesis focuses on understanding different ways in which music recommender systems can affect artists. It proposes alternatives to mitigate such problems and to make better recommendations with a special focus on the artists' perspective but also considering the interests of the users. Thus, creating fairer systems.

1.1 Computer Ethics

Understanding the ethical implications of an algorithm is a topic that gained more attention in the last years. However, the first ideas of what later will be called *computer ethics* were introduced by Wiener (2019) in 1948. It was in 1973 when the Association of Computer Machinery (ACM) published the first version of the code of ethics (ACM, 1972). Later, Maner (1980) introduced the term *computer ethics* and defined it as a branch of applied ethics. The essay from Moor (1985) that defines what it is *computer ethics* gained a lot of attention at the time. The same year, the book from Johnson (1985) was

published and was established as a reference to study this topic. Unlike the previous works, Johnson argues that computers did not create new ethical problems but instead are new versions of already familiar issues. However, Johnson agrees with Moor that should be considered as part of the field of applied ethics.

Applied ethics is a branch of philosophy that studies the application of moral principles or ethical theories to solve moral problems that arise in practical fields. There is still an ongoing debate about the uniqueness of the ethical issues raised by computers (i.e. *uniqueness debate*) that was started by Johnson and Moore (Tavani, 2002). The importance of this debate is that it argues whether we should apply the fundamental ethical theories to answer the ethical questions that arise in computer ethics.

In the last years, with the increase of computational power, more access to larger amounts of information and advances in machine learning techniques, many concerns have been raised by researchers and also in the media about the potential harms that can be produced by applications of machine learning (Russell, 2019; Bostrom, 2017; O'neil, 2016). This gave more attention to the topic and more researchers from the fields of computer science, mathematics and statistics started to focus on ways to define how fair models should behave (Corbett-Davies et al., 2017). However, in the attempt to formalize the idea of a fair system, multiple definitions had been proposed in the last years (Narayanan, 2018), these definitions seem to be detached from the discussions in the *uniqueness debate*.

According to Binns (2018) some of these definitions of fairness can also be related to the principles of justice, non-discrimination and egalitarianism. Binns suggests that in the machine learning community the term *fairness* is a placeholder for a variety of normative egalitarian considerations.

The multiple definitions of fairness sometimes can be contradictory and finding a trade-off between them is not just a technical decision. Thus, fairness definitions can not be applied in a general way and we need to consider the context in which the algorithm is used and the people affected, involving researchers from different fields such as computer science, philosophy, ethics, human-computer interaction and law.

1.2 Communication of Machine Learning Research and Operationalization of Fairness

The machine learning research community is debating if the dissemination of scientific advances should be restricted because of the negative impact that it may have in our societies (Hutson, 2021). This questions the established

procedure followed in the computer science research community through peer-reviewed conferences and journals to disseminate the advances. For example, recently the Conference on Neural Information Processing Systems (neurISP) adopted a procedure to assess if contributions could be rejected because of ethical considerations (Hsuan-Tien Lin, Maria Florina Balcan & Ranzato, 2020).

Johnson & Verdicchio (2017) mentions potential limitations that can be generated in the research of artificial intelligence because of public understanding. They highlight the importance of presenting and communicating correctly the ethical issues in AI. In particular, the importance of distinguishing the autonomy that AI has from the human actors in the design and deployment of the systems. Therefore, we should not focus on the potential negative consequences of an algorithm but we should focus more on how these algorithms are used in practice. This is related to the difficulties that had been raised in operationalizing the bias restrictions of ML algorithms (Cramer et al., 2018, 2019a).

1.3 Social Implication and Ethical Issues in Music Recommendations

Since streaming platforms have a large number of users, their recommender systems may have implications for society. These implications are not necessarily negative but it is important to take them into account since they affect the way people consume music.

Some aspects related to the implications of music streaming platforms and their algorithms that were covered by the media in the past years include:

- Monitoring users: There is an increasing concern about streaming platforms using sensors to monitor people's behaviour (Minsker, 2021; Graves, 2021). Additionally, multiple research works by North & Hargreaves (2007a,b,c) associated users' listening behavior with their personalty, potentially revealing private information of the users.

- Remuneration: Multiple media articles cover the business models of the different streaming platforms and how unfair they are for music artists (Mulligan, 2021; Shah, 2020; Hogan, 2020; McDermott, 2020)

- Discrimination and biases: Chayka (2019) mentions multiple cases where users identify a predominant presence of male artists recommended by the algorithms of a streaming platform. In another case, the singer Rosalía denounces in an interview the discrimination that women artists suffer in the music industry which is related to algorithms (Serrano, 2020).

- Shaping music consumption: Multiple media articles describe how music streaming platforms shape and define what gets more promotion. Turk (2021) describes how the *feedback loop* of streaming platforms does not allow users to discover different music styles. Kirn (2019) mentions that artists publicly sharing the statistics about user listening activity provided every year by the platform promotes the consumption of music as a commodity. Pelly (2018) describes how streaming platforms had changed the way people consume music due to the popularity of playlists. Arcand (2017) argues that Youtube's algorithm is already shaping the music industry. Maicki (2020) presents a recent functionality from Spotify that allows artists to promote more their music through recommendations in exchange for lower royalty payments.

There are multiple decisions that the designers of such recommender systems make that would affect different groups of people related to the music platform. Recently, it was introduced the notion of multi-stakeholders recommender systems, with the goal of considering the interest of the different groups when generating the recommendations. Important decisions have to be made when designing a system that balances the interests of different groups of stakeholders. However, there is a lack of research in understanding how the artists get affected by music recommender systems.

The use of audio-based music description in the field of Music Information Retrieval (MIR) has been developed for many years and solutions for recommending new and less popular music had been proposed. Further research in this direction is needed for a fairer exploration of music content and also more aligned with the artists' interests. A big limitation is the availability of open datasets that can be used to combine content and collaborative information and can be used publicly by researchers to compare new methods.

It is also important to mention that even if laws are currently starting to regulate the use artificial intelligence, usually the music domain is considered as a minimal risk, which does not require legal obligations of auditing and adhere to voluntary codes of conduct from the companies (European Commission, 2021).

1.4 Inter-disciplinary Research

It is important to highlight the need to study the problem raised in *computer ethics* not only by researchers from different disciplines working together (i.e. multi-disciplinary) but also by combining the tools and methods of the different disciplines (i.e. inter-disciplinary) (Brey, 2000). An important effort

was dedicated in this thesis to learn and apply techniques and methods from different disciplines to find optimal solutions.

With the goal of understanding the perspective of people affected by the recommendations Qualitative Content Analysis (Mayring, 2004) was applied. This method is commonly used in the field of Human-Computer Interaction to understand the users' opinion of a particular system. The method was used in this thesis to understand artists' opinions about streaming platforms and their recommender systems.

Other aspects that needed to be considered in this thesis were the advantages and disadvantages of the different algorithms currently used to make recommendations, including the techniques used to evaluate such systems studied in the recommender systems community for many years. In this thesis, we also had to consider how to evaluate qualitatively the systems considering multi-stakeholders.

Also, the automatic extraction of information from the music content is a fundamental part of the systems that should be considered. Researchers in the MIR community had developed techniques in this direction for many years. In this thesis, a significant effort was dedicated in reviewing and exploring further such techniques. A large-scale dataset was released and a new method that can be used for multiple task was developed.

1.5 Objectives of the Thesis

Since this thesis focuses on some ethical aspects of music recommendations, one of its goals is to contribute to society, by raising questions that are important to discuss not only from a technical point of view but also need to be discussed by society in general.

The first goal of this thesis is to understand the artists' perspective regarding music streaming platforms and the role that automatic recommendations play in those platforms.

After identifying some of the issues that affect the artists in the streaming services, the goal is to address some of them by understanding the effects of different algorithms used for music recommendation and proposing concrete solutions to mitigate the negative effects. In summary, in this thesis we:

- Study how music recommendation algorithms behave regarding different groups of artists. Analyze the effect that the algorithms could have from the artists' perspective in the long term and propose actions to reduce unfair behaviors.

- Leverage multiple signals of user interaction in order to produce recommendations that maximize the engagement that users have with artists

- Study the longitudinal effect of multiple algorithms for session-based recommendations. These algorithms rely on few interactions of the user in a listening session in order to produce recommendations, potentially having a negative effect in terms of coverage of the catalog, recommending more the popular items.

One of the main limitations for conducting research in the field of music recommendations is the lack of open datasets that combine audio information with user interaction data due to commercial licensing of the content. One of the goals of this thesis is the creation of an open dataset of commercial music, avoiding licensing issues. To do avoid copyright issues, we released representations of the audio (such as mel-spectrograms) with enough information suitable for content-based methods while, at the same time, the quality in the reconstruction of the original audio is severely affected. For that we reduced the information available in the mel-spectrograms of the audio and we studied how it affects the performance on the tasks where this dataset can be used.

Finally, recommending new music without sufficient user interaction data (i.e., suffering a cold-start problem) is an important topic that was addressed in the past by automatically learning a representation of the music using multiple types of information in a multimodal approach (e.g: from the audio of the songs, text extracted from the web like social networks or Wikipedia, crowd-sourced tags related to songs or other metadata) (Oramas et al., 2017b; Van den Oord et al., 2013). Another goal of this thesis is the exploration of a new method that takes advantage of multi-modal music data, such as semantic metadata and collaborative filtering information. An improved representation of the information related to music should give a better performance in multiple tasks of MIR (such as automatic playlist generation, genre classification and automatic-tagging from audio) compared to other state-of-the-art approaches.

1.6 General Structure of the Thesis

Since this thesis covers topics from different fields each chapter is self-contained, focusing on a particular problem, describing the methodology followed, giving a description of the work carried out, details of the results and finally some conclusions. However, the different chapters are connected, the findings presented in one chapter are important for developing the other chapters. The chosen order of the chapters follows the logical connection between the results of the previous chapters as described in Figure 1.1.

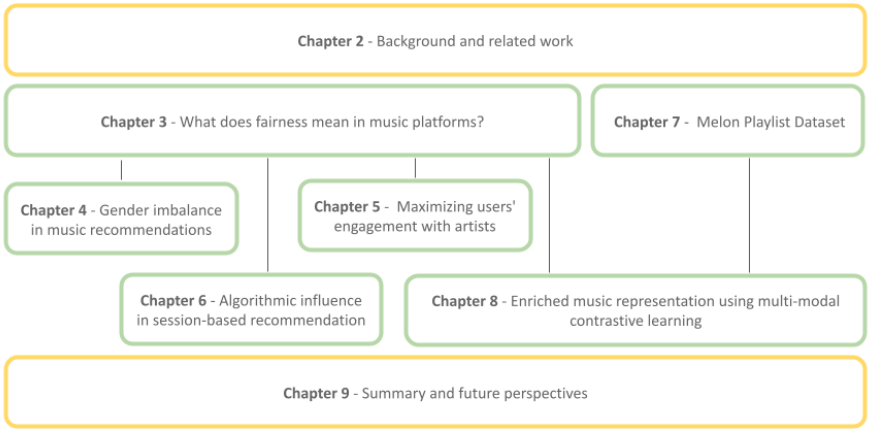

Figure 1.1: Diagram with structure of the thesis.

Chapter 2 presents the previous work on which this thesis is based, including different concepts and methods from the relevant areas covered by this thesis.

Chapter 3 covers a missing gap in the research literature regarding the effects of streaming platforms and algorithmic recommendations from the point of view of the artists, which is obtained from interviews with some music artists. One of the aspects identified from the results presented in **Chapter 3** is the gender imbalance, this is studied more in depth in **Chapter 4** following a qualitative and also quantitative approach. More specifically, we study and propose a way to mitigate the problem of gender imbalance in the recommendations generated with a collaborative filtering algorithm.

Chapter 5 describes a method to make recommendations that give more value to the users and the artists by measuring the engagement that a user will have with the recommended artist. This method considers a) the number of days a user listened to an artist b) the number of songs of the artist listened to by the user and c) the number of times a day that the user listened to the artist.

Chapter 6 focuses on session-based algorithms, simulating the effects that these algorithms can have in the long term on what users listen to. From the different families of algorithms, it is possible to see a reduction in the coverage and also an increase in the recommendation of already more popular items. Based on simulations, a method is proposed that mitigates the negative impact of the algorithms.

In **Chapter 7** we focus on the problem of limited public datasets available that can be used by researchers for recommending music in a cold-start situation. To this end, we collaborated with a Korean music streaming service

and published a dataset that contains playlists information. It includes multiple types of data such as audio representations, track-playlists interactions and genre annotations. Using this dataset, in **Chapter 8** we propose a new method for learning enriched music representations from multiple modalities of information which can be used for multiple tasks such as playlists continuation, automatic tagging and audio classification.

Finally, **Chapter 9** summarizes the findings of the thesis and also and gives the final conclusions. This chapter describes the main contributions of the thesis including the datasets, software and publications. Also, in this chapter is discussed the dissemination of the thesis outside the research community and media coverage. Finally, in this chapter possible future work is discussed.

CHAPTER 2

Background

This chapter describes the previous works on which this thesis is based. Multiple topics are covered from the fields of music information retrieval, recommender systems, human-computer interaction and fairness in machine learning.

We first cover some methods for generating music recommendations and the characteristics of such methods, including the public datasets that are commonly used. Then, we describe the metrics used for the evaluation of recommendations and some issues in offline evaluation.

After that, we introduce the topic of algorithmic fairness and multi-stakeholder recommendations. We also cover the qualitative studies regarding the perception of algorithmic fairness. Finally, we describe studies related to systems, gender and discrimination, including an overview of gender bias in the music domain.

2.1 Methods for Recommendation

The methods to generate recommendations are usually divided into Collaborative Filtering (CF) and Content-based (CB) (Ricci et al., 2010). The CF methods are based on the interactions between the items and the users to generate the recommendations. The CB methods use the information of the items to generate the recommendations (e.g., audio) or information related to the context of the items (e.g., reviews, tags, lyrics or biographies), some authors consider Content and Context-Based as a different category.

The interactions between the users and items that are used in the CF methods could be implicit or explicit (Ricci et al., 2010). An example of explicit feedback is when the users likes or dislikes a songs. An example of implicit feedback is when the information of play-counts is used. In this thesis we focus on implicit feedback since it is the most common to find in publicly available music datasets.

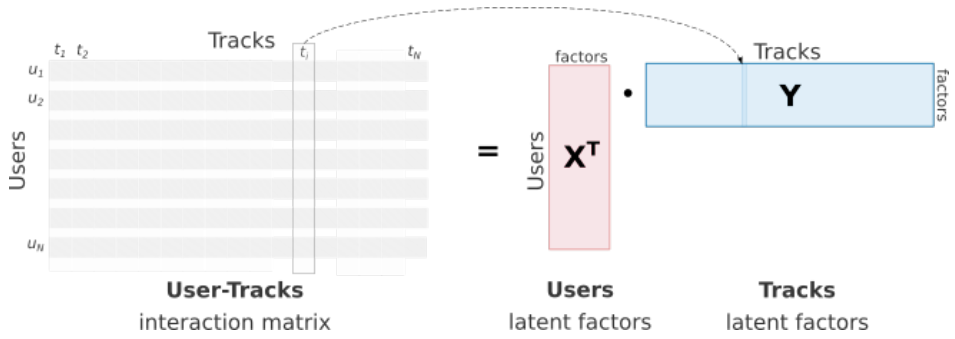

Figure 2.1: Matrix Factorization of user-tracks iteractions.

The term collaborative filtering was coined by the Tapestry project at Xerox PARC (Goldberg et al., 1992). Since then multiple types of CF methods had been proposed, some are based on finding similarity neighborhoods among the items or users, others are based on dimensionality reduction to find a representation of the items and the users (Ricci et al., 2010).

There are multiple methods of CF based on dimensionality reduction, among the most used are based on Singular Value Decomposition and Non-negative Matrix Factorization. The method proposed by Hu et al. (2008) is commonly used for implicit feedback. Figure 2.1 shows how Matrix Factorization works, the original interactions matrix is decomposed into two matrices, the *User latent factors* and the *Item latent factors*. If we multiply the user and item latent factors we regenerate the original interactions in the matrix, predicting the missing values.

Model-based methods such as Matrix Factorization described in Figure 2.1, require a loss function and a optimization method to train the model. Hu et al. (2008) proposed Alternating Least Squares (ALS), an iterative optimization method that learns user and item latent vectors. For one iteration it keeps constant the user factors and trains the item factors, then in the following iteration keeps constant the item factors and learns the user factors. The loss function to minimize is:

$$min_{x*,y*} \sum_{u,i} c_{ui}(p_{ui} - x_u^T y_i)^2 + \lambda(\sum_u ||x||^2 + \sum_i ||y||^2)$$

Where x_u and y_i are the user latent vector and item latent vector respectively. Additionally, this loss function allows to indicate for a given interaction the confidence metric (c_{ui}) and a preference metric (p_{ui}). Finally, λ is used for regularization of the model.

Bayesian Personalized Ranking (BPR) is another popular method proposed by Rendle et al. (2009), the authors stress the importance of optimizing con-

sidering pairs of items in BPR instead of single items like in ALS. The authors argue that BPR is more aligned with the goal of ranking items for a user instead of predicting the original interactions.

More recent methods for recommendations based on deep learning claim that can achieve higher performance than traditional methods (e.g., Zheng et al. (2018); Ebesu et al. (2018); He et al. (2018); Zhang et al. (2018)). However, Dacrema et al. (2021) claims that not much improvement has been achieved in the last years for top-n recommendations with the introduction of deep learning approaches. Dacrema et al. (2021) reproduce 12 algorithms based on deep learning and shows that they do not perform better than many baselines, highlighting methodological issues in the evaluations of the methods and problems with reproducibility.

2.1.1 Session-based Recommender Systems

Session-based recommender systems emerged as a particular setting where there is not long-term information of the user, it is only known the last items consumed in the ongoing session and the next element to be consumed have to be predicted. There are multiple real-world applications of this setting such as news recommendations, e-commerce and also music recommendation. One of the advantages of these methods is than can adapt well to the interest context of the user which is something that is important in the music domain.

Methods for session-based recommendations are typically based on previous method for recommendations such as nearest-neighbor, Matrix Factorization and also deep learning approaches (Wang et al., 2019b).

- Nearest-neighbor-based approaches are simple but effective methods for session-based recommendations (Ludewig et al., 2019). Session-based kNN (Ludewig & Jannach, 2018) (sknn) compares the items in the current session with the items in the previous sessions to determine the most similar session in the training data.

- Multiple deep learning approaches had been proposed for session-based recommendations. Recurrent Neural Networks (RNN) are commonly used since allow to model sequences or items in the sessions. GRU4REC (Hidasi et al., 2016) is based on RNNs with Gated Recurrent Units (GRU) with the goal of predicting the probability of next events given an input sequence. It is the first widely-used neural approach designed for session-based recommendations. Other methods based on deep learning incorporate attention mechanism to better capture the level of relevance of different items in the sessions. NARM (Li et al., 2017) combines RNNs

with an attention mechanism and showed competitive results in multiple datasets (Ludewig et al., 2019).

- Factorization methods are adapted to predict the next action of a user by combining with different methods such as Markov Chains (FPMC by Rendle et al. (2010)) or FPMC-LR by Cheng et al. (2013).
- Other baseline methods such as rule-based had been proposed in the past proving to be effective. For example, CAGH is a simple yet often effective baseline proposed in Bonnin & Jannach (2014), which recommends the greatest hits of artists that are similar to those appearing in the seed tracks.

2.1.2 Limitations of Collaborative Filtering

The main problems with the CF methods are that (a) if there is no information about the interaction with an item is impossible to recommend it (i.e., item cold-start problem) and (b) if there is no information about the interaction of a user then is impossible to generate recommendations for that user (i.e., user cold-start problem). In this thesis, we will address the cold-start problem related to the items and not the user cold-start problem.

Another problem with the CF methods is that they also reinforce the popularity of the items, which means that they tend to choose more popular items. The problem of recommending music in the long-tail was first raised by Celma (2009) and proposed multiple novel ideas that were followed later by other researchers. The author proposed metrics to evaluate a music recommender system taking into account the popularity of the items (Celma & Herrera, 2008). By comparing the performance of different types of recommender systems, (Celma & Herrera, 2008) shows that using a CF system based on last.fm[1] data the performance is better than CB but the results reinforce popular artists and discard less known music.

2.1.3 Capturing User Behavior

In the music domain, previous work uses different signals as a way to describe the behavior of the users from what they listen to. Farrahi et al. (2014) compares the listening activities of the users in terms of playcounts, diversity and *mainstreamness*. Vigliensoni & Fujinaga (2016) goes beyond and computes the *exploratoryness*, *mainstreamness* and *genderness* from the listeners' activity which defines how the users interact with the content. Oliveira et al. (2017)

[1] https://www.last.fm

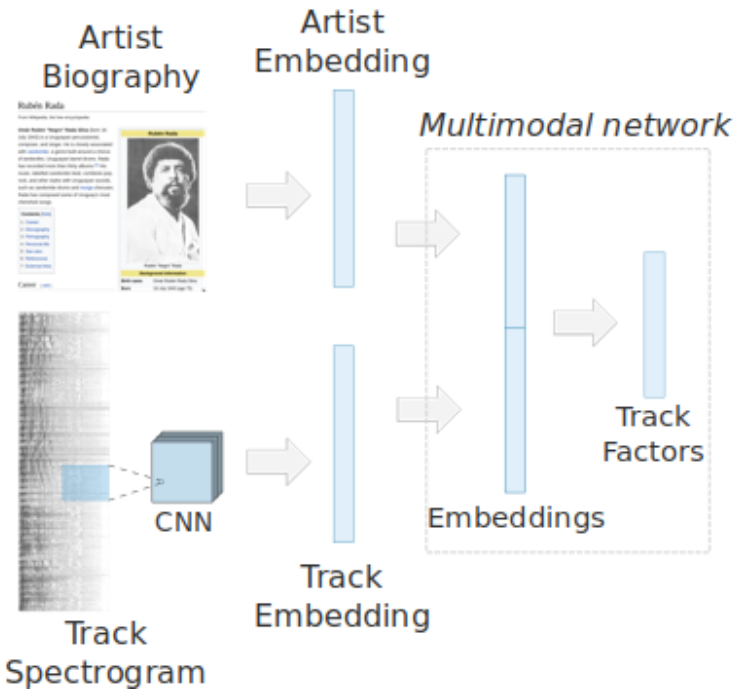

Figure 2.2: Multimodal architecture for cold-start music recommendation (Oramas et al., 2017b).

propose a multiobjective optimization approach to find a balance for diversity in the recommendation. However, as far as we know there is no prior work that tries to capture using the implicit feedback signals how much a user is engaged with a music artist. Combining for how long the user listened to the artist, including how many times a day and how many different songs of the artists listened to.

2.1.4 Long-tail and Cold-start Music Recommendation

Given the complexity of the music domain, it is clear that to give good recommendations in the long-tail or cold-start is important to have multiple sources of data or types of information (Oramas et al., 2017b) that can cover as many aspects of the music as possible. Then, the problem is how to extract relevant information that can be used for the recommendations and how to combine it.

In order to make recommendations, multiple studies try to define a similarity measure from different aspects of the music (Knees & Schedl, 2013, 2016; Bogdanov et al., 2011; Ferraro et al., 2019a).

Some recent methods automatically learn a representation from the content or context of the music. The recent advances in this direction showed good

accuracy of audio-based recommendations (Van den Oord et al., 2013), text and semantic information for the recommendations (Oramas et al., 2017b) (shown in Figure 2.2).

In the task of playlist generation and continuation, many new approaches had been proposed for the RecSys Challenge (Chen et al., 2018). In particular, the 'creative' track of the challenge was restricted for solutions that other sources of information apart from the data provided by the organizers. Therefore, some researchers proposed solutions using multimodal approaches (Ferraro et al., 2018; Zamani et al., 2019) by learning embeddings for the playlists, artists or songs.

Sequence-aware recommender systems (Quadrana et al., 2018) also can be applied to music recommendation, and some methods are based on learning embeddings or latent representations for the items, especially the ones based on recent advances in deep learning. In this field, Vall et al. (2019a) study give an example of using different types of information for improving the playlist generation from songs better represented, especially for very infrequent songs. Also, Vall et al. (2019b) investigate the importance of order, song context and popularity bias in music playlist continuation task.

It is also relevant to mention some methods that incorporate information from the user context (Schedl et al., 2012; Schedl & Schnitzer, 2013; Forsblom et al., 2012) to give better recommendations, this could allow us to understand the user needs at a particular moment, in this way we could reduce the effect of filter bubbles and it proves the importance of considering the temporal aspect of the recommendation (Schedl et al., 2014).

Finally, the advances in music auto-tagging and genre classification (Oramas et al., 2017a; Dieleman & Schrauwen, 2014; Lee & Nam, 2017; Lee et al., 2017; Pons et al., 2018; Choi et al., 2016, 2017b) can lead to better recommendations, since these models capture relevant information from the content or contextual data. Therefore, the models can be used to obtain multiple representations of the same songs that could be combined to improve the results (Oramas et al., 2018b).

2.1.5 Methods for Auto-Tagging of Music With Audio

Current state-of-the-art systems for music auto-tagging using audio are based on deep learning, in particular convolutional neural networks (CNNs), following two different approaches, one directly using the audio as input (end-to-end models) (Lee et al., 2017) and the other using the spectrograms as input (Dieleman & Schrauwen, 2014; Choi et al., 2017a; Won et al., 2020a). Previous works (e.g., Pons et al. (2018); Won et al. (2020b)) suggest that two approaches can have a comparative performance when they are applied on large datasets.

# dataset	tracks with annotation	total annotations
MASD	17,785	24,623
MAGD	23,496	37,237
top-MAGD	22,535	34,867

Table 2.1: Number of files with genre annotations in Lakh.

We can distinguish two architectures for the spectrogram-based CNN solutions, depending on whether they use multiple convolutional layers of small filters (Choi et al., 2017b, 2016) or if they use multiple filter shapes (Pons et al., 2017; Pons & Serra, 2017; Pons et al., 2018). The former is borrowed from the computer vision field (VGG Simonyan & Zisserman (2014)) and gives a good performance without prior domain knowledge, while the latter is based on such knowledge and employs filters designed to capture information relevant for music auto-tagging such as timbre or rhythm. Commonly mel-spectrograms are used with such architectures although constant-Q (Oramas et al., 2018a; Choi et al., 2017a), raw waveform approaches (Lee et al., 2017; Choi et al., 2018b) and raw short-time Fourier transform (STFT) (Choi et al., 2018b) can be also applied.

2.1.6 Genre Classification From Symbolic Music

Most of the previous works on genre classification use the audio as input. However, there are also several works based on symbolic music for genre classification. As described by Corrêa & Rodrigues (2016), the classification has been based on meta-level features, such as note duration and musical key distributions, and on identifying repeated note patterns using different features and representations. Most of the previous methods where develop in small datasets, recently the release of Lakh dataset (Raffel, 2016) of MIDI files that is mapped to Million Song Dataset (MSD) enabled larger scale studies. Table 2.1 shows the number of tracks that have at least one annotation of genre considering the different datasets asociated to MSD.

Our previous work (Ferraro & Lemström, 2018) was the first to apply a method for symbolic music genre classification in a large scale. We apply two algorithms, SIA (Meredith et al., 2002) and P2 (Ukkonen et al., 2003), they both work with polyphonic music represented geometrically as points in a Euclidean space. SIA was originally developed to discover recurring patterns, P2 to efficiently find occurrences of a query pattern in a corpus. SIA and P2 run in $O(n^2 \log n)$ and $O(mn \log m)$ time, respectively, where n and m represent the number of notes in the corpus and the number of notes in the query pattern.

Once the patterns are discovered for each track we use logistic regression for the classification.

Recent works apply deep learning techniques for large scale genre recognition in symbolic music. Dervakos et al. (2021) compares multiple CNN architectures in terms of number of layers, kernel size and size of layers. Qiu et al. (2021) proposed a method that first converts the MIDI in a vector sequence and then use the vectors as input for a deep bidirectional transformer-based masked predictive encoder. Qiu et al. (2021) method is first trained to reconstruct the MIDI tracks in an unsupervised way, then the trained encoder is used with a CNN to predict the classes. Both methods use the same Lakh dataset with top-MAGD annotations for the evaluation.

2.1.7 Datasets for Auto-tagging and Playlists Continuation

Table 2.2 summarizes the existing datasets for the tasks of music auto-tagging and automatic playlist continuation.

MagnaTagATune (MTAT) by Law et al. (2009) is commonly used for auto-tagging, but mainly for prototyping because of its small size. The Million Song Dataset (MSD) by Bertin-Mahieux et al. (2011) contains audio features extracted for one million songs, it was expanded by the MIR community with additional metadata, including collaborative tags from Last.fm. and other information such as lyrics and annotations. It was previously possible to download 30-second audio previews for MSD through the 7digital service, but it is no longer accessible. Another limitation of this dataset is the noise in the tags (Choi et al., 2018a).

To address the issue of open access to audio, the FMA (Defferrard et al., 2017) and MTG-Jamendo datasets (Bogdanov et al., 2019) were proposed for auto-tagging, both containing audio under Creative Commons licenses. The former is based on poorly structured music archives with inconsistent annotations and low-quality recordings. The latter tries to address this issue, focusing on a free music collection maintained for a commercial use-case, thus containing better quality audio and annotations. This annotations are provided by the content creators. Yet, their content is different from commercial music streaming platforms.

Recently the Million Playlist Dataset (MPD) from Chen et al. (2018) was released by Spotify. This dataset contains information about one million playlists created by their U.S. users. However, it does not include the tracks' audio information. Even if it may be possible to download 30-second audio previews with the Spotify API, it is unclear if it is legal to redistribute them. Also, there can be inconsistencies when trying to download audio previews in the future (e.g., due to songs changing their identifier or restricted access to some

Dataset	Tracks	Tags	Playlists	Audio (official)
MTAT	5,405	188	–	30 s previews
MSD	505,216	522,366	–	–
FMA	106,574	161	–	full CC tracks
MTG-J	55,609	195	–	full CC tracks
MPD	2,262,292	–	1,000,000	some previews through API
MPSD	1,993,607	–	74,996	–

Table 2.2: Public datasets for automatic playlists continuation and auto-tagging compared to streaming-platform. CC stands for audio available under Creative Commons licenses.

of the previews in different countries). These limitations significantly affect the reproducibility and complicate the use of MPD for audio research.

The Million Playlists Songs Dataset (MPSD) by Falcao & Mélo (2017) combines multiple smaller datasets (Art of The Mix by McFee & Lanckriet (2012), #nowplaying by Pichl et al. (2015), and 30Music by Turrin et al. (2015)). Similar to MPD, this dataset does not provide audio nor its representations for the songs. Since it contains playlists collected from different sources, there can be noise in the data due to song matching inconsistencies between multiple sources. Also, one of the source datasets, 30Music, was originally created for session-based recommendations instead of playlist continuation.

2.2 Metrics and Evaluation

2.2.1 Accuracy Metrics

For offline evaluation, multiple metrics had been proposed to evaluate recommender systems. To measuring the accuracy of the system, some metrics are Root Mean Squared Error (RMSE) or Mean Average Error (MAE) (Ricci et al., 2010; Schedl et al., 2018).

$$RMSE = \sqrt{\frac{1}{|T|} \sum_{(u,i) \in T} (\hat{r}_{ui} - r_{ui})^2}$$

$$MAE = \frac{1}{|T|} \sum_{(u,i) \in T} |\hat{r}_{ui} - r_{ui}|$$

Where r_{ui} is the interaction value in the test set (T) for the user u and the item i and $\hat{r}_{ui}$ is the value predicted by the system.

These metrics can be used when we want the system to predict the value rated by the user. This could be the case of explicit ratings, or if we transform from the implicit data to a scale that we want to predict.

Other systems instead of predicting the exact value focus on the ranking between the items, since the user perceives better accurate prediction on the highly-rated items rather than on the lower-rated predictions. For this purpose, some metrics from the field of Information Retrieval are used, for example, Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (NDCG), Mean Reciprocal Rank (MRR) or Precision and Recall at some cut-off (k) (Schedl et al., 2018).

To calculate Precision at k (P@k), for a user u only the top k items of the recommendations are considered:

$$P_u@k = \frac{|L_u \cap \hat{L_u}|}{|\hat{L_u}|}$$

Where L_u is the list of relevant items for the user (u) in the test set and $\hat{L_u}$ is the list of the top k items recommended for the user. Then, the values for all the users are averaged to obtain an overall Precision at k value.

Similarly, the Recall at k (R@k) is calculated using the following equation:

$$R_u@k = \frac{|L_u \cap \hat{L_u}|}{|L_u|}$$

The ranking metrics consider the order between the items in the recommendation. To measure the MAP we have to compute the Average Precision (AP) for each user and then compute the average. AP is measured using the following equation:

$$AP_u@k = \frac{1}{N} \sum_{i=1}^{|k|} P_u@i \cdot r_i$$

Where N is the number of relevant items in the test set for that user and k is the cutoff considered for the recommendation. The value of r_i is equals to 1 if the item i is relevant and 0 if not. An alternative to computing the (AP) is to use the following equation:

$$AP_u@k = \frac{1}{min(N,k)} \sum_{i=1}^{|k|} P_u@i \cdot r_i$$

The difference between the two ways of measuring AP is that when there are not many values to recommend (i.e., small value of k) the former equation penalizes the scores.

The metric NDCG was originally proposed for measuring the quality of the ranking for an Information Retrieval system, now is also used to measure the recommendations assuming that the recommended items are sorted according to the relevance for the user. NDCG is calculated from Discounted Cumulative Gain (DCG) and ideal DCG (IDCG) for each user (u) and then averaged for all the users using the following equations:

$$NDCG_u = \frac{DCG_u}{IDCG_u}$$

where

$$DCG_u = \sum_{i=1}^{|R|} \frac{rel_i}{\log_2(i+1)}$$

$$IDCG_u = \sum_{i=1}^{|G|} \frac{1}{\log_2(i+1)}$$

Where R is the list of the recommended tracks for a user, and G contains the ground-truth tracks for the user. $|.|$ denotes the length of the list of tracks and rel_i is the interaction value for the current user (u) and the given track.

Mean Reciprocal Rank (MRR) measures the first relevant item for each user and is calculated using the equation:

$$MRR = \frac{1}{|U|} \sum_{u \in U} \frac{1}{rank_u}$$

Where $rank_u$ returns the position of the first relevant value for each user (u).

2.2.2 Beyond Accuracy

In the previous section, some metrics for measuring the accuracy of the recommendations were mentioned. However, previous works (Ziegler et al., 2005; Knijnenburg et al., 2012) show that users might not always perceive as better when a recommender system has higher accuracy. Therefore, in this section, we describe alternative metrics that can be used to evaluate the recommendations offline.

In the case of music long-tail recommendation, Celma & Herrera (2008) evaluates a CF and CB systems from two perspectives: the quality perceived by the user of the recommendations and from the items similarity, taking into

account the popularity of the items. Bogdanov et al. (2013a) also evaluates different methods based on CF and CB with users and proposes a method for visualization of the users preference.

Also in the field of music recommendations, some metrics beyond the accuracy are summarized by Schedl et al. (2018), these metrics are Spread, Coverage, Novelty, Serendipity and Diversity:

- The measure Spread is related to the number of times each item is recommended. The perfect value of Spread is when all the items are recommended the same number of times. Therefore, Spread is calculated as the entropy of the distribution.
- The measure Coverage can be computed as the percentage of items that the system recommends globally. While the measure Novelty (or also referred to as Freshness) is measured for each user and takes into account the global popularity of the items.
- The concept of Serendipity is related to how unexpected and useful is a recommendation for a user. Therefore, we need to quantify the unexpectedness and the usefulness of an item for a user. A way of measuring the unexpectedness is by using a distance metric between what usually the user listens. For measuring the usefulness, the user could either explicitly rate the recommendation or it can be inferred from the listening history.
- For measuring Diversity, we also need to define a distance measure between the items. Some metrics that can be used are inverse cosine similarity, inverse Pearson correlation, or Hamming distance.

Furthermore, Vargas & Castells (2011) defines other metrics for Novelty and Diversity in the context of ranked lists of items. The author compares the performance of multiple recommender systems according to these metrics and how they are related to some accuracy metrics. In another work, Vargas & Castells (2013) re-rank the recommendations to change Diversity for different profiles of users and they show improved accuracies. Similarly, Schedl & Hauger (2015) shows that is possible to categorize users according to their preference regarding Novelty, Diversity and Mainstreamness (or Popularity), improving the overall accuracy of the recommendations.

Some works (Ziegler et al., 2005; Hu & Pu, 2011; Adamopoulos & Tuzhilin, 2015) show the importance of Novelty and Diversity to give good recommendations and increase user satisfaction.

Catalog coverage is a traditional quality factor in recommender systems and measures how many of the available items are presented to users in their top-n lists (Herlocker et al., 2004). They propose to measure it by creating the union

of the top-10 recommendations at a given point in time, and they emphasize that the metric should be combined with accuracy measures.

In other works in the information systems literature, catalog coverage is often referred to as *aggregate diversity* (e.g., Adomavicius & Kwon (2012)). One assumption is that higher aggregate diversity will ultimately lead to higher *sales diversity*, as investigated in Lee & Hosanagar (2019).

Besides aggregate diversity, researchers also frequently investigate diversity at the level of the individual user (Castells et al., 2015). Aggregate and individual diversity are, however, not necessarily correlated (Lee & Hosanagar, 2019). One can, for example, recommend the same set of highly diverse items to everyone, which does however not lead to high aggregate diversity (Wang et al., 2019a).

Alternative model-based strategies for counteracting in particular popularity biases for traditional recommendation scenarios were also proposed (e.g., Abdollahpouri et al. (2017); Jannach et al. (2015b)).

2.2.3 Provider Metrics

Traditionally, the evaluation of recommender systems is user-centered, most of the metrics described in the previous sections are intended to capture different dimension regarding the user satisfaction. Recently, Abdollahpouri et al. (2020a) summarizes some alternative metrics to capture the behavior of the recommender from individual providers

In our case, artists are the providers of the content. The metrics described by Abdollahpouri et al. (2020a) could be applied for the individual artists or also to the groups of songs according to the desired attribute (e.g., genre, style, country or year). In the following metrics p corresponds to a given provider, $i_p \in I_p$ are the items associated with p. $\mathcal{L}$ is the list of recommendations for n users. T is the set of r_{ij} ratings in the test set where $\mathcal{L}$ is calculated. $\mathbb{1}$ is the indicator function and $m(r_{ij}, \hat{r_{ij}})$ is an accuracy metric. $g_p(i)$ is a boolean function that returns true if user i is the target of provider p.

- Exposure: Count the number of recommendations given across all the p's items

$$Exposure(p) = \sum_{L_i \in \mathcal{L}} \sum_{j \in L_i} \mathbb{1}(j \in I_p)$$

- Hits: Count the number of hits in recommendation lists for all the p's items

$$Hits(p) = \sum_{L_i \in \mathcal{L}} \sum_{j \in L_i} \mathbb{1}(j \in I_p \wedge r_{ij} \in T)$$

- Reach: Count how many users get at least one i_p item recommended

$$Reach(p) = \sum_{L_i \in \mathcal{L}} \mathbb{1}(|I_p \cap L_i| > 0)$$

- TargetReach: Count how many users in p's target set get at least one i_p item recommended

$$TargetReach(p) = \sum_{L_i \in \mathcal{L}} \mathbb{1}(|I_p \cap L_i| > 0 \wedge g_p(i))$$

- PAccuracy: Average metric m score for predictions of p's items

$$PAccuracy(p,m) = \frac{\sum_{r_{ij} \in T_p} m(r_{ij}, \hat{r_{ij}})}{|T_p|}$$

2.3 Popularity Bias and the Long-Tail

Already long before music streaming platforms were the main source for music consumption, Anderson (2004, 2006) popularized the concept of the long-tail economy. He proposed that in the online business (e.g., Amazon, Yahoo, Apple), using recommender systems would create a big opportunity to sell more items by exploiting small niches (i.e., selling small amounts of a lot of products that form the tail in the popularity curve); physical stores, in comparison, have a limited stock capacity and typically restrict their stock to the more popular items. Since the early 2000s, there is an ongoing debate (Benghozi & Benhamou, 2010) about whether these 'new' online platforms indeed allow users to consume more of the long tail or, on the contrary, accentuate the 'superstar phenomenon' (Rosen, 1981) where the distribution of popularity of artist gets more skewed in the long-term (Coelho & Mendes, 2019; Bauer et al., 2017).

The work from Celma (2010) was the first to study how different recommender systems may promote or not the less popular items in the music domain. The work from Fleder & Hosanagar (2009) focuses on the effect of recommenders but not only in the music domain. The results suggest that with the recommender systems users tend to consume a reduced number of items, reducing the diversity. At the same time, the authors mention that the system is helping the users to discover new items. While these works focus on the consumers'

behavior, the findings indicate that recommenders have an impact on item providers. In other words, it is important to consider the impact from the item providers' perspective, otherwise, we risk strong negative effects in music and other domains.

In another recent work from Spotify, Anderson et al. (2020) identify that when users follow more algorithmically-generated recommendations the diversity of the content they consume is also reduced. In contrast, users who consume increasingly more diverse content, are those who reduce the algorithmic consumption and increase their organic consumption. Anderson et al. (2020) conclude that—in the context of Spotify—, recommendations are more effective for users with lower diversity.

2.3.1 Simulations of Recommender Systems

In recommender systems literature, a dataset of listening events of users is typically used as ground truth to evaluate offline if the recommendations are accurate or not. However, there are strong limitations in such offline evaluations, since a) This does not allow us to understand how the users would have behaved with a different set of recommendations, i.e., recommending something outside of the ground truth items does not mean that the user would not like it; b) Offline evaluation usually has a popularity bias, favoring the algorithm that recommends more popular items (Bellogin et al., 2011; Steck, 2011; Park & Tuzhilin, 2008; Ferraro et al., 2019b), which can also vary depending on demographic aspects of the users (Ekstrand et al., 2018a), and can have a disparate bias for different users groups (Lin et al., 2019). Therefore, if these systems are optimized for accuracy they may privilege popular items, which are not necessarily more satisfying for the users (Konstan & Riedl, 2012; McNee et al., 2006). Given the limited possibilities for evaluating longitudinal effects of recommender systems in the field, several studies rely on simulating the feedback loops to analyze these potential effects (Jannach et al., 2015b; Zhang et al., 2020a; Ferraro et al., 2020c).

Using simulation-based techniques as a research method, e.g., in the form of agent-based modeling, has a long tradition in various fields, like in managerial science (Wall, 2016). Simulation-based research is however comparably rare in the field of recommender systems. Recently, Zhang et al. (2020a) used agent-based simulation to analyze longitudinal effects of recommender systems. Among other aspects, their simulations revealed a phenomenon called *performance paradox*, where it turned out that a strong reliance by users on the recommendations may lead to suboptimal performance development over time. It was also found that recommender systems can concentrate on a small set of items. Such concentration biases, measured in terms of the Gini index, were previously explored in Jannach et al. (2015b). Both Zhang et al. (2020a)

and Jannach et al. (2015b) observed in their simulation approach that some algorithms may lead to a concentration over time. However, some algorithms were also suited to increase catalog coverage and to decrease the Gini index over time.

2.3.2 Locality and the Long-tail

Based on the idea that many local artists tend to be obscure long-tail artists, there are research endeavors (e.g., Turnbull & Waldner (2018); Akimchuk et al. (2019)) to "localify" recommendations with the goal to promote local artists. The authors provide technical approaches to localize music recommendations. While Turnbull & Waldner (2018) and Akimchuk et al. (2019) take "small geographic region (e.g., 10 square miles)" as their point of interest, a recent work by Spotify (Way et al., 2020) studies music consumption on a country level. The latter study investigates how users from a country consume content from another country, and how this consumption pattern evolves and changes over time. The authors identify that language and geographical proximity have an impact on the consumption between countries.

2.4 Algorithmic Fairness and Multi-Stakeholders

It is clear that platforms (such as Amazon, Spotify or Google) have an important impact on multiple groups of people and also business, therefore, different behaviors of the recommender systems in these platforms will affect these groups either positively or negatively. Recently, in the field of recommender systems, the idea of considering the impact of the systems on different stakeholders has increasingly gained more attention (Burke, 2017; Abdollahpouri et al., 2020a). The considered groups of stakeholders are the item providers, item consumers and the platforms. The typical approach is to define the metrics that capture the interests of each stakeholder and then optimize them together, finding a balance between the interests of all the groups of people involved.

While fairness for various consumer groups increasingly gains attention in recommender systems research (e.g., Ekstrand et al. (2018a); Yao & Huang (2017); Farnadi et al. (2018)), studies on fairness considering other stakeholders are scarce (e.g., Ekstrand et al. (2018b)).

For the music domain, Mehrotra et al. (2018) consider the interests of the music consumers and the artists when generating recommendations. To achieve algorithmic fairness, they define a 'fairness metric for artists' based on the popularity distribution in the recommendations. While Mehrotra et al. (2018)

point out that there are different ways of defining fairness, it is not justified why the chosen metric is aligned with what the artists expect from a fair system.

More generally, multiple works define algorithmic fairness in the field of Artificial Intelligence (Hutchinson & Mitchell, 2019; Lee et al., 2020b). Maybe the most common definitions refer to 'group' and 'individual' fairness. Individual fairness reflects that similar individuals should be treated similarly. Group fairness ensures that people of a protected group should be treated in the same way as the rest of the population. Dwork et al. (2012) clearly distinguishes those two concepts, however, other works (e.g., Binns (2020)) suggest that individual and group fairness are not contradictory and may be achieved simultaneously. In the field of information retrieval, fairness is also defined in terms of *exposure* (Biega et al., 2018; Sapiezynski et al., 2019) or *attention* (Singh & Joachims, 2018). Thereby, Biega et al. (2018) focus on individual fairness, whereas other works typically consider group fairness (Singh & Joachims, 2018; Sapiezynski et al., 2019). While the idea of exposure of the artists could also be adopted for the music domain, it is not yet answered how a *fair exposure* should be operationalized or—on an even deeper level—what fairness means in the context of streaming platforms.

2.5 Qualitative Studies Regarding Perception of Algorithmic Fairness

The general perception of fairness on automated decision-making systems was studied by Helberger et al. (2020) using surveys. Respondents consider that systems can make more objective decisions and also process larger amounts of information compared to humans, which allows them to make fairer decisions. However, respondents agree the systems are limited in the generalizability and modeling of reality. Harrison et al. (2020) compares the human perception of fairness in realistically-imperfect systems using surveys. From the findings, the authors highlight how respondents have contradictory opinions regarding the preference of the systems, showing that is not possible to achieve a broad acceptance in society regarding the "right" fairness definition. The authors also highlight that there is a general preference towards human judges compared to the systems, even if participants consider the systems fair or unbiased.

Wang et al. (2020) conduct an online experiment where participants can rate algorithms according to their perception of fairness. The authors find that participants rate systems as more fair if they are favored by the prediction, even when the algorithms were explicitly described to the participants as very biased to a particular demographic group. The authors find this effect in different levels depending on education level, gender and other aspects of the participants.

According to Srivastava et al. (2019), people prefer more simple definitions of fairness compared to complex ones.

Woodruff et al. (2018) conducted group discussions and interviews with populations from traditionally marginalized of the US to understand their perceptions of fairness in algorithms. In this work, the authors highlight that the opinions of the participants regarding the fairness of algorithms vary depending on individual factors, context, different stakeholders' perspectives and different framing of fairness. Therefore, this can help to explain why different studies appear inconsistent regarding some findings. Also, this shows that the contextual factors should be taken into account when studying algorithmic fairness, supporting the idea of considering the interests of the different groups of stakeholders of the systems.

In another related study, Pierson (2017) found that women are likely to oppose the inclusion of gender as a feature in course recommendation algorithms if such algorithms are less likely to recommend science courses to female students.

Smith et al. (2020) raise the fundamental question, "what is fair in the context of recommendation—particularly when there are multiple stakeholders?". In their initial study, the authors interview users to understand their ideas about fair treatment in recommendations, deriving common topics from the participants' answers.

In the music domain, to understand what fairness is from the artists' perspective, we need to involve artists. To the best of our knowledge, there is no public research work that reaches out to artists to identify how they feel affected by current music platforms and how they believe the embedded recommenders systems could be improved to be fair for them.

2.6 Systems, Gender and Discrimination

Gender bias has been of particular research interest. Studies have shown that the design of software (Vorvoreanu et al., 2019; Burnett et al., 2011; Beckwith et al., 2005, 2006) or websites (Metaxa-Kakavouli et al., 2018) may introduce bias for users of different genders.

Vorvoreanu et al. (2019) studies the gender bias in the design of software. They derived design changes to a particular piece of software to favor inclusiveness and then ran a study with participants to compare both versions. The results show that women succeeded more often in the new version than in the original, removing the gender gap. Metaxa-Kakavouli et al. (2018) show that the design of a website can introduce a bias for users of different genders, which may have a negative impact in the interest of different groups of people, such as young females, to study computer science broadly.

Another thread of research investigates gender bias in algorithmic decision-making. For instance, Keyes (2018) criticizes how gender is defined in previous works on automatic gender recognition in the field of human-computer interaction and analyzes the papers that were published in this field to understand how they define gender. The current implicit understanding of gender and how it is implemented in such systems negatively affect transgender people. The authors suggest how the field can move forward to a more trans-inclusive treatment of gender and propose alternatives to automatic gender recognition.

More recently, research on gender bias in recommender systems has gained attention. For instance, studies have shown that various algorithms perform differently for different demographic user groups (Ekstrand et al., 2018a) and that common collaborative filtering algorithms differ in the (book authors') gender distribution in the computed recommendation lists (Ekstrand et al., 2018b). The analyzed algorithms provide different distributions of recommendations for each gender.

In this research thread, several metrics have been introduced for evaluating algorithms for gender biases (Ekstrand et al., 2019; Ekstrand & Mahant, 2017; Singh & Joachims, 2018; Lin et al., 2019).

2.6.1 Gender Bias in the Music Domain

In music business research, several works study the discrimination and bias related to gender. Schmutz & Faupel (2010) investigate the factors that influence reaching an audience. They conclude that for female artists it is less likely to reach an audience only for being female. Smith et al. (2018) analyze the gender and ethnicity of the artists of the top 600 songs in the Billboard charts from 2012 to 2017. The authors show many important differences between genders across the different years; e.g., in the year 2017, there was a particularly high imbalance with a ratio of 4.9 male artists to every 1 female artist. Also, female artists are underrepresented in Grammy nominations, with only 9% of the nominees being female. Watson (2020) analyzes the airplay of country music from metadata in Canadian radios in the last 15 years, considering the gender and country of the artists. The author found that less air time is dedicated to female artists compared to male artists, and that female artists are typically aired in hours with less audience.

Recent works studied the role of music recommender systems on gender bias. For example, Oliveira et al. (2017) investigate gender diversity in music recommendations, where they compare methods for multi-objective optimization in music recommendations. In this work, the authors consider the type of the artist (i.e., 'band', 'orchestra', 'solo', etc.) as a gender.

Shakespeare et al. (2020) evaluates offline the recommendations of multiple

collaborative filtering approaches in terms of gender distribution. Their approaches consider the gender of the users and the gender of the artists, concluding that these methods propagate the gender bias that is present in the dataset.

Also, on the streaming platform Spotify, users tend to listen more to male artists. Epps-Darling et al. (2020) select a group of users and manually annotate the gender of all the artists they listen to in a month, considering different gender identities. The authors compare organically generated listening events on Spotify with the ones that are generated algorithmically, considering 'algorithmic' in a wide sense (e.g., including a user's most listened to songs presented on the platform's homepage). Epps-Darling et al. (2020) found that female artists and mixed-gender groups are particularly present in the lowest level of popularity. Yet, they found that female artists are also present in the highest level of popularity, which the authors mainly attribute to the different gender representations in different music styles. Similarly, Aguiar et al. (2018) analyzes playlists on Spotify regarding gender bias. Authors identify that females account for between a seventh and nearly a third of streaming. They found some evidence for bias (e.g., in favor of women at Today's Top Hits and in the New Music rankings, yet against women at global playlists). Still, playlist inclusion did little to explain the low female share of streaming on Spotify. The authors speculate that the low share is attributed to the relatively low share of female songs entering the platform. The authors analyze the presence of female artists in two playlists rankings. In one of them, it is shown a higher presence of female artists and on the other a lower presence, but the authors do not find here an explanation for such high difference between male and female in the number of streams.

2.6.2 Artist Exposure Biases in Collaborative Filtering for Music Recommendation

To make sure that streaming platforms give an opportunity to all the artists of reaching an audience is important to consider the exposure that the recommendations give to the different groups of artists. In that way, we can see if the systems do not discriminate or have undesired bias.

We are not aware of previous studies of other authors evaluating the potential impact of music recommendations on user behavior and artist exposure, which consider different types of artists. Therefore, in a previous work (Ferraro et al., 2020d) we focus on the exposure of music artists and analyze the recommendations made with Collaborative Filtering to quantify how differently the system promotes various types of artists. We consider grouping the artists using different attributes: location, gender, period, and artists type (e.g., solo, band, orchestra). We analyze the amount of users reached by each artist group in

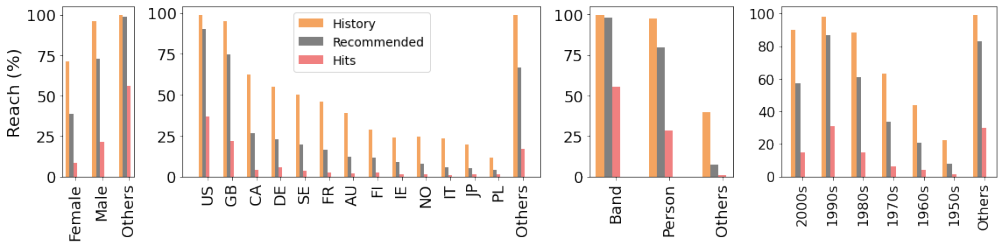

Figure 2.3: A comparison between user history, recommendations, and hits in terms reach of different artists grouped by gender, country, type, and period.

general. For each artist, we compare their exposure in the recommendations with previous listening data to see whether the system promotes or punishes certain groups.

Figure 2.3 presents a comparison of different groups of artists by the how many users they reach calculated for both the listening history and the recommendations from lfm-360k dataset. In this figure, we see that recommendations increase the biases for the reach metric compared to the users history.

CHAPTER 3

Understanding Fairness in Music Streaming Platforms

3.1 Introduction

Music streaming platforms are currently among the main sources of music consumption. What the users consume is strongly influenced by what is offered on the music platforms, and what is promoted through the algorithmic recommendations in particular. What the users consume, in turn, shapes the music streaming ecosystem at large. Traditionally, the platforms were focused on maximizing user's satisfaction since the users are the main source of income through monthly subscriptions or advertisements. However, as described in Section 2.4, there is an increasing research interest in considering how the different stakeholders are affected by the platforms and their recommendations.

A recent literature review by Milano et al. (2020) demonstrates the research gap in considering the provider's interests (and the interests of the society at large) when assessing the ethical impact of recommender systems. They stress the need to consider the provider's interests and the interests of society at large while existing literature tends to consider the implications for the receivers of the recommendations only.

As recommender systems play an important role in connecting users with the artists and their music on such music platforms, therefore, it is important to understand the societal and ethical implications of these systems. Previous works (e.g., Cramer et al. (2018, 2019a)) have demonstrated that fairness can not be operationalized with a lack of a definition or understanding of what fairness means in a certain context. Therefore, a purely technical approach is not sufficient for defining and operationalizing fairness in practice.

The long history of definitions of fairness in various disciplines shows that the notion of fairness has evolved over time (Hutchinson & Mitchell, 2019); and

particularly recent literature (e.g., Holstein et al. (2019); Selbst et al. (2019)) emphasizes that for developing a good understanding of fairness in a given context, it is crucial to take this context into account, taking a multidisciplinary point of view and listening to the opinions of the many groups of stakeholders involved and affected.

In this chapter, we address the research gap by taking an exploratory approach and focus on the music artists as the main item providers for music platforms. In this chapter, we focus on understanding the different dimensions of music platforms and their recommender systems that define fairness from the artists' perspective.[2]

To this end, we conducted semi-structured interviews with music artists from different countries and with varying popularity. Following a Qualitative Content Analysis as proposed by Mayring (2004), we have (i) identified multiple aspects that show how the artists feel affected by the current music platforms and their integrated information retrieval and recommender systems components, and (ii) how such music platforms and systems could be improved with respect to fairness. Only such understanding can ultimately lead to fairer music platforms that are not only optimized for the interests of the platform providers or the consumers.

The contributions of this chapter are to (i) understand the impact that such music platforms have on artists from their perspective and (ii) to understand what the artists consider fair for them.

3.2 Methods

This section describes the methods followed to investigate the impact of music streaming platforms on artists and to explore viable solutions for fairer music platforms and music recommender systems, focusing on understanding the artists' perspective on what represents a fair music platform.

In this section, we first describe how the interviews were designed and carried out. Then we give some details about the participants of the study and finally we explain how the information collected during the interviews was processed and analyzed. All the material used for the interviews is available on Zenodo[3], including the invitation letter, consent form, initial version of questions and final version of questions.

[2]The findings described in the following sections are based on our published work presented in Ferraro et al. (2021d)

[3]https://doi.org/10.5281/zenodo.4793395

3.2.1 Interviews

Between December 2019 throughout March 2020 we carried out 9 semi-structured interviews with music artists. According to research practice (Creswell & Poth, 2016; Morse, 1994), the sample size is adequate and, more importantly, we reached a high level of thematic saturation (Guest et al., 2006) with the same topics being repeatedly mentioned across the interviews. This means that we could identify that there was a high agreement between the topics that the participants mentioned and that the same comments started to be repeated again and again.

The interviews were designed to last one hour in total. We used the initial 10 minutes of the interview to explain the purpose of the project and to give a brief and general introduction to music recommender systems and how these are integrated with current music platforms. For the interview part, we used open questions with the goal that the artists could bring up their own new ideas that might not have been considered in the field before. In addition, we proposed some specific situations to gather the artists' opinions about how the systems should behave according to their perspective.

Following, we detail the interview process. Prior to the interviews, we informed that the data and the results of the interviews were going to be kept anonymous at all times, which was important so the artists could feel free to share any concerns with regard to the music platforms or other issues related to the music industry. As some of the participants sometimes used strong language and addressed delicate issues, we believe that we were able to maintain a comfortable atmosphere with a high level of trust. In addition to a consent form, we asked the participants to fill out a short form with optional information about themselves that would be used to refer to their answers (i.e., age, gender, country, genre, popularity level, number of records/singles, years active in the music industry, contracts with labels).

We defined a guiding protocol to be used during the interviews which included a set of tentative questions and follow-up questions to encourage the interviewees to elaborate further. The protocol was developed so that each guiding question addressed and explored various different topics and issues of how the music platforms might affect the artists. As the first step in elaborating this guide, we started with collecting information and ideas about the general aspects of current music platforms that could affect the artists. In a second step, we identified many potential questions that could be used in the interviews and formulated these questions to address the identified issues. Since the rich collection of potentially interesting issues had to be reduced to fit the time scope of an interview and to have a narrower focus, in the third step, each of the participating researchers gave a score between 1 and 3 to each of the original 36 questions according to priority, with 1 being the highest priority.

Then, the team discussed the 14 questions with differing scores until consensus was reached.

We relied on the extensive expertise of our research team conducting user studies and, in particular, qualitative studies to score the candidate questions based on the evidence collected during this process.

As a result, based on the agreement between the scores, we defined a list of guiding 21 questions composed of 11 as main questions and 10 as sub-questions to be used in the interview protocol and the order of the questions. Table 3.1 provides an overview of the guiding questions used in the interviews. Note that all interviews were held in Spanish and the materials provided to participants were all in Spanish.

We note some limitations of our interview design. First, while we cater for diversity, in our work we do not intend to represent the opinion of the whole population of music artists. The findings we report should, thus, be viewed as a deep exploration of our sample's beliefs and attitudes, but not as generalizing to the music artists population as a whole. Moreover, the lack of representation of many groups of minorities within the participants limits the scope of our findings. Nevertheless, after processing the results we found a saturation in the answers to the questions, indicating that we could expect that more interviews would not affect the results. Second, while we assure anonymity throughout the process, the interviewer naturally knows the participants' identity, which may have influenced participants such that some issues may not have been voiced. As some of the participants sometimes used strong language and addressed delicate issues, we believe that we were able to maintain a comfortable atmosphere with a high level of trust.

3.2.2 Participants

We recruited 9 music artists that we consider diverse in the kind of music they perform, their popularity, location, age, and gender. Four of the participants are between 26–35 years old, four are between 36–45, and one is within the range 46–55. Seven are male and two are female. Most of these artists have many projects in parallel (one is a solo artist, the others work with many bands). Regarding the location, the artists were born in four different countries (i.e., Australia, Spain, Russia, Uruguay) and started their careers in three different countries (i.e., Spain, Uruguay, Cuba).

The genres that the artists consider their music sum up to a total of 24. Some examples are: folk, pop, ska, punk-rock, dubstep, drum & bass, world music, etc. The number of years in the industry is between 4 and 25. The number of albums released ranges between one and ten. Five participants consider themselves 'independent artist', three have a contract with one of the three major

Table 3.1: Guiding questions in the interview protocol.

No.	Topic	Guiding question
1	Convey interest, gain trust	Do you use any platform to listen to music? What's your experience with it?
2	Reflecting	Do you think your career would be much different without these systems?
3	Lack of control	Which of your music tracks should be recommended more and which less?
4	Bias to more popular	There are groups of artists that are not recommended by the system because of different reasons. Do you see any alternatives for this?
5	Diversity	Music considered "niche music" is not recommended to many users, should the system nurture diversity (e.g., in terms of genres, styles, artists from all over the world, popular and not-yet-popular) or focus more on recommending what the user is familiar with?
6	Size of repertoire	If artist X has more music than the artists Y, do you think the system should recommend more music by artist X—or should the recommendation be independent of an artist's repertoire?
7	New artists	For a given user out of 100 recommendations, how many do you think should be new artists?
8	New music	Should your older songs be promoted more than your newer songs?
9	Country quotas	In a music platform that has more users from country X but more artists from country Y, the artists from X could be recommended more than artists from Y. What is the behavior that you expect from the system in that case?
10	Influencing the users	Currently, K-pop is the 7th most listened genre, over R&B and classical music. Such a popularity distribution could also refer to gender, country, or other aspects of the music. Do you think the systems should try to reproduce this behavior, or should try to provoke a change on it?
11	Income distribution	Do you think the current model based on number of streams is good or there could be a better model?

Table 3.2: Information about the participants.

ID	Country	Age	Music Styles	Audience	Gender	Contract
P1	Uruguay	46-55	Rock, Folk, Hip-Hop, Electronic	International	Male	Major Label
P2	Uruguay	26-35	Rock, Hip-Hop, Reggae, Dub	Local	Male	Major Label
P3	Uruguay	36-45	Ska, Punk/Rock, Dub, Dubstep, Drum & Bass	International	Male	Major Label
P4	Spain	26-35	Indie, Rock	Local	Male	Independent
PF1	Cuba	36-45	World, Jazz, Cuban Music, Electronic	International	Female	Independent
PF2	Spain	26-35	Indie Pop, Singer-songwriter	Local	Female	Indie Label
PN1	Uruguay	26-35	Alternative Rock/Indie, Progressive Rock	Local	Male	Independent
PJ1	Spain	36-45	Jazz, Free Improvisation	International	Male	Independent
PR1	Spain	36-45	Rap/Hip-Hop, Reggae, Blues, Salsa, Flamenco	International	Male	Indie Label

labels (i.e., Sony, Universal, Warner), two have a contract with an independent label. We also asked to the artists if they considered that are known in a global, regional or country level. From the answers we found that five artist consider themselves internationally known while four are known within their country.

The specific information of each participant is given in Table 3.2, including the identifier that we use to refer to participants for quotes.

3.2.3 Processing and Analysis of Interviews

Following the methodology of Qualitative Content Analysis (Mayring, 2004), the interviews were recorded and transcribed. An annotation scheme (i.e., coding) was defined inductively from the transcriptions of the interviews and was used to annotate the transcriptions accordingly.

The total duration of the recordings is 420 minutes, which transcribed corresponds to 33,669 words. The annotation scheme was developed inductively from the transcriptions, where statements were used as the level annotations. Often, statements were on sentence level; yet, many sentences include two or more statements. Note that the annotation scheme was developed in English while the transcriptions were kept in their original language (i.e, Spanish). The transcriptions were manually processed in order to annotate each section according to the topic that it was identified. The development of the annotation scheme and the annotation of statements itself was an iterative process where we assigned a topic to a specific sentence and if it did not fit to any of the previous topics then a new one was defined. We iterate a total of 4 times and the final annotations were reviewed by a different person than the annotator in order to increase intercoder reliability. The original topics were also refined

Table 3.3: Details on the annotation scheme.

Topic	Description	Example of annotation
User view	The participants comments on how or when they use a music platform in the role of a music consumer (user).	*PN1: "I usually read the artists biography and the influences of an artists."*
Artist view	The participant expresses an opinion from the point of view of an artist.	*P2: "[...] it would not hurt if the systems were more random, not that obvious—if you like 'Beatles,' I recommend 'Rolling Stones."'*
Lack of control	Reference to giving more control either to the artists over the music presented, or to users over what they get recommended.	*P2: "As an artist, I would love to have more freedom of action over my music on the platform."*
Diversity	Related to any aspect of diversity in the recommendation.	*P3: "[As an artist] I would expect that the music platforms promote more diverse content."*
Context of music	Aspects related to information and presentation apart from the music and the artist themselves; the context that it is embedded in.	*P1: "There are songs that have history, [you cannot ignore that]."*
New music	The participants refers to new music of existing artists or to artists that are new on a music platform.	*P2: "In my opinion it makes sense if half of the recommendations [made by the music platform are songs of] new artists."*
Popularity bias	The participant refers to aspects related to popularity bias of recommender systems or the music business.	*P4: "The problem is [that] the recommender system systematically ignores all those potential artists because is easier to recommend [what is more popular]."*
Influencing users' behavior/ taste	The participants express an opinion regarding the music platforms' opportunities to influence users' listening behavior or musical taste.	*PF1: "In my opinion you can't impose some [specific] music to the users."*
Transparency	Refers to the need of more information about how a music platform works and its recommender system makes decisions.	*PN1: "If a human makes that decision, if he says 'I have a small store, I am going to put it this way,' I will understand it better than if an algorithm does it."*
Labels'/ platforms' interests	The participant refers to the interests of stakeholders such as the music platform providers or the record companies.	*P2: "[A platform] needs to take responsibility for its recommender system—to understand the situation. [...] Obviously, they are commercially not capable or not interested."*
Size of artists' repertoire	The participant distinguishes between artists with more or less songs or albums.	*PR1: "If you have more songs you have more chances to satisfy different audiences."*
Quotas for local music	The participant mentions regulations for local music such as minimum quotas for local artists on the music platforms.	*PR1: "Otherwise you won't know what there is in your country."*
Gender balance	The participant talks about gender bias in the music industry or the music platforms, or how the recommendations might be fair(er) from a gender perspective.	*PN1: "[...] the population of the world is 50% women. So it would be ridiculous if the system wouldn't recommend it."*
Regulation of recommendations	The participant refers to regulations or policies about the music platform or its recommender systems.	*P3: "[...] the question is if it should be imposed by the state [to promote local music]."*
Artists' income distribution	The participant refers to royalties generated on the music platforms and how they are distributed among artists.	*PF1: "It is absurd what [the platform] pays to the artists."*

taking in consideration the general topics identified when defining the guide for the interviews summarized in the Table 3.1.

The final annotation scheme includes a total of 15 overall topics which we obtained from 752 annotated text sections annotated. The topics defined for

Table 3.4: Statistics about annotations.

Topic	P4	PF1	P3	P2	P1	PN1	PJ1	PR1	PF2	**Total**
User view	8	1	5	8	6	5	4	2	7	**46**
Artist view	15	12	25	22	14	17	4	9	13	**131**
Lack of control	15	4	8	14	6	5	0	2	8	**62**
Diversity	5	2	8	6	4	2	1	3	8	**39**
Context of music	7	6	8	2	17	1	0	5	0	**46**
New music	12	6	7	8	6	14	6	10	13	**82**
Popularity bias	6	0	8	5	1	4	2	5	1	**32**
Influencing user behavior/taste	7	2	17	7	16	3	1	7	7	**67**
Transparency	7	0	11	13	2	6	0	2	2	**43**
Labels'/platforms' interests	8	4	14	17	13	5	5	11	5	**82**
Size of artists' repertoire	2	1	2	2	2	4	1	2	1	**17**
Quotas for local music	3	3	8	3	5	5	2	2	3	**34**
Gender balance	5	3	5	4	0	2	1	1	2	**23**
Regulations of recommendations	3	0	3	0	4	0	0	0	0	**10**
Artists' income distribution	3	1	4	9	4	4	3	7	3	**38**
Total	106	45	133	120	100	77	30	68	73	**752**

the annotations are: (use/show information of) context of music, new music, diversity, lack of control, transparency, popularity bias, size of artist's repertoire, quotas for local music, influencing users' behavior, gender balance, artists income distribution, regulations/policies about recommendations, labels/platforms' interests, user view, and artists view. In Table 3.3, we provide a description of the topics with an example of an annotation for each topic. Note, quotes are given as translations to English, whereas the interviews were held in Spanish. Table 3.4 presents the number of annotations per participant per topic. The annotations allowed us to analyze to which degree the artists agree on the topics that they discuss. As we can see in the Table 3.4 most of the topics where mentioned multiple times by each participant.

3.3 Discussion of Findings

In this section, we describe the main findings that emerged from our analysis. In the first part of this section (Subsections 3.3.1 to 3.3.3), we present how the artists feel affected by current music platforms. Subsequently (Subsections 3.3.4 to 3.3.8), we discuss those topics that give direction on what the artists consider fair music platforms; we only report those topics that were addressed by *all* participants. Table 3.5 provides an overview of these topics and summarizes what the participating artists deem necessary for future fair music recommender systems.

3.3.1 Fragmented Presentation

The participants report that they do not find adequate the way they are presented on the music platforms. The artists also mention that they feel negatively affected by how their message and public image are presented.

Two artists (P2 and P3) mention that their artist profiles on the music platforms show some of their tracks that are very old at the top, just because those are the most listened items over the years.

> *P3: "But it is something that I have done 10 years ago. [The platform] puts the most listened tracks. When new users reach our profile to know the band they listen to these tracks that I do not feel identified"*

Also, the context in which artists and their music are presented may affect their public image; e.g., the artist P3 refers to a feature that serves users with an automatically generated playlist (called "radio"). The "radio" of an artist is a infinite playlist that includes tracks of this artist and also music by other artists. P3 reports that the "radio" based on his band includes music that he does not like and features artists that he distances himself from ideologically. In this case, the artist mentioned that he found on his band's "radio" music that was used for the campaign of a right-wing Argentinian party and the ideology of their band is left-wing.

> *P3: "I see things that I do not like and that I reject ideologically. Why does [Band X] appear?"*

P3 explains that the band works hard on creating a certain public image—with the lyrics, the music, the art. The music platform then mixes it with something completely different.

> *P3: "I don't think the same people listen to it. That bothers me as an artist."*

The artist P1 states that the current music platforms disconnect the music from its context. He points out that a song is inseparable from its social context.

> *P1: "A song represents a universe of culture, people, social context, etc. [...]. Music—art—is a representation of people's sensitivity, it's a diary of people telling what happens."*

> *P1: "Listening to hip-hop from the 90's [is tied to] the slums of Los Angeles, of New York, and [to] what was happening at that time. That music comes from a place. It doesn't come out of nowhere."*

P1 thinks that current music platforms do not provide much context of the music and emphasizes that including such information would enhance the experience.

> *P1: "[...] there are songs that have their history, their function. So the more information there is—who are the people who made that song, with whom, how, where, why they sing it—, I know it would be much richer."*

Yet, P1 adds that some music may not convey any deeper message but exists for business reasons only.

> *P1: "[They are] made to sell more records."*

He puts the example of the genre Reggaeton, for which he thinks that frequently the explicit video is the selling argument and not the music. In such a case, adding information about the context would not add much to the user's experience.

3.3.2 Reaching an Audience

Another frequently mentioned issue is the difficulty to reach a larger audience, either because the artists are newcomers or when established artists enter a new music platform. This issue is usually known in recommender systems as the item cold-start problem. While it has become easier than ever before to access an enormous amount of music, the artists (P2, PN1, PF1, PF2, PR1) state that it is not easy to discover less popular artists with current music platforms; it requires the user actively looking for those artists when encountering them via other sources such as magazines or interviews. PF1 feels that it is very difficult to reach a larger audience.

> *PF1: "If [a track] is not listened to a number of times, it does not show up, and that means that it is not being recommended."*

In particular, it is not possible to get into a larger audience for artists that are new in the platform.

> *PF2: "[...] if [the music platform] does not recommend things that [...] have not been listened to much, then it enters a circle that never ends... it always goes... you will never be listened more. Then you're stuck there until someone pays for you to have promotion."*

P2 also participates in a less-popular band project. He affirms that music platforms make it difficult for the users to reach the band.

> *P2: "[...] it would not hurt if the systems were more random, not that obvious—if you like 'Beatles,' I recommend 'Rolling Stones."'*

PN1 underlines that the emergence of large music platforms changed the entire music industry, making it even more difficult for new artists to reach a larger audience with their music.

> *PN1: "Before you could go [to a label,] with something super weird but super interesting that could catch their attention and take you on a tour [...]. [But now] it's not the music that sells. [If] you don't have followers, you don't have content, you have nothing, you're nobody, and that's why you won't appear [in the recommendations]. You have to grow in a different way, through Instagram for example, which doesn't have anything to do [with music], and the value of music gets lost."*

Also, PR1 points out the difficulty of getting visibility on the music platforms if the artist is not popular. Similar to PN1, PR1 argues that it was also hard for an artist to reach visibility before the emergence of music platforms.

> *PR1: "[...] in this business, there have always been many traps. The musician wants to make music, but it's a business. At first you don't want to see it because you want to make music and you are happy. So for example, the old record companies used to have a monopoly before the [social] networks. You were only played on the radio if they paid for it. They said, '[This artist] has sold twenty thousand copies' but it was the record company itself that bought twenty thousand copies. [...] then everyone wanted to hear that [artist]."*

> *PR1: "On YouTube, when it started, there were companies that made it to reach the million [...] And I think that [happens] today too. You can buy visits [...]."*

P3 adds that artists being excluded, making it hard to reach an audience, has happened ever since, and he provides an anecdote of Bob Marley going to a radio station to force them to play his songs.

> *P3: "They had Bob Marley and they didn't play it!"*

P3 argues that the artists that really want to reach an audience will find a way to do it. Yet, probably not through a music platform but using another digital medium.

In conclusion, there is no clear consensus about the actions that music platforms should take in order to be fair in this context. Among the mentioned

alternatives, we see that some artists (e.g., P1, P2, P4, PN1) suggest that the music platforms should have a minimum quota of starting artists that are recommended to all the users alike. Others suggest that new artists should use alternative ways to reach an audience.

3.3.3 Transparency

Several artists (P3, PN1, P1, P2, PR1, PF2) mention that the music platforms should be more transparent. P3 states that he does not understand how exactly the music platform promotes some artists more than others; e.g., in the automatic playlists or the curated playlists.

> *P3: "It would be nice if the platform was equitable, or fair, for everyone in that sense, because if I'm one of the largest bands in Uruguay, why am I not in many of the playlists there? Is it the platform that doesn't want me there? Is it me doing something wrong? [...] Maybe the platform doesn't get much benefit with what we do, so they discard us."*

PF2 thinks that the music platforms should be more transparent towards the artists about how their recommendation system works and what the artists have to do for being recommended more often. PF2 thinks that this is particularly important for independent artists.

> *PF2: "[...] you are a bit naked there. You put your music on Spotify and mention in the concerts that they can listen [to your music], but you don't see any change. For example, no one explains you that it is important that other people add your songs to their playlists, so that the algorithm will recommend you."*

PJ1 feels that music platforms are profitable only for some of the stakeholders. The artist wonders what the goals of the music platforms are.

> *PJ1: "Is the goal for people to listen to music or is the goal to make money from it?"*

> *PJ1: "The people who invest in these things at the [...] powerful industry level, these people do make money from this. The others do not. [...] there is no middle class. There is an upper class and a lower class."*

Regarding transparency in algorithmic decisions, the artist PN1 mentions that although both humans and algorithms may be biased, a non-ideal decision made by a human may be easier to accept than one taken by an algorithm.

PN1: "[...] if a human makes that decision, if he says 'I have a small store, I am going to put it this way,' I will understand it much more than if an algorithm does it."

3.3.4 Influencing Users' Listening Behavior

One of our interests was to learn what the artists think about the music platforms' opportunities and power to influence the users' listening behavior. For example, with tailoring the music recommendations, a music platform could try to balance some styles that are not listened to among the recommended artists or also based on the gender of the artist. In an open question, which did not mention the gender, all artists came up with the issue that content by female artists is not well represented. We found that all artists agreed that the music platforms should promote content by female artists to reach a gender balance in what users consume.

While there was clear consensus to influence users' listening behavior with respect to the artists' gender—to reach a balance—, there was also a clear agreement *not* to do so with respect to the music style. For the latter, they think it is better not to influence the users. P2 suggests a gender balance in the recommendations.

P2: "[Platforms have] a huge responsibility in making recommendations."

About gender balance, PN1 states,

PN1: "[...] the population of the world is 50% women. So it would be ridiculous if the system wouldn't recommend them."

PN1 suggests a progressive change, which he thinks will prevent users to perceive it as something negative and leave the music platform. PF1 states that the system could enforce a 50% balance of male and female artists because there are many other factors different from gender, e.g. the music style, that define whether someone will like what is recommended. Finally, PF2 considers that using quotas alone would not be enough, as there is a need for a change in education to have a bigger impact. However, she agrees that artists could be better off with quotas.

PF2: "As a female artist I would like the system to recommend my music to someone that only listens to music of male artists."

3.3.5 Popularity Bias

We asked the artists how they relate to popularity in the recommendations. Researchers have put a lot of effort in reducing the popularity bias and improving the recommendations with respect to items in the long-tail of the popularity distribution (e.g., Vall et al. (2019b)). Reaching out to the artists, we wanted to explore their perspective on that issue; whether or not they feel that current systems give the users the opportunity to access items of less popular artists; and how this could be improved in the future.

P4 states that it may be easier for the music platforms to recommend what is popular because it can satisfy the majority of the users. Yet, P4 points out,

> *P4: "The problem is [that] the recommender system systematically ignores all those potential artists [...]"*

P4 thinks that some users may be happy with the recommendation of the generally popular items, whereas others may not and they may probably leave the music platform. The users who are more passionate about music will not see any advantage in using such recommendations.

> *P4: "[...] it's wrong that you don't have the option to explore that long-tail. [You can't] take advantage of a recommender system if you want to [explore the long-tail]."*

PR1 considers that the music platforms may generate higher revenues if they recommended popular artists, and may therefore not be interested in promoting less popular content.

Although strongly advocating the promotion of diverse content, PF2 speculates that this may again lead to have users listen to music that is widely popular. Therefore, she claims that the music platforms should prevent it,

> *PF2: "[...] otherwise you will always end up listening to American music."*

All the interviewed artists agree that it is crucial that the music platforms also recommend less popular music. They believe that the music platforms will harm the music culture if recommendations are limited to the most popular artists.

3.3.6 Artists' Repertoire Size

While popularity bias is a widely researched topic in the recommender systems community, and for music recommender systems in particular, little attention

is paid towards how the size of an artists' repertoire affects the probability of their songs being recommended to users. The artists' opinions are divided with respect to how a music platform should reflect the differences in the repertoire sizes. While three artists think that artists with larger repertoires should be more represented in recommendations (P1, PF1, PR1), four artists do not support that idea (P2, P4, P3, PF2), and two artists where indecisive (PN1, PJ1).

P1 argues that the higher number of records leading to an increased likelihood for an artists' items being recommended reflects what happens outside the music platforms.

> *P1: "[...] that's fine, the same thing happens in real life if someone makes 25 records, you will surely come across it at some point."*

PJ1, P2, and P4, in contrast, argue that having more records should not be a reason for being recommended more often.

> *P2: "[...] I know so many amazing bands with only one album, it has 10 songs. And you will never get to those [being recommended]."*

> *PJ1: "This is delicate because there are big artists that have one album, or the opposite."*

> *P4: "Intuitively, I think it is unfair that an artist with more music is recommended more. [...] if an artist has 30 albums but they are all completely different to the previous ones, then it make sense. But if there is an artist whose albums are all the same ...[it does not]."*

PR1 points out that artists with more songs are probably more diverse in their repertoire. So, it is more about the diversity than the size of the repertoire.

> *PR1: "If you have more songs you have more chances to satisfy different audiences."*

P3 states that an artist profile with more songs may leave the impression that the artist is in the music business for a long time, which may be a reason to recommend the artist more. But he adds that repertoire size should not be given such an importance. PF2 adds that sometimes not all tracks of an artist are available on a particular music platform, so the repertoire size on the music platform would not reflect reality.

PN1 raises the issue that it may depend on the users' goals whether the artists' repertoire size matters. If a user wants to explore a new artist, then it is not

useful to recommend to this user an artist with only a few songs. For users exploring on a track-basis, the artists' repertoire size is irrelevant.

3.3.7 Quotas for Local Music

Today's most prominent music platforms operate in several countries, more or less globally. With the widely adopted CF approaches for recommendations ("people who like that..., also like that..."), it may happen that the music preferences of users in countries with a large number of users may influence the algorithms' outputs globally. As a result, artists that are popular in countries with a smaller user number will have less chances to be recommended than artists that are popular in countries with a large number of users.

While there are studies investigating the existence of local trends on global music streaming (e.g., Way et al. (2020)) and recommendation approaches that account for country-specific music preferences (e.g., Bauer & Schedl (2019)), it is not clear whether and how current music platforms consider local repertoire. Similarly, outside the music platforms, some countries define quotas of local content for radio stations while other countries do not. Yet, such quotas for radio do not apply for online music platforms, specifically not for automatic recommendations.

We asked the artists about quotas, the desirability and applicability to have quotas on online music platforms and for automatic recommendations in particular. We also asked for potential alternative solutions to deal with local content. Overall, three artists mention that the recommenders should have quotas for local music. Other five artists were not sure whether quotas were the right solution but emphasized that it is important that the music platforms promote local content.

PR1 agrees with quotas and indicates that

> *PR1: "otherwise you won't know what there is in your country."*

PJ1 is ambivalent. On the one hand, he believes that it is not right to base such decisions on where a person is from, on the other hand, he sees the need to promote local artists because they are at a disadvantage to begin with; and with quotas they could make up for it over time.

P1 suggests to abandon the idea of defining locality in terms of countries because national borders are not necessarily cultural borders. He proposes to define locality within a radius of a user: Artists that are in a certain radius (e.g., within a radius of 5000 km) should be given a higher weight than artists outside that radius; and within the radius, different weights again, with higher weights for the closest artists. Yet, the same artist (P1) emphasizes that quotas

are a necessary measure in some countries because otherwise local artists would not be able to make a living solely from music. Accordingly, he suggests that quotas should also apply to automatic recommendations and proposes to use a combination of country and radius.

> *P1: "[...] if you go to the border [between Uruguay and Brazil] [...], the Brazilian influence is greater than the Uruguayan one. So it seems to me that the radius is more representative for culture."*

PN1 considers peculiar that there is a higher chance for the user to be presented US artists compared to local ones, even if the latter are locally famous and popular. PN1 calls for more transparency and draws an analogy to the news sector.

> *PN1: "[...] it is like reading the news in the New York Times instead of the local newspaper. [...] you know that you're reading the New York Times or the local news. But you don't know that, if it is an algorithm that makes the recommendation."*

PF2 and P4 argue against quotas because this could cause users to leave the music platform if they do not like local music. Yet, both emphasize to give importance to locality. P4 suggests to give individual users the chance to choose the degree of locality they want to have. PF2 proposes to promote local content by letting users indicate what countries they would like to receive recommendations from besides their own country. Artists could also be allowed to indicate in which countries they would like their music to be recommended. This would enable artists to reach other countries.

> *PF2: "As an artist you could reach more countries if you are interested."*

P2 is unsure whether quotas are the ideal measure, but emphasizes that the music platforms have responsibility for what their algorithms recommend.

> *P2: "[A platform provider] needs to take responsibility for its recommender system—[...] understand the situation. [...] Obviously, they are commercially not capable or not interested [...] but it would be great [...] if [they] find a way to link [...] a Yankee band with a Uruguayan band [...] make a connection that contributes. [...] [The platforms] have their share of responsibility for what they are showing or recommending. I don't know if there should be quotas [...] but [...] it would be great if there was something."*

P3 questions whether quotas should be enforced by law and suggests that music platforms take the responsibility for it.

> *P3: "The question is whether it should indeed be state-imposed. For things to be that way, do we have to impose it?"*

> *P3: "[the platform] should do what is ethically correct. [...] If I were Spotify [...] I would [promote local content] in every country."*

While P3 voices concerns whether the music platforms should be trusted in deciding what to promote more or what to promote less. Yet, for gender fairness and local content, P3 is confident that the music platforms could find the right balance.

Different to the common understanding of local quotas, PF1 suggests that it should be the opposite: Instead of having quotas for local music in smaller markets, there should be quotas in larger markets to include music from those smaller markets. In addition, PF1 points out that giving users the possibility to explore the country-specific music scene would be more beneficial for the artists.

> *PF1: "[...] provide the possibility to listen to what you have not listened to before. For example, 'what have I not listened to from Colombia?' 'What is underground in such a country?' If you give the local artists a voice and let them tell the story behind their music, that would be more interesting."*

3.3.8 New Music

In this section, we describe the artists' reflection on two related concepts that we group under the term 'new music'. New music may refer to (i) artists that are new to a user (thus, discovery of artists) and also to (ii) a new track or album released by an artist that had already been part of the music platform.

Most of the interviewed artists agree that the artists should be in control of what tracks or albums get more recommended. In the case that they are not in control themselves, they strongly prefer a recommender system that puts more weight on their latest releases. PR1 states that every artist wants their new music to be promoted so that the world finds out they have a new release.

> *PR1: "[...] you do a promotion campaign to tell that you released more [content]. To tell the world, 'Hey! There is a new album!' [...] Like saying, 'Hello, I'm here."'*

Also, some artists feel more identified with what they are doing now compared to music they released many years ago, which is another reason for them to prefer the promotion of the latest release.

> *P3: "[...] it is something that I have done 10 years ago. [...] I don't know if I feel identified [with it]."*

With respect to allowing users to discover artists that they are not yet familiar with, there is no clear consent either. While all agree—in varying degrees, though—that the music platforms should allow users to discover artists that they do not know, it remains unclear how the music platforms should do that. Most of the interviewed artists state that the user should be in control, having the opportunity to indicate that they want to discover new artists, and when they want to do so.

In general the artist P2 think that the platform should show music to the user that he/she doesn't know (discovery). About that the artist says

> *P2: "[...] being able to choose would be good. A button that says 'I'm open to new stuff' or 'Let me listen to what I want.' Because sometimes you want new stuff and sometimes you want something very specific."*

In particular, the artist says that as a user he is limited to discover new music from some places

> *P2: "I do not have a way to say I want to listen to indie music from Japan or reggae from Italy, that is missing"."*

3.4 Conclusions

We reached out to music artists to understand how they are affected by current music platforms and what improvements they deem necessary with respect to those music platforms being considered *fair* from the artists' perspective. Thereby, we paid particular attention to music recommender systems that are an integral part of today's music platforms. We conclude that the participating artists' perceptions and ideas are well aligned. The concrete aspects that could be identified in which the music platforms could be more beneficial for the artists are summarized in Table 3.5.

After identifying some aspects of the music platforms that are considered relevant by the artists to be fair we derive the following aspects: popularity bias, size of the artists' repertoire, quotas for local music, and dealing with new music.

We found a strong alignment among the artists in order to make the music platforms fairer regarding the following aspects: 1) The artists call for better promotion of local music; 2) they also agree that gender balance in the recommendations is expected; 3) the artists voice that music items in the long

Table 3.5: Aspects to improve with agreement of most of the artists.

Topic	Description
Quotas for local music	Music platforms should promote more local music.
Gender balance	Artists mention the clear expectation of gender balance in the recommendations.
Popularity bias	Artists agree on the importance of the recommendation of items in the long tail, not only the most popular artists.
Lack of control	Artists mention the need for more control concerning the tracks that are promoted. Otherwise, if they are not in control, then they generally favor the promotion of their latest releases.
Transparency	Artists agree on the importance of transparency about how the algorithms work, to understand why their music is recommended or not.
Influencing users' taste	The system should not influence the user's taste.
Music in context	Artists would appreciate if their music would be presented to users with information about its context.

tail of the popularity distribution (not only the most popular artists) have to be included in the recommendations shown to users; 4) the participants state that is important to give more control to the artists concerning the tracks that are more promoted or higher weighted in recommendations (if they are not directly in control, then they generally favor the promotion of their latest releases); 5) they request transparency about how the algorithms work, to understand why their music is recommended or not; 6) the artists consider a system that influences a user's taste (or attempts to do so) an undesired misuse; and 7) they would appreciate if the music platforms would be enriched with information that puts the music into context.

Besides the consensus on many topics, our findings indicate that for some topics there is no clear direction: 1) There is no agreement with respect to quotas as a potential solution for local content; 2) no clear agreement about the size of the artists' repertoire should be reflected in recommendations; 3) while the artists seem to agree that new artists should be given space on the music platforms, there is no agreement on how to operationalize this; and there is 4) no clear agreement whether a music platform should promote the discovery of artists previously unknown by a user but most of them agree that it should be in control of the user when to listen to new music and 5) there is no clear opinion if the music platforms should promote more diversity in the recommendations.

Overall, while there is a prevailing belief that online music platforms would give access to items in the long-tail to users and encourage them to consume more of those items, the interviews suggest that the long-tail items and artists remain obscure and do not appear in recommendations. The artists report that it is difficult for their long-tail items to be discovered by users or appear

in recommendations computed by music recommender systems. They indicate that it is difficult for users to discover or reach less popular artists on the music platforms unless the users explicitly search for those artists. The artists express their concerns with respect to reaching a larger audience, which they do not consider fair.

The advances in the MIR community had been fundamental to improve the content-based approaches for recommendations that allow promoting new and less popular artists. From the interviews, we understand that artists see it highly important to use more these techniques. Thus, further advances can have a direct impact on society.

Based on these findings, in the following chapters, we address some of the issues mentioned by the artists related to recommender systems such as the gender imbalance, reducing the influence on the users' musical taste and more promotion of less popular content through improved recommendations techniques.

CHAPTER 4

Gender Imbalance in Music Recommendations

4.1 Introduction

The goal of a recommender system is to predict which items a user might like given the user's previous ratings or interactions with items. This may lead to situations where users only see a narrow subset of the entire range of available recommendations (Sun et al., 2019), a phenomenon known as the 'filter bubble' (Pariser, 2011). Users will respond to those recommendations, which will then be used as input for future recommendations; with this feedback loop, the recommender system will learn to recommend increasingly similar items (Sun et al., 2019; Chaney et al., 2018; D'Amour et al., 2020; Ferraro et al., 2020c). However, the strategy of maximizing the users' experience may have negative effects on other humans involved in and affected by recommender systems: the item providers (Bauer, 2019).

In the case of music platforms the recommender systems strongly influence what people listen to; and that, in turn, defines what is going to be the next hit song, how much exposure an artist gets, or a music genre at large. The music recommender systems may privilege the content of a small group of artists when maximizing user satisfaction. As a consequence, there are groups of artists that the system recommends less, unfairly reducing the exposure of that group, and limiting some artists' chances to reach a larger audience particularly due to the feedback loop.

In Chapter 3 we reached out to artists to understand their perspective on *fair recommendations*. From the interviews presented in Chapter 3, we understand that one of the main problems artists see in the music streaming platforms is the gender imbalance. In fact, gender bias in the music business is a relevant research track and imbalances were repeatedly evidenced (Schmutz & Faupel,

2010; Smith et al., 2018; Wang & Horvát, 2019). Also, concerns about bias and discrimination are repeatedly voiced in the media (e.g., Youngs (2019); de Revere (2015); Mitchum & Garcia-Olano (2018)).

Yet, not many previous works analyze the recommender systems from the artists' gender point of view. Therefore, in this chapter we address this problem in two steps: First, we analyze a widely-used collaborative filtering approach concerning the artists' gender. Second, based on the insights from the artists' interviews and the results of the first step of the analysis, we propose a progressive re-ranking method to achieve gender balance. For the evaluation of the proposed method we rely on a simulation of feedback loops to provide an in-depth analysis of the longitudinal effects considering state-of-the-art performance measures, and metrics concerning gender fairness.

This work distinguishes from previous work in these aspects: (i) We incorporate the opinion of those people concerned (i.e., artists) in finding a solution for gender bias on music platforms. (ii) We propose a way to gradually mitigate this bias following the insights from the interviews with artists. (iii) We use a simulation approach considering feedback loops to understand the longitudinal effects.[4]

4.2 Insights From Interviews

In this section, we give a summary of the conducted interviews to music artists and we make a more in-deep analysis of the artists' opinions on what regards gender representation in streaming platforms.

4.2.1 Summary About Interviews Participants, Material, and Focus

As described in Chapter 3, we conducted semi-structured interviews with a first set of 11 guiding questions on 9 artists that we consider diverse in the kind of music they perform (including folk, pop, punk-rock, dubstep, jazz, flamenco, progressive rock, hip-hop, and reggae), their popularity (2 globally known, 4 known within their country, 3 regionally popular), experience (4 to 25 years active in the music industry; 1 to 10 records released), age (26 to 55 years), and gender (2 female, 7 male). According to research practice (Creswell & Poth, 2016; Morse, 1994), we reached a high level of thematic saturation (Guest et al., 2006) with the same topics being repeatedly mentioned across the interviews.

[4]The findings described in the following sections are based on our published work presented in Ferraro et al. (2021c)

These interviews were designed to last one hour each, whereof we used the first 10 minutes to explain the purpose of the project and to give a general introduction to recommender systems and music platforms. We asked open questions with the goal that the artists contribute their own ideas, that might not been considered in the field before. In addition, we proposed some specific alternatives to learn the artists' opinions on those.

Prior to the interviews, the participants were informed that the data and results of the interviews were going to be kept anonymous at all times, which was important so that the artists could feel free to share any concern regarding streaming platforms or other issues related to the music industry. In addition to the consent form, we asked the participants to fill out a short form with some optional information about themselves that would be used to refer to their answers (i.e., age, gender, country, genre, popularity level), number of records or singles, years active in the music industry and contracts with labels.

The recordings of the interviews amount to 420 minutes with a total number of 33,669 words in the transcriptions. Following the methodology of Qualitative Content Analysis (Mayring, 2004), we developed an annotation (i.e., coding) scheme inductively from raw data (i.e., the transcriptions), defined the topics and annotated the transcriptions accordingly.

In the interviews, the artists addressed a wide variety of topics (e.g., lack of control, context of music, transparency), with gender fairness representing one major concern.

In this Chapter, we focus on the parts of the interviews related to the topic of *gender balance*. In particular, we asked the artists whether and how music streaming platforms should intervene and influence users concerning the music that they consume. Only in a follow-up questions, we asked for the concrete scenarios, whether and how a platform should attempt to intervene and pointed with respect to genres that are not widely consumed and with respect to the representation of the gender of artists.

4.2.2 Interview Results

Overall, we found that the artists agree that the music platforms should try to promote more content of female artists to reach a gender balance in the music that users listen to. This stands in sheer contrast to the artists' opinion on influencing users concerning music style or artists' location. In general, the interviewed artists had a strong tendency against influencing users concerning music style. One participant argues,

> *"In the case that we talked about before, about recommending different types of music, I don't see why we should tell the users which genres they*

should listen to."

In contrast to this, there was a clear agreement among all participants—even though the sample is male-dominated—that it is important to promote more content of female artists to reach gender fairness. One participant argues,

> *"This is in contradiction with what I said before. Before I thought that we can't say to the user what to listen to but this doesn't help to solve some problems that are in the system where the recommendation has an influence. I think there should be applied some actions to correct some biases, the question is in which cases it should be corrected and in which not. In Heavy metal music I imagine that there are not many female singers, maybe we could give them more visibility, otherwise they would never be seen. It could start slowly, not 50/50 from the beginning. In this case I think it is reasonable[...]."*

Another artist thinks that the recommendations should be divided into different spaces: the 50% should be 'spontaneous' where is defined by what the user listened to, then in the other 50% you could try to provoke a change depending on the goal. But the user should be aware of this, about the space (which 50%) that it is using. The artist says:

> *"It would be nice that the platform was transparent about how it works and then you could be more conscious about what you are listening to: 'I'm recommending this because these are your neighbors that play music in your town'."*

The artist says that the platforms should do good to people.

> *"It's a cultural good, it connects people with experiences, with their lives, with their emotions. The platforms should contribute in a way that makes the experience more relevant. There is a connection between the links of communities, about how people interact with music. [...] We are in an age where if it works it is not enough, we need something with more sense, more values."*

Another artist suggests that there should be a gender balance in the recommendations. He states,

> *"[Platforms have] a huge responsibility in [gender balance] when they make recommendations."*

He suggests that the platforms should be more transparent about the recommendations that they are giving to the user and they should also allow the users to control what proportions in gender they want to receive.

A newcomer artist thinks that the platform should not try to influence what people listen to in terms of the genres or styles that people listen to. However, he voices the need to do so concerning gender balance:

> *"[...] the population of the world is 50% women. So it would be ridiculous if the system wouldn't recommend them."*

This artist argues that the change should be made progressively towards gender balance,

> *"[...] otherwise the users could perceive it as something bad and leave the platform."*

Another female artist proposes to enforce a 50% gender balance, because many factors other than gender (e.g., the music style) define whether a user will like a recommendation. Another artist argues for 50% of female music, and likewise 20–30% of local artists, and suggests to consider also proportions for other minorities (e.g., ethnicity, sexual orientation). However, he questions how to define the proportions:

> *"If we consider all the minorities then, what amount of the music is black music? And how much Caucasian? How much sexual diversity? [...] In summary, I do believe that it should be [imposed to users]."*

The artist thinks that while the platforms could influence the distributions in the recommendations in a positive way and use it as a means to demonstrate that they act ethically, it is out of control what they indeed do. Yet, the artist recognizes that this thought could be unnecessarily pessimistic. The artist summarizes it as

> *"I think that it should be that way [there should be quotas] but I don't trust the ethical relationship of the platform with the user."*

A female artist considers that quotas in recommendations would not be sufficient. She emphasizes that a change in education would be necessary to have a bigger impact than quotas could. However, she agrees that this could be one step forward. As a female artist, she would like to have a system that recommends music of female artists to someone that previously only listened to music of male artists.

A jazz artist argues that every recommendation influences a user in one way or another. Therefore, the artist advocates for recommending more balanced content:

> *"It is impossible to be impartial; so it is better to do it as equally as possible."*

A rap artist explains that the system should not attempt to change the users' musical taste but agrees that the system could create some balance in gender.

In summary, the results suggest that the interviewed artists are concerned about gender fairness. They would like to see a balanced gender representation in the recommendations. The artists voiced that the recommendations could be used as a means to change the consumers' listening behavior by promoting more content of female artists and suggest gender balance in the recommendations (i.e., *positive disparate treatment*). At the same time, they emphasize the need to increase this proportion only gradually until gender balance is reached to avoid the users' negative reaction towards the change.

4.3 Quantitative Approach

We build on the interview results (Section 4.2) with a two-part quantitative analysis. In the first part, we evaluate a recommendation algorithm that is widely used in the music domain with respect to gender fairness. For this evaluation, we device two large real-word datasets of music listening events (Section 4.3.1). The goal of this analysis is to understand (i) how the datasets are distributed in terms of the artists' gender and (ii) how the algorithm performs for those distributions with respect to gender fairness.

As collaborative filtering is a common approach for recommendations in the music domain, we choose a state-of-the-art Implicit Matrix Factorization optimized with ALS (Hu et al., 2008) for our analysis. As the number of tracks per artist may vary per gender both in the dataset as well as in the recommendations, we evaluate—where possible—for both, recommendations on the artist level (*LFM-360K* dataset and *LFM-1b* dataset) as well as recommendations on the track level (*LFM-1b* dataset). In addition, we compare *ALS* with two baselines, one that generates random recommendations (*RND*) and one that recommends the same most popular items to all the users (*POP*).

As we want to understand the impact that the recommendations can have on users in the longer term, we use a simulation to mimic feedback loops that allows us to study how the recommendations can affect user's behavior, following the same approach used in previous works (Ferraro et al., 2019b; Jannach

et al., 2015b; Zhang et al., 2020a). For each user, we first generate recommendations and then we assume that the top-10 items are listened to by the user. We correspondingly extend the dataset with these new interactions and we re-train the underlying model. After the retraining steps, we generate new recommendations for the next iteration, repeatedly for a total of 20 iterations. In the second part of the quantitative analysis, we show that by employing a simple re-ranking mechanism, we can break the feedback loop and gradually increase the exposure of female artists. By controlling the parameters of the re-ranking mechanism, the impact in the users behavior can be more or less gradual. We provide an in-depth analysis, using a set of metrics as described in Section 4.3.2, and we compare the re-ranking mechanism to the baseline without re-ranking.

4.3.1 Datasets

We use two public datasets obtained from Last.fm. *LFM-1b* (Schedl, 2016), a large dataset of more than one billion listening events containing playcounts with timestamp by 120K users covering 32M tracks by 3M artists. The second dataset is *LFM-360k* (Celma, 2010), which contains 17M interactions between users and artists (359K users and 260K artists). In the absence of a dataset on music interactions containing gender information of artists, we had to create our own ones by enriching existing datasets. We extended the datasets with gender information of the artists collected from MusicBrainz.org [5] (*MB*). For this, we first query the *Last.fm API*[6] to get the *MB* identifier of the artist. For complexity reduction, we focus on 'solo' artists—thus, where the artist is an individual person—and we consider those artists for which *MB* reports the gender (in *MB*: female or male). While we are aware that this binary gender classification is inapt to reflect the multitude of gender identities (Spiel et al., 2019), to the best of our knowledge, there is no dataset that goes beyond this binary gender classification. For *LFM-1b*, we collected the gender of 64,745 artists, whereof 15,055 are classified female and 49,690 are male. For *LFM-360k*, we collected gender information for 46,469 artists, whereof 10,535 are female and 35,922 are male (Ferraro et al., 2020a). Note, the gender imbalance in the datasets reflect the current reality in the music business (Youngs, 2019; Epps-Darling et al., 2020).

We consider only users and tracks or artists, respectively, with more than 30 interactions to have sufficient data for training and evaluation. Thus, we remain with the following data: For *LFM-1b*, we have 112,291 users and 465,064 tracks by 33,325 artists, and for *LFM-360k*, we have 220,444 users and 12,900

[5] http://musicbrainz.org

[6] http://ws.audioscrobbler.com/2.0/?method=artist.getinfo

artists. For both datasets, we split in train and test set by randomly selecting for each user 80% of the items for training and 20% for test.

4.3.2 Metrics

Following the literature of multi-stakeholder recommendation (Abdollahpouri et al., 2020a), we apply metrics that allow us to gain a better understanding of the system's behavior in different situations, where we focus on the artists' perspective. In particular, we apply multiple metrics to understand the system's behavior from different perspectives. We use metrics that assess the probability of female artists being recommended. We particularly focus on the position in the recommendation rankings of content by female and male artists because users typically interact more frequently with only the top-ranked items (i.e., position bias Collins et al. (2018)). To this end, we average for each user the *position of the first occurrence of content by a female* (with the highest rank on position 0) in the recommendation ranking and the *percentage of content by females* in the recommendations. We use *Hellinger distance* following the approach in Abdollahpouri et al. (2019, 2020b) to measure the similarity of the gender distribution in the recommendations compared to the users' original listening behavior. In addition, we use the *Gini index* which measures how concentrated the recommendations are on fewer artists. A Gini value of 1 would indicate that all the recommendations are the same for all the users, whereas a value of 0 means that they are all different. With *Coverage*, we measure the number of different artists (or tracks) globally recommended (differentiated by gender).

We use precision and *nDCG* (Ricci et al., 2010) to measure the accuracy of the algorithms. We report precision for all recommendations and also separately by gender. Given a track (t) and a user (u), $hit@K(t,u)$ returns 1 only if t is recommended in the top-K to the user u and is in the test set for that user. We follow these steps: 1) Generate ranked recommendations of tracks for user u, referred to as $A = \{t_1, t_2, ..., t_n\}$); 2) divide items in A that are by the female artists ($F = \{f_1, f_2, ..., f_i\}$) and items by male artists ($M = \{m_1, m_2, ..., m_j\}$), 3) for each user, the precision is computed as:

$$P@k = \frac{1}{|K|} \sum_{t \in T} hit@K(t,u)$$

where the group of items J corresponds to: A when we compute $P@K_{all}$, F when we compute $P@K_{female}$ and M when we compute $P@K_{male}$. Thus, $P@K_{female}$ and $P@K_{male}$ add up to $P@K_{all}$. We evaluate recommendations for multiple K (i.e., 1, 3, 10, 100).

4.4 Gender in Music Recommendation

In this section, we report our analysis of the performance of the algorithm *ALS* with respect to gender fairness and compare it against the two baselines (*POP* and *RND*). To this end, we considering the top-100 recommendations from the gender perspective. We analyze the recommendations on the artist level using both datasets—*LFM-1b* and *LFM-360k*—(Section 4.4.1), and on the track level using the *LFM-1b* dataset (Section 4.4.2). Finally, we present the results of the simulation of artist recommendations using the *LFM-1b* dataset (Section 4.4.3). Note that for all analyses we run the experiments three times and see stable results in all cases.[7]

4.4.1 Gender Fairness on the Artist Level

Table 4.1 summarizes the results of the analysis of the recommendations on the artist level, considering the top-10 artists recommended by the algorithms.

Comparing the average position of the first female and the first male artist shows that, using *ALS* for *LFM-1b*, the first female artist is recommended on average on the position 6.7717, whereas the first male artist is on average on position 0.6142. Using *ALS* on the *LFM-360k* dataset, the results show an even worse presence of female artists in the recommendations, where the average first position of a female artist is 8.3165 whereas for a male artist it is 0.7136. Compared to the baselines, *ALS* delivers the largest gender gap concerning the average first position of a female artist.

Using *ALS* and *LFM-1b*, 25.44% of the recommendations are female artists, which is close to what is reflected in the users' previous listening behavior, the users listened to 25.26%. This indicates that the algorithm has statistical parity in this aspect. For both datasets, the Hellinger distance suggests that the recommendations computed via *ALS* are the closest to the gender distribution as reflected in the users' previous listening behavior. Whereas the gender distribution in the recommendations generated via *RND* and *POP* deviates more from the original distribution.

The last three columns of Table 4.1 show the performance of the analyzed algorithms. For each of the datasets, the parameters of *ALS* were optimized to provide a higher precision. Consequently, we used a 300-dimensional space in *LFM-1b*, and a 200-dimensional space in *LFM-360k*. In both cases, the results clearly suggest that although *POP* gives more balanced recommendation in terms of gender fairness, the performance with respect to precision and nDCG are below *ALS*.

[7] https://github.com/andrebola/gender-recs

An additional analysis for coverage (considering the top-10 recommendations for each user using *LFM-1b*) shows a far lower coverage using *POP* compared to *ALS*. Using *POP*, the total number of artist appearing in the top-100 recommendations are 336, whereas for *ALS* the total number is 15,194. Likely, the low coverage using *POP* is not in the interest of the overall artist population. As expected, *RND* gives the highest value of coverage.

4.4.2 Gender Fairness on the Track Level

We conducted the same experiment but on the track level instead of the artist level. Table 4.2 summarizes the results on the track level considering the top-100 recommendations of the algorithms, using the *LFM-1b* dataset. On the track level, *ALS* shows a large gender gap in the average first position (4.6993 vs. 24.9162 for male vs. female artists, respectively); by far larger than on the artist the level. Using *RND* provides a similar picture as on the artist level, and using *POP* results in similar positions for male and female artists when analyzed on the track level.

From Table 4.2, we can see that *ALS* provides slightly more recommendations of content by female artists (28.99%) than what the users listened to before (25.33%). However, *ALS*'s percentage of content by female artists in its recommendation is closest to the users' original listening behavior compared to the baselines. *POP* delivers a far higher percentage of content by female artists (66.66%) compared to other approaches. *RND* provides a lower percentage of female artists (21.72%).

The Hellinger distance indicates that —also on a track level— the recommendations generated via *ALS* delivers recommendations that are the closest to the gender distribution as reflected in the users' previous listening behavior. Also *RND* comes close to the original distribution, while *POP* does not.

While the last three columns of Table 4.2 present the performance metrics for all three algorithms for all artists, we show these metrics differentiated by the artists' gender in Table 4.3.

For computing the accuracy as presented in Table 4.3, we take the recommendations for each user at the given cutoff (i.e. top-1, top-3, top-5, and top-10 recommendations). The results suggest for all precision metrics as well as for the ranking quality (nDCG) that lower performance is achieved for recommended female artists than for male artists when using the *ALS* algorithm. Using *POP* flips those results.

The last column of Table 4.3 shows the number of unique items that are recommended—for each algorithm, and differentiated by the artists' gender. An additional analysis shows that the recommendations generated with *POP* cover the limited number of 130 different tracks by female artists, compared

Table 4.1: Results for artist recommendation (both datasets).

	Algo	Avg position		% females	Hellinger	Precision		nDCG
		1st female	1st male	rec.	distance	P@1	P@10	@10
LFM-1b	ALS	6.7717	0.6142	25.44	0.0988	0.4505	0.2997	0.3409
	POP	0.1325	1.7299	32.44	0.1577	0.1033	0.0919	0.1118
	RND	3.3015	0.3046	23.30	0.1346	0.0010	0.0010	0.0019
LFM-360k	ALS	8.3165	0.7136	26.27	0.2102	0.1781	0.0863	0.2804
	POP	0.9191	0.2713	29.31	0.2670	0.0247	0.0205	0.0978
	RND	3.3973	0.2951	22.77	0.2597	0.0003	0.0003	0.0025

Table 4.2: Results of track recommendation (*LFM-1b*).

Algo	Avg position		% females rec.	Hellinger distance	Precision		nDCG @100
	1st female	1st male			P@1	P@10	
ALS	24.9162	4.6993	28.99	0.1374	0.4730	0.3237	0.2392
POP	0.8726	0.8239	66.66	0.3404	0.0509	0.0310	0.0239
RND	3.6422	0.2819	21.72	0.1507	0.0002	0.0002	0.0002

Table 4.3: Performance of track recommendation (*LFM-1b*).

Algo	P@1		P@3		P@5		P@10		nDCG@100		Coverage@100	
	female	male	female	male	female	male	female	male	female	male	female	male
ALS	0.1696	0.3176	0.1505	0.2810	0.1360	0.2567	0.1140	0.2193	0.1322	0.1803	18,825	52,513
POP	0.0200	0.0329	0.0377	0.0113	0.0347	0.0075	0.0261	0.0073	0.0317	0.0092	130	102
RND	0.0001	0.0002	0.0001	0.0002	0.0001	0.0002	0.0001	0.0002	0.0001	0.0002	100,722	362,887

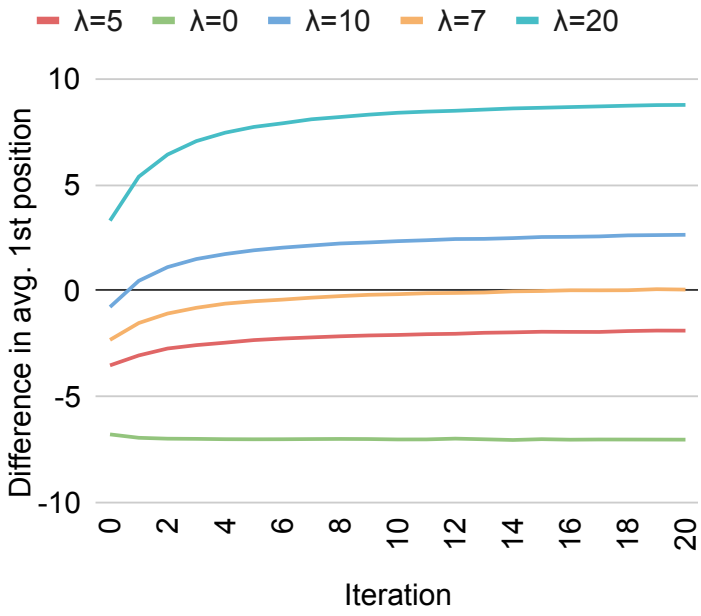

Figure 4.1: Average difference between first position of female and male artist.

to a coverage of 18,825 tracks by female artists with *ALS* and 100,722 with *RND*.

4.4.3 Simulating Feedback Loops

We propose an ad-hoc approach to improve the exposure of female artists by penalizing male artists by moving them λ positions in the ranking. We study the impact of different values for λ on the exposure of female artists in the long term. Thereby, $\lambda = 0$ represents the baseline *ALS* without re-ranking. To this end, we use recommendation on the artist level and simulate the interaction of users with the top-10 recommendations for each iteration. We visualize two different aspects of exposure: First, Figure 4.1 shows the difference between the average first position of female and male artists for each iteration. Increasing λ gives a more balanced exposure to female artists compared to the baseline without re-ranking ($\lambda = 0$). Depending on how fast the change is desired, different values of λ may be preferred. Using $\lambda = 7$ seems to achieve a good balance in the long-term, which is aligned with the idea expressed by the artists of progressively inducing a change in the behavior.

Second, Figure 4.2 shows the evolution of the average percentage of female artists across the iterations for the different values of λ, and compares those to the consumers' listening behavior according to the *LFM-1b* dataset. Compared to the users' current listening behavior, using $\lambda = 7$, the percentage increases

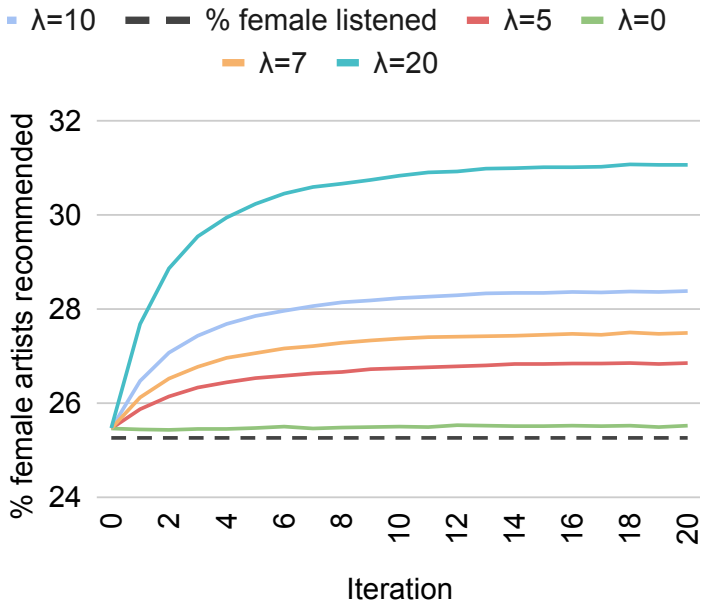

Figure 4.2: Percentage of female artists recommended and listened to by the users.

by almost 2 percentage points, whereas with $\lambda = 20$, it increases by more than 6 percentage points. Considering both views on gender fairness (Figures 4.1 and 4.2) it provides a good basis to decide on a λ value. Using $\lambda = 7$ achieves a good balance in the long-term, which is aligned with the idea expressed by the artists (see Section 4.2.2) of progressively inducing a change in the behavior to a balanced exposure of female and male artists. An even higher exposure of female artists could be achieved with $\lambda > 7$.

To investigate the potential performance loss when increasing the exposure of female artists, we compare the prediction accuracy achieved with the baseline *ALS* without re-ranking (i.e., $\lambda = 0$) and those achieved with different values for λ for each iteration. Our analysis suggests that, in comparison to the baseline ($\lambda = 0$), the nDCG@10 is, on average, reduced by 2.2% for $\lambda = 5$, 4.9% for $\lambda = 7$, 6.7% for $\lambda = 10$, and 15.0% for $\lambda = 20$.

In addition, we analyzed the intervention of the re-ranking by looking at the average number of items that are re-ranked for each user in each iteration. Figure 4.3 shows for each λ value, the evolution of the number of items where the rank is modified. Considering either only the users that have any modification (*mod*; e.g., λ7-*mod*) and all users (*all*; e.g., λ10-*all*). Results suggest that the number of re-ranked items decreases with increasing iterations. In short, *ALS* starts recommending more females over time compared to the initial recommendations, and the effect of the feedback loop decreases once the users start changing their behavior.

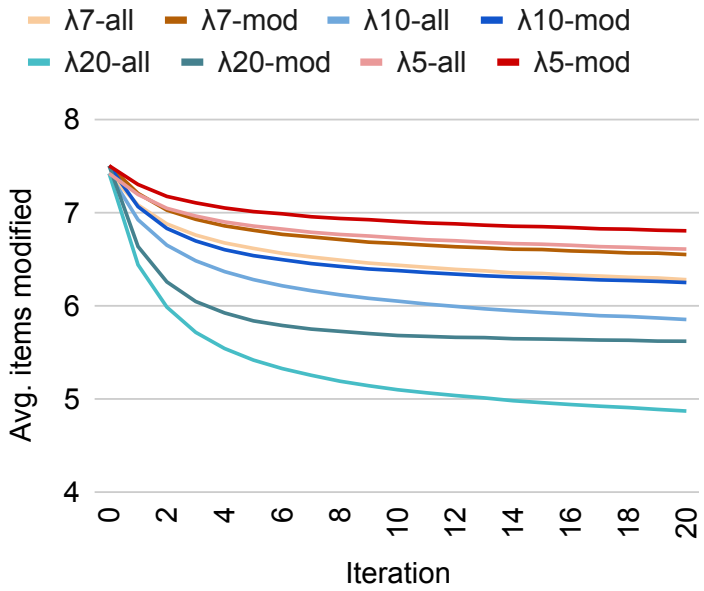

Figure 4.3: Average number of items modified for each user using different values of λ.

4.5 Conclusion

In this chapter, we have considered how music recommender systems affect artists from a gender point of view. We zoom on the conducted interviews with music artists, to understand their attitudes and beliefs regarding the current music platforms and the embedded recommender systems regarding artists' gender.

The results suggest that interviewed artists are concerned about gender fairness on current music platforms and it is in the best interest of many artists to give more balanced recommendations in terms of the artists' gender. Motivated by this finding, we investigated how the effects of the bias is affecting a collaborative filtering recommender approach on gender fairness in an offline setting.

Even though the Collaborative Filtering algorithm that was used leads to a similar representation of content by female artists in the recommendations compared to what consumers usually listen to, results suggest that there is a considerable difference with respect to the average first position of female and male artists in the recommendation ranking. In short, the exposure of content by female and male artists is not well balanced which means that the exposure is not even. Moreover, we followed the interviewed artists' expressed request to gradually give more exposure to female artists and we propose a simple re-ranking approach.

By simulating the user interactions with the recommendations we show that gender can be better balanced in the long term when gradually increasing the exposure of female artists in the recommendations. This balance is achieved without severely affecting performance.

By analyzing how recommender systems perform for different genders, we can better understand the systems' performance from the artists' perspective. Giving a more balanced recommendation could be misaligned with some short-term oriented business values (Jannach & Jugovac, 2019) since it may not meet the consumers' expectations with regard to what they are used to be served with on platforms or the world market. Yet, we also have to consider the benefits that these platforms contribute to society in a broader sense and we have to consider what impact they have on how people consume music. Platforms have the *responsibility* to act ethically (Milano et al., 2020).

CHAPTER 5

Maximizing Users' Engagement With Artists

5.1 Introduction

Recommender systems are a fundamental part of music streaming platforms, allowing users to explore the platform's music collections. Recent studies in the field of Recommender Systems (Abdollahpouri et al., 2020a; Ferraro et al., 2019b; Ferraro, 2019) show the importance of taking into account the interests of all the stakeholders involved when making recommendations (e.g., users, artists, record labels or the service itself).

In this chapter, we explore how implicit feedback provided by users can be leveraged by a recommender system to provide more value to both the user and the artists.

We assume that different users could bring more or less value to the artists the system recommends to them. Music artists likely prefer to be recommended to those users who will actively engage more with their production, such as listening to their music, attending their concerts and buying their merchandising. Unfortunately, strong engagement metrics between users and artists can hardly be tracked by the existing music streaming platforms, which instead rely on implicit interaction signals such as play counts or session lengths to quantify the engagement and satisfaction of users with the recommended content (Mehrotra et al., 2019). Regarding artist recommendation, most music recommender systems consider the number of times a user plays a track or an artist (the *playcount*) as the main engagement signal (Jannach et al., 2018).

Playcounts alone can hardly cover all the different ways in which listeners "consume" an artist's production. For example, a listener who frequently listens to only the same few tracks is unlikely to be attracted by new releases by that artist or attend their concerts. On the same line, listeners who played a few

albums by the artist only for a few days in the past are likely less engaged with the artist than listeners who constantly listen to the artist's tracks over long periods of time. For these reasons, it is important to use implicit interaction signals beyond the simple frequency of interaction with the artist, i.e., the number of days that the user interacted with artists, how many different songs the user listened from an artist and the number of times a user listened to an artist. Therefore, we introduce novel signals that capture both the *breadth* of the listener's engagement with the artist's production, computed as the number of distinct artist's tracks played by the listener (*trackcounts*) and the *temporal extent* over which the listener engaged with the artist, computed as the number of days a user listens to an artist (*daycounts*). As far as we know there is no prior work that tries to capture using the implicit feedback signals how much a user is engaged with a music artist.

In this chapter, we study the behavior of state-of-the-art Implicit Matrix Factorization optimized with ALS (Hu et al., 2008) over these new engagement signals, both from the listener's and artist's perspective. We evaluate both the case in which these relevance functions are used as implicit relevance values to train ALS, and when they are used as evaluation metrics over test data in combination with other traditional offline evaluation metrics (such as MAP and NDCG). We run experiments over four different datasets in the music domain in order to understand better the quality of the engagement of the users with the recommended artists.[8]

5.2 Implicit Engagement Signals

In this section, we describe the implicit engagement signals that we use to train and we evaluate artist recommendations generated by ALS. We consider both *raw signals* that are extracted directly from the user's listening logs, and *composite signals* which are combinations of the raw ones.

5.2.1 Raw Signals

Given a listener u and artist a, we extract the following raw signals from listening logs:

- $playcounts(u,a)$ is the number of tracks of a played by u;
- $binary(u,a)$ is the binarized playcount, i.e. $binary(u,a) = \mathbb{1}\{placounts(u,a) \geq 1\}$;
- $trackcounts(u,a)$ is the number of distinct tracks of a played by u;

[8] The findings described in the following sections are based on our published work presented in Ferraro et al. (2020e)

- *daycounts*(u,a) is the number of distinct days in which u listened to at least one track by a.

5.2.2 Composite Signals

To capture multiple aspects of the listener's behavior with the artist in a single implicit signal, we combine the raw signals into two novel composite signals named *engagement* and *fidelity*.

- *engagement*(u,a) is a discounted weighted combination of the playcounts accumulated by the listener u over the days they have listened to a. Specifically, weight plays on the first days of listening less than plays happening later down by using the following formula:

$$engagement(u,a) = \sum_{d=0}^{D} playcounts(u,a,d) * log(d)$$

  Where D is the number of days a user listens to an artist and $playcounts(u,a,d)$ is the number of tracks of a played by u on day d.

- $fidelity(u,a)$ combines *engagement* with *trackcounts* into a single metric in the following way:

$$fidelity(u,a) = \alpha * trackcounts(u,a) + (1-\alpha) * engagement(u,a)$$

The motivation for the given definition of Engagement is that we want to weight play interactions differently, according to the day they were played. We assume that plays on the first day are less valuable for the artist than plays on the subsequent days. We apply logarithm to the number of days to soften the impact of large day numbers, making this factor more determinant in the first days of listening. For example, the difference between the first and the third day of listening is larger than between the tenth and the twentieth.

Fidelity combines the three raw signals by a linear combination of *trackcounts* and *engagement*, which is already combining *playcounts* and *daycounts*.

5.3 Evaluation Metrics

If we want to generate recommendations that will lead to more fans for an artist, the first difficulty that we have to solve is how to measure that. Measuring how much a user is a fan of an artist is a complex task. The most common signal used for measuring how much a user likes an artist is playcounts (Ricci

et al., 2010). However, this signal doesn't capture relevant information about the interaction of users with artists (e.g., during how much time, how many times a day, how many songs of the artist). We propose to use multiple metrics that capture characteristics that define how much a user is involved with an artist and are necessary to understand if the user will attend to the artist's shows and purchase the artist's merchandising. The metrics that we propose are: 1) During how much time a user listens to an artist; 2) how many times a day the user listens to the artist; and 3) how many songs of the artist the user listens to (if they cover more of the artist's repertoire or not).

To understand the quality of the recommendations both from the listeners' and from the artists' perspectives, we compute both *listener-centric* and *artist-centric* metrics. Listener-centric metrics are the following traditional offline accuracy metrics: Mean Average Precision (MAP) and normalized Discounted Cumulative Gain (*nDCG*) (Ricci et al., 2010).

The artist-centric metrics are instead the average values of playcounts (PLAYS@K), trackcounts (TRACKS@K) and daycounts (DAYS@K) over all the artists that were recommended. We also compute the coverage (C@10) of the recommended artists as in Oramas et al. (2016). Let A and U be the sets of artists and users in the dataset respectively, $A = \{a_1, a_2, ..., a_n\}$ and $U = \{u_1, u_2, ..., u_n\}$. We define the above artist-centric metrics as follows:

$$PLAYS@K = \frac{1}{|A|} \sum_{a \in A} \frac{\sum_{u \in U} hit@K(u,a) \cdot playcounts(u,a)}{\sum_{u \in U} hit@K(u,a)}$$

In a similar way we define the TRACKS@K and DAYS@K:

$$TRACKS@K = \frac{1}{|A|} \sum_{a \in A} \frac{\sum_{u \in U} hit@K(u,a) \cdot trackcounts(u,a)}{\sum_{u \in U} hit@K(u,a)}$$

$$DAYS@K = \frac{1}{|A|} \sum_{a \in A} \frac{\sum_{u \in U} hit@K(u,a) \cdot daycounts(u,a)}{\sum_{u \in U} hit@K(u,a)}$$

where $hit(a,u)$ returns 1 if and only if a was recommended to u *and* belongs to the test set of that user. Finally, given L_u as the top-k artists recommended to user u, we define the catalog coverage of recommended artists as

$$C@K = \frac{\bigcup_u L_u}{|A|}$$

Note that $playcounts(u,a)$ is the number of times that the user u listened to the artist a, $trackcounts(u,a)$ is the number of tracks that the user u listened to from the artist a and $daycounts(u,a)$ is the number of days that the user u listened to the artist a.

5.4 Datasets

For this experiment, we use four datasets of user-artist interactions with timestamps (see Table 5.1). To reduce noise, the original datasets are filtered according to the following constraints. First, we discard all artists having less than 3 interactions and all users that interacted with less than 10 artists. Having users with less of these interactions would make the recommendations harder to evaluate. Then, we split each dataset on a temporal-basis by first sorting the interactions by timestamp, and then we assign the first 80% of the events to the training set and the remaining 20% to the test set. Finally, since our goal is to study the impact of recommendations of artists that were not previously listened to by the user, and to understand how artists can reach new audiences through recommendations, for each user we removed from the test set all the artists occurring in their training set. The resulting number of artists and users on each dataset is detailed in Table 5.2[9].

Table 5.1: Datasets used in the comparison.

lfm-1b	Large dataset with over a billion listening events containing playcounts and timestamp extracted from last.fm (Schedl, 2016)
Streaming-service	Dataset obtained from a music streaming service for 6 months in 2019
Nowplaying	Dataset containing listening logs collected from Twitter (Zangerle et al., 2014). We use a subset of the original dataset published by Ludewig et al. (2019)
30music	Dataset collected from last.fm (Turrin et al., 2015) with the main purpose of session recommendations

Table 5.2: Information about the datasets.

Dataset	Users	Artists		Density	# Days		User-Artist Interaction	
		train	test	train	train	test	train	test
lfm-1b	119,692	693,436	111,086	0.0007	3073	122	61,443,465	517,903
Streaming-service	25,981	22,667	17,189	0.0028	147	36	1,655,600	235,653
Nowplaying	7,198	13,213	7,921	0.0033	428	107	318,250	25,124
30music	33,462	112,354	97,274	0.0010	292	73	3,888,882	1,307,575

In Figure 5.1, we show the distribution of playcounts for user-artist interactions with a value bigger than zero. We see that each dataset has a different distribution, which may lead to a different behaviour in the recommendations.

[9]For reproducibility purposes code is provided: https://github.com/andrebola/artist-engagement

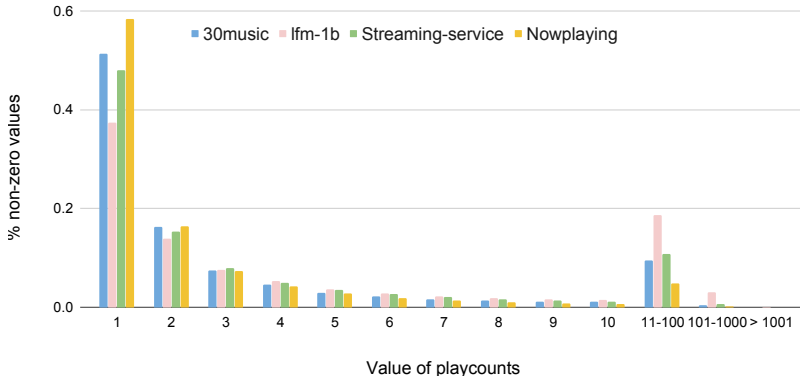

Figure 5.1: Distribution of playcount values in the datasets.

If we analyze the distribution of playcounts for user-artist interactions, we observe that some datasets have a higher distribution of values bigger than 10, such as the lfm-1b, which have much more values between 11 and 1000 compared with the other datasets. This observation implies a richer interaction between users and artists with respect to other datasets, such as Nowplaying, where there are a majority of playcount values closer to one. Considering the different distributions of playcounts may give some light when interpreting the results of the offline evaluation over the different datasets.

5.5 Correlations Between the Raw Signals

We provide here an analysis of the correlations between raw signals introduced in Section 5.2. We hypothesize that these signals capture different and complementary aspects of how listeners engage with artists.

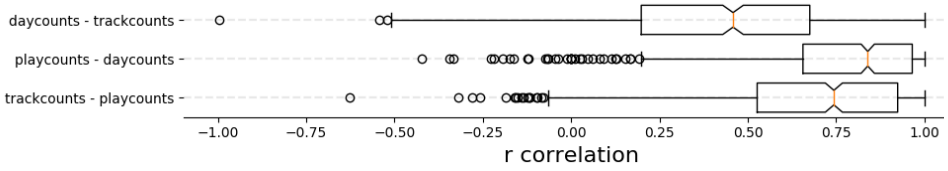

Figure 5.2: Correlation for *playcounts*, *daycounts* and *trackcounts* on 1000 artists.

We measured the correlation (r) between the described raw signals for 1000 random artists of the lfm-1b dataset. In Figure 5.2 we see that the highest average correlation is between *playcounts* and *daycounts*. However, for some artists, these values are not very correlated, which means that for those artists there could be a higher benefit of using other signals than only *playcounts*. Also note that for *daycounts* and *trackcounts* the correlation is lower for most of the artists.

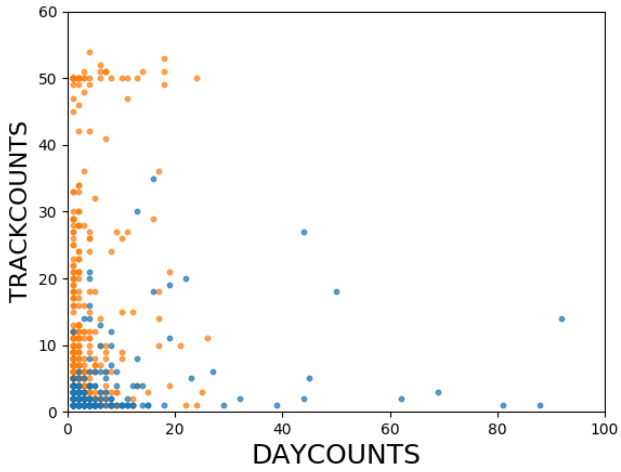

Figure 5.3: Distribution of Users' *trackcounts* and *daycounts* raw signals for 'The Honeycombs' (blue) and 'Roland Pontinen' (orange) in the LFM-1b dataset.

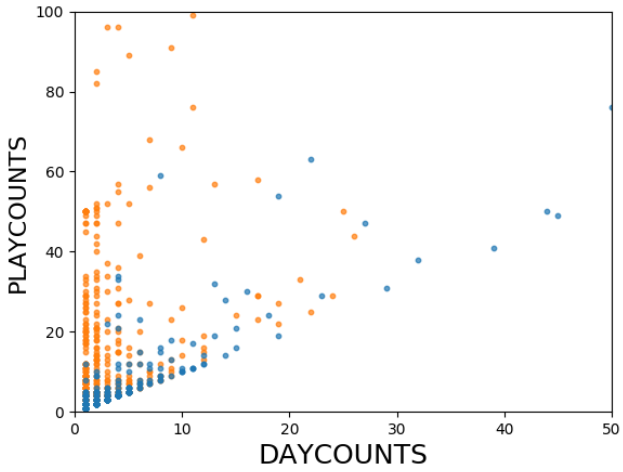

Figure 5.4: Distribution of Users' *playcounts* and *daycounts* for 'The Honeycombs' (blue) and 'Roland Pontinen' (orange) in the LFM-1b dataset.

We further illustrate this with two artists taken from lfm-1b dataset, which have similar popularity (i.e., number of users) but different music styles. (a) Roland Pontinen is a pianist and composer of chamber music from Sweden and (b) The Honeycombs was a British band from the '60s influenced by The Beatles. We can expect a different behavior between the users that listen to these artists, for this we plot the correlation between the raw signals for the users that interacted with the artists (a) and (b).

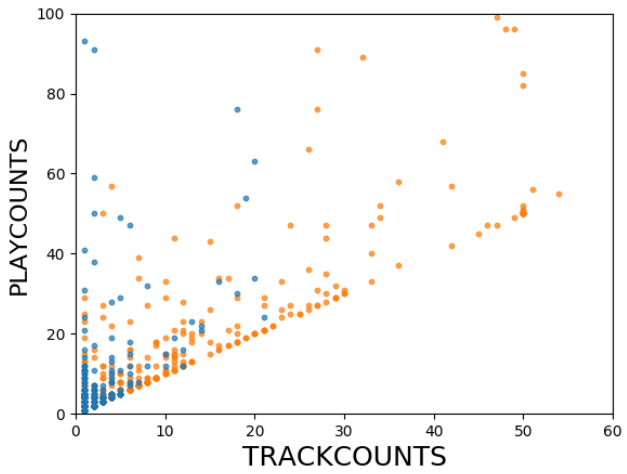

Figure 5.5: Distribution of Users' *playcounts* and *trackcounts* values. for 'The Honeycombs' (blue) and 'Roland Pontinen' (orange) in the LFM-1b dataset.

The Figure 5.3 shows the users' distribution of *trackcounts* and *daycounts*, Figure 5.4 shows the users' distribution of *playcounts* and *daycounts* and Figure 5.5 shows the users' distribution of *playcounts* and *trackcounts*. These figures highlight some interesting differences in the way users engage with both artists. We can see that *trackcounts* and *playcounts* are more correlated for (b) (r=0.32) than for (a) (r=0.06), whereas *playcounts* and *daycounts* are more correlated for (a) (r=0.99) than for (b) (r=0.43).

In these two analyses we observe that the correlation of the raw signals can be different between artists. Therefore, they provide complementary information that could be useful for generating recommendations. Furthermore, this analyses motivates the use of the different signals as input for the recommendations since we understand that the fans of the artists can behave differently. Therefore, we can obtain useful information that might not be correlated with the typically-used *playcounts*.

5.6 Recommendations Using Engagement Signals

In this section we analyze the performance of the recommender system for each dataset to understand if the implicit engagement signals that we proposed in Section 5.2 can be favorable to artists while keeping acceptable levels of (offline) recommendation quality to listeners. For these reasons, we decided to study the behavior of the Implicit Matrix Factorization with Alternating Least Squares (ALS) (Hu et al., 2008) in this scenario. ALS is known to be

one of the most used collaborative filtering algorithms and a *de-facto* industrial standard. While we cannot know what algorithms are used by the various online music services available nowadays, the choice of ALS surely extends the applicability of our experimental results to many real-world music recommendation scenarios. We trained ALS[10] on all training datasets with each of the implicit engagement signals defined in Section 5.2 as relevance functions. For the case of *fidelity*, we decided to give the same weight to *engagement* and *trackcounts* (α=0.5) to simplify the experiments, but further optimization of these weights may lead to improved results. To measure the performance of the recommendations, we generate a list of 10 artists for each user (K=10) and we use the metrics defined in Section 5.3. We tuned only the number of latent dimensions for each (dataset, relevance function) combination. The final number of dimensions used are 200, 200, 50 and 30 for lfm-1b, Streaming-service, 30music and nowplaying respectively.

In Table 5.3, we show the performance according to the listener-centric metrics (MAP@10 and nDCG@10) and the artist-centric metrics (PLAYS@10, TRACKS@10, DAYS@10 and C@10).

The results show that there is no single relevance function that performs the best on all the datasets in terms of listener-centric metrics. Training the model with the function *daycounts* performs the best for the 30music dataset, while training the model with *trackcounts* performs better for the nowplaying dataset. Training the model with the *engagement* function, which is a discounted weighted combination of an artist's *trackcounts* over the days it was played by the listener, performs the best on the Streaming-service dataset. Interestingly, *binary* input obtains the worst performance for all the previous three datasets and the highest performance for the lfm-1b dataset. A possible explanation for this is that lfm-1b dataset presents a higher proportion of values greater than 10, as it was noticed in Figure 5.1, which might be beneficial for *binary* in this dataset according to the listener-centric metrics.

From these results, we cannot say that an implicit engagement signal will give always the best accuracy from the listener's perspective. We hypothesize that it is related to the nature of each dataset, as they all present a different distribution of values.

However, we see more consistent results in all datasets according to artist-centric metrics. Training the model using *engagement* relevance function outperforms training with all the other relevance functions in terms of C@10 on each dataset. This suggests that training with the *engagement* as relevance function increases the fraction of artists that are effectively recommended in the top-10 in what respects all the other relevance functions. Function *engagement* performs particularly well also in terms of DAYS@10, for which it

[10]We used the implementation available at https://implicit.readthedocs.io/en/latest/

Table 5.3: Evaluation of the recommendations in all the datasets

		listener-centric		artist-centric			
	Rel. fun.	MAP@10	nDCG@10	PLAYS@10	TRACKS@10	DAYS@10	C@10
lfm-1b	*binary*	**0.0290**	**0.0640**	6.0294	3.6433	1.7735	0.0128
	playcounts	0.0256	0.0580	**8.5717**	**4.3151**	1.8706	0.0291
	daycounts	0.0287	0.0632	7.1185	3.8260	1.8619	0.0235
	trackcounts	0.0279	0.0623	7.6089	4.2643	1.8420	0.0210
	engagement	0.0240	0.0545	8.5211	4.2640	**1.8887**	**0.0324**
	fidelity	0.0253	0.0574	8.3988	4.1912	1.8483	0.0292
Streaming-service	*binary*	0.0519	0.1024	2.7665	1.7372	1.6663	0.1099
	playcounts	0.0610	0.1177	3.4088	1.9597	1.7535	0.1697
	daycounts	0.0585	0.1136	3.1650	1.8468	1.7200	0.1560
	trackcounts	0.0573	0.1122	3.3188	**2.0064**	1.6983	0.1372
	engagement	**0.0619**	**0.1193**	**3.4272**	1.9669	**1.7620**	**0.1826**
	fidelity	0.0615	0.1185	3.2919	1.9666	1.7209	0.1695
nowplaying	*binary*	0.0527	0.1019	2.5673	1.9380	1.6365	0.0673
	playcounts	0.0553	0.1071	3.2953	2.0027	1.7635	0.0892
	daycounts	0.0540	0.1044	2.6502	1.8291	1.7155	0.0901
	trackcounts	**0.0563**	**0.1088**	3.2008	**2.1385**	1.6348	0.0745
	engagement	0.0522	0.1019	**3.9596**	2.0120	**1.8485**	**0.0951**
	fidelity	0.0553	0.1063	3.9416	2.0553	1.7928	0.0892
30music	*binary*	0.0659	0.1360	4.0627	3.9759	1.4631	0.0123
	playcounts	0.0679	0.1403	5.5051	5.2998	1.4678	0.0214
	daycounts	**0.0703**	**0.1432**	4.7363	4.5380	**1.5012**	0.0177
	trackcounts	0.0680	0.1406	5.5297	5.3305	1.4702	0.0213
	engagement	0.0669	0.1384	**5.7237**	**5.4944**	1.4921	**0.0228**
	fidelity	0.0687	0.1411	4.9575	4.7234	1.4903	0.0134

is the best function in all but the 30music, where it is the second-best. The *engagement* also performs particularly well in terms of PLAYS@10, for which it is the best function in all but the lfm-1b, where it is the second-best. This suggests that *engagement* tends to recommend artists to users who will likely engage with them for *longer* time and *more frequently*. While we do not observe consistently the same behavior for TRACKS@10 on all datasets, training the model with the *trackcounts* relevance function seems to give the highest performance and also *engagement* has a notable performance on this metric as well. Interestingly, *fidelity* does not outperform the other relevance functions in any of the datasets and metrics. This suggests that the simple linear combination of *engagement* and *trackcounts* is not sufficient, and it should be investigated in future works. However, training the model with the proposed *fidelity* relevance function somehow balances all the metrics that we evaluated from the artists' perspective and from the listeners' perspective.

5.7 Conclusions

In this chapter, we proposed new signals for listener's engagement in music recommender systems. We used these signals both as relevance functions to train Implicit Matrix Factorization, and also as evaluation metrics to gauge how traditional "listener-centric" recommender systems impact listeners and artists differently.

Our results suggest that listener-centric quality is highly dependent on the choice of the relevance function and of the dataset they are tested on. It is therefore an important parameter to optimize when designing a music recommender system.

Looking at the results from the artists' perspective, the proposed *engagement* relevance function, which combines *playcounts* and *daycounts*, performs better in most datasets. *Engagement* provides, in general, a higher average consumption of the artists' music in terms of the number of plays and number of days. It also notably increases the fraction of recommended artists overall. However, regarding distinct tracks played per artist, *trackcounts* still performs better in some datasets, suggesting that it is an important implicit signal to capture when optimizing for a wider consumption of the artists' catalog. More investigation is needed to properly combine the three individual implicit signals in a single one.

CHAPTER 6

Algorithmic Influence in Session-Based Recommendation

6.1 Introduction

Recommender systems can have a major influence on the information that we receive and—at least to a certain extent—also on the decisions that we make, since they determine the content that we see. Recommender systems could allow the user to discover previously unknown content by exposing elements that the user might be interested in. However, an algorithm may have undesired effects when exposing some items repeatedly, for example, recommending too many already popular items. In music recommendation, it is a known behavior of algorithms based on collaborative filtering to expose more the already popular items (Celma, 2009). Also in domains related to e-commerce it was investigated that collaborative filtering approaches may decrease sales diversity because of such "blockbuster effects" (Fleder & Hosanagar, 2009; Jannach et al., 2015b), and they were also observed in field tests (Lee & Hosanagar, 2019). On news websites or on social media sites, for example, an over-emphasis on certain types of content may lead to a biased distribution of information, and to effects like filter bubbles (Pariser, 2012). Solutions had been proposed to mitigate the popularity biases in the recommendations leading to a higher coverage by a small sacrifice of the accuracy (Abdollahpouri et al., 2017; Jannach et al., 2015b). Understanding such undesired effects is, however, important in many domains.

The problem of such biases might be particularly pronounced in the context of *session-based* recommendation scenarios, where the system has to deal with anonymous or first-time users (Hidasi et al., 2016; Quadrana et al., 2018). Such

situation is very common and highly relevant in practice when many of the recommendations of the system are based only on the few observed interactions in the ongoing session. The level of personalization may therefore be lower and lead to the effect that many of the provided recommendations consist mainly of generally popular items. Over time, the repeated recommendation of these items may then again result in a reinforcement effect.

From the results described in Chapter 3, we see that such effects of algorithmic recommendations are not aligned with two aspects that the artists expressed as requirements to make the music platforms fair from their perspective. First, the artists voice that not only the most popular artists have to be included in the recommendations shown to users. The system should also give more exposure to music items in the long tail of the popularity distribution. Second, the artists consider that systems should not influence the users' taste. For this reason, in this chapter, we further investigate the longitudinal effects of session-based recommendations with a simulation-based approach. Starting from real-world data of recorded user interactions from the music domain, we first generate session-based recommendations from a selected set of seed tracks with different algorithms. We then assume that a certain fraction of the recommendations are listened to by the users, and we correspondingly extend the dataset with these new interactions. This process is repeated in the simulation and from time to time we re-train the underlying models. After these retraining steps, we measure if the recommendations have changed on a system-wide level. In particular, we analyze if the recommendations have become more concentrated on a small subset of the items or not. Finally, in this chapter, we also investigate the effectiveness of applying reranking strategies to mitigate the concentration effects identified in the recommendations.[11]

It is important to note that we simulate users' interactions with the recommendations but we do not consider that users could also consume items that are not recommended by the algorithm. Even if this differs from a real-world case, our goal is to focus on the effect that the algorithms would have if the users follow the recommendation. Therefore, we propose as future work incorporating in the simulations other interactions to reduce the effect of the algorithms.

6.2 Methodology

We investigate how many of the available items are presented to users in their top-n lists by different algorithms. Herlocker et al. (2004) and others refer to this measurement as *catalog coverage*. We adopt the definition from Herlocker

[11] The findings described in the following sections are based on our published work presented in Ferraro et al. (2020c)

et al. (2004) and additionally measure how coverage develops over time. They propose to measure it by creating the union of the top-10 recommendations at a given point in time, and they emphasize that the metric should be combined with accuracy measures. We rely on their definition but additionally measure how coverage develops over time. In this work, we also assume that higher aggregate diversity (i.e., coverage) in the recommendation lead to higher consumption diversity, as investigated for example in Lee & Hosanagar (2019).

Our general research methodology is based on simulation principles that were followed in Chapter 4 and also adopted in Zhang et al. (2020b) and Jannach et al. (2015b). As a starting point for our simulations, we use two datasets containing real user interactions on the music domain that include session information. One contains listening histories from the online music service *last.fm* called *30Music* (Turrin et al., 2015); the other one, called *#nowplaying*, was extracted from music-related Twitter messages by Zangerle et al. (2014). Regarding dataset characteristics, the #nowplaying dataset comprises 1.2M listening events in 146K sessions for 61K items. The 30Music dataset is even larger, with 2.8M events 166K sessions for 446K items, i.e., sessions here are generally longer as well.

Note that both datasets contain user IDs and that long-term listening histories are available. Since we focus on session-based recommendation problems, we do not take these long-term models into account when recommending. Therefore, we treat each session of a user as an independent session.

The main simulation procedure is as follows:

1. For each session in the dataset, we select one track as a seed for a new listening session.
2. We generate playlists from the seed track using different session-based algorithms and measure the characteristics of the recommendations at a system-wide level.
3. We assume that a certain fraction of the tracks in the playlist are listened to by the users and we add these simulated listening events to the dataset.
4. We update the models at defined intervals and continue with Step (1).

In our experiments, we used the following configuration:

- As seeds in **Step 1**, we randomly selected one of the tracks played in each session. We also made experiments with other seeds, e.g, the most frequently played track in a session or the track that would receive the highest rating prediction by a matrix factorization algorithm. The obtained results were similar to the random seeds in terms of the general characteristics, which is why we omit them.

- In **Step 2**, we created recommendations of list length 30. If all 30 tracks are listened, this roughly corresponds to a two hours music experience in case of pop songs.
- In **Step 3**, we assume that the users on average consume about one third of the recommendations, i.e., 10 tracks, based on the observations regarding adoption rates in the music domain in Kamehkhosh et al. (2020). The selection of the tracks was done randomly. In general, modifying this parameter would result in slower or faster changes in the behavior of the recommender, but it would not impact the general characteristics of the emerging phenomena.
- We retrain the models in **Step 4** after having generated artificial playlists 3 times. Assuming that models in real-world deployments are retrained over night, there would be 3 sessions per day before the models are updated.

Measurements are taken in Step 2 after each re-training step. To see how recommendations change over time for a given set of items, we repeatedly took the following measures for the tracks that were used in the first simulation round:

- the *Gini* index to assess the concentration of the recommendations on certain items. Higher values mean higher concentration (Zhang et al., 2020b; Jannach et al., 2015b);
- catalog *coverage* in terms of the absolute number of different items appearing in the top-n lists;
- the average item *popularity* in terms of the number of times a recommended track appears in the dataset;
- the information retrieval measures *precision*, *recall*, and F1 at list length 10 as accuracy measures.

In this chapter, we seek to understand differences between session-based recommender algorithms in terms of their longitudinal effects. Depending on the application domain, the choice of the algorithm can then be based on these observations. We consider algorithms from different families in our simulations, as shown in Table 6.1. The hyper-parameters of the algorithms were optimized for accuracy on the training data. To ensure reproducibility we share all code and data used in the experiments online, including the configuration for splitting the data and the parameters used for each algorithm.[12]

[12] https://github.com/andrebola/session-rec-effect

Table 6.1: Algorithms used in the Comparison

GRU4REC	The first widely-used neural approach to session-based recommendation, based on RNNS (Hidasi et al., 2016).
NARM	An attention-based neural method by Li et al. (2017), often leading to competitive results (Ludewig et al., 2019).
SKNN	A nearest neighbor technique that shows competitive results in a number of domains (Ludewig & Jannach, 2018).
CAGH	A simple yet often effective baseline proposed in Bonnin & Jannach (2014), which recommends the greatest hits of artists that are similar to those appearing in the seed tracks.

6.3 Results

We report results for three experiments, discussed in Section 6.3.1 to Section 6.3.3.

6.3.1 Experiment 1: Analysis of Initial Recommendations

In the first measurement, we determine precision and recall for the different algorithms for the first round of recommendations, i.e., on the original data, and we additionally measure the Gini index, coverage, and the average popularity of the recommended items.

Table 6.2: Results for first simulation round for the #nowplaying dataset.

Algorithm	F1	Precision	Recall	Gini	Popularity (abs.)	Popularity (rel.)	Coverage
SKNN	**0.1550**	0.1482	**0.1624**	0.4782	57,7683	39,8138	**61,161**
NARM	0.1481	**0.1490**	0.1472	0.5982	65,7828	48,0066	59,578
GRU4REC	0.1227	0.1175	0.1283	**0.4169**	**22,7044**	**4,7920**	61,119
CAGH	0.0002	0.0005	0.0001	0.9301	171.7474	153.6176	24,718
Random	0.0000	0.0002	0.0000	0.0667	17,9516	-0.1179	61,220

The results for the #nowplaying dataset are shown in Table 6.2. In terms of accuracy measures, we find that SKNN and NARM are working best in this experiment. GRU4REC works slightly worse in this setup, where we only use the first element of a session to create the playlist. CAGH, although often competitive in alternative setups, does not work in a satisfactory way in such a cold-start setup. Regarding the other metrics, we find that among the well-performing techniques, NARM has both the highest concentration bias and the strongest tendency to recommend popular items (see column "Popularity (abs.)"). This is important because it shows that deep learning based techniques such as GRU4REC and NARM, despite comparable performance in

accuracy, can recommend largely different items to users in their top-10 lists. [13] SKNN represents the middle ground here, but still leans quite strongly to recommend popular items. The column "Popularity (rel.)" shows the difference between the average popularity of the recommendations and the seed track. All algorithms recommend tracks that are more popular than the seed tracks, with GRU4REC being the best in terms of retaining the original popularity level. CAGH, by design, is worst here, as it only recommends greatest hits of artists. The differences in terms of *coverage* among SKNN, NARM, and GRU4REC are low and all of them recommend almost all of the about 61k different tracks at least once. Notably, the coverage of the random recommender is at about the same level. The coverage measure therefore seems to be not as informative as the Gini index, because even if one item only appears once in the thousands of generated playlists, it will increase this measure. The results for the 30Music datasets shown in Table 6.3 are similar in terms of the accuracy measures. On this dataset, however, SKNN is also favorable in terms of the beyond-accuracy measures.

Table 6.3: Results for the first simulation round for the 30Music dataset.

Algorithm	F1	Precision	Recall	Gini	Popularity (abs.)	Popularity (rel.)	Coverage
SKNN	**0.1988**	**0.1802**	0.2218	**0.5629**	**21,6084**	**15,5684**	**429,338**
NARM	0.1955	0.1697	**0.2306**	0.7116	23.9657	17.9276	365,736
GRU4REC	0.1537	0.1318	0.1844	0.6547	24,0323	18,0633	397,470
CAGH	0.0000	0.0000	0.0000	0.9340	83.2055	77.0883	141,257
Random	0.0000	0.0000	0.0000	0.1678	5.9608	0.0177	446,769

6.3.2 Experiment 2: Longitudinal Analysis of Concentration, Coverage, and Popularity Effects

In this experiment, we repeatedly generated playlists by randomly choosing seed tracks for each training session. In total, we made 30 simulation rounds. After each round we added the simulated interactions to the dataset and we re-trained the models in every third round, leading to the 10 iterations shown in Figure 6.1.

Looking at the Gini index (Figure 6.1), we observe that all algorithms in this comparison lead to an increased concentration effect over time. As expected from the results after the initial recommendations, the concentration increases

[13] We could run NARM only for five iterations before running out of memory. The tendency is however clear, and a simulation with 10% subsamples of the datasets confirmed the trends observed on the full data.

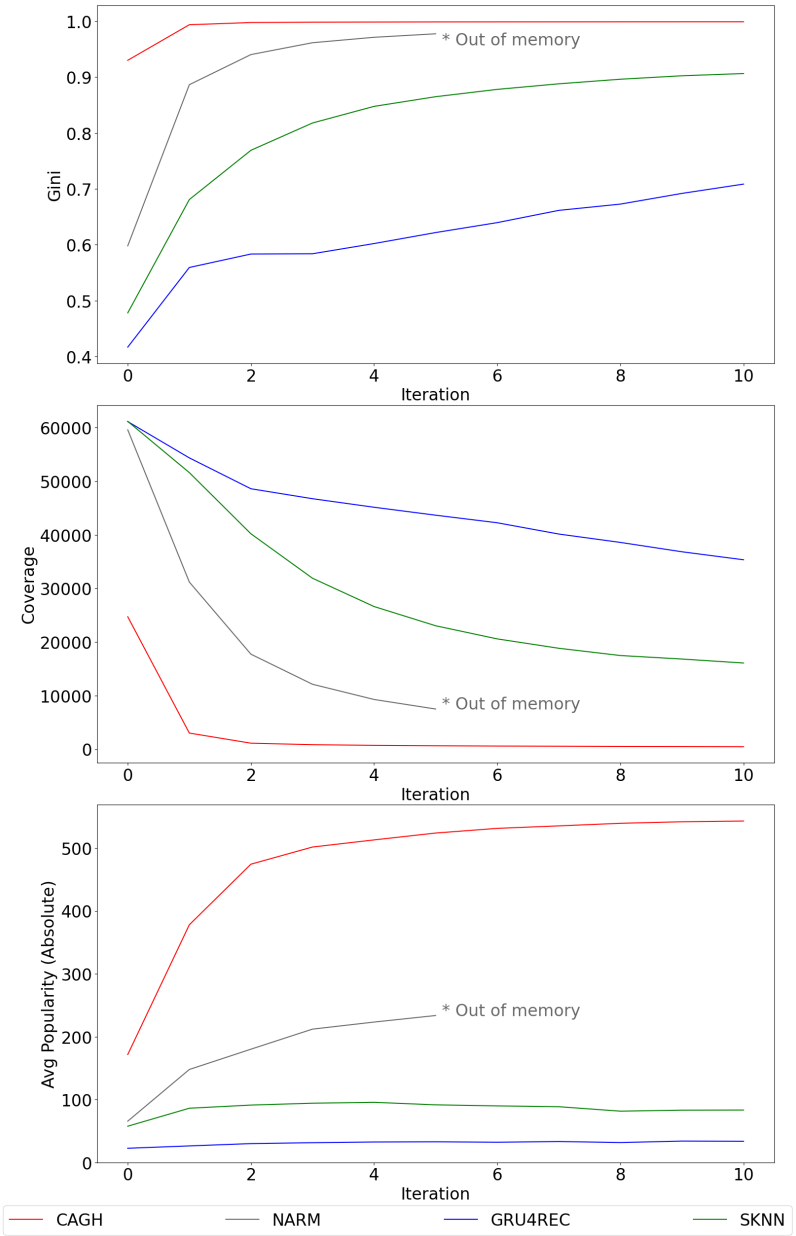

Figure 6.1: Simulation Results for the #nowplaying Dataset. NARM ran out of memory (>64 GB) after 5 iterations as we add more data to the training set. Additional simulations (not shown here) in which we created playlists for only 20% of the data in each round confirmed the trends observed for the full datasets.

most strongly when using NARM (excluding again CAGH), and it increases more slowly for GRU4REC. The development of the coverage metric follows this trend, as shown in Figure 6.1, i.e., the coverage of all algorithms goes down

steadily during the simulation, with the strongest effect observed for NARM and the weakest for GRU4REC. Interestingly, the average popularity level remains mostly constant for SKNN and GRU4REC, with GRU4REC generally having a lower popularity bias than SKNN. For NARM, the popularity bias however increases over time.

We run all the experiments with 64GB of memory but for the case of NARM we could not retrain after the fifth round because it requires much more memory than the other algorithms. Therefore, we sub-sample the sessions to only a 20%.

The simulation results for 30Music are shown in Figure 6.2. The general observations are similar to those obtained for the #nowplaying dataset, i.e., all algorithms lead to a concentration over time and to a reduced coverage. As expected from the results of the initial recommendations (Table 6.2 and Table 6.3), we can however see that SKNN behaves favorably on this dataset in terms of concentration and coverage.

We see that there are some differences across the datasets with respect to individual algorithms. In particular the concentration effect of GRU4REC is much higher on #nowplaying dataset, which indicates that certain developments over time depend on the characteristics of the underlying datasets.

6.3.3 Experiment 3: Longitudinal Effects of Using Reranking as a Countermeasure

In a final set of experiments our goal was to investigate the effectiveness of applying reranking strategies to avoid concentration effects as was done, e.g., by Adomavicius & Kwon (2012). We analyzed the effects of two basic heuristics:

- In *Reranking Strategy 1*, our goal was to avoid recommending too often tracks that were recommended frequently in previous rounds to all users. Technically, in each round we count the number of times an item i was recommended, denoted as *previous_recs(i)*. In the following round, we penalize frequently recommended tracks by moving them back in the recommendation lists. The penalty p in terms of the number positions to move the item i back is computed heuristically as:

$$p = 10 * log(\text{previous\_recs}(i))$$

- *Reranking Strategy 2* is personalized, and it tries to avoid recommendations that an individual user has consumed previously in the same session. Note that while our focus is on session-based recommendation, where

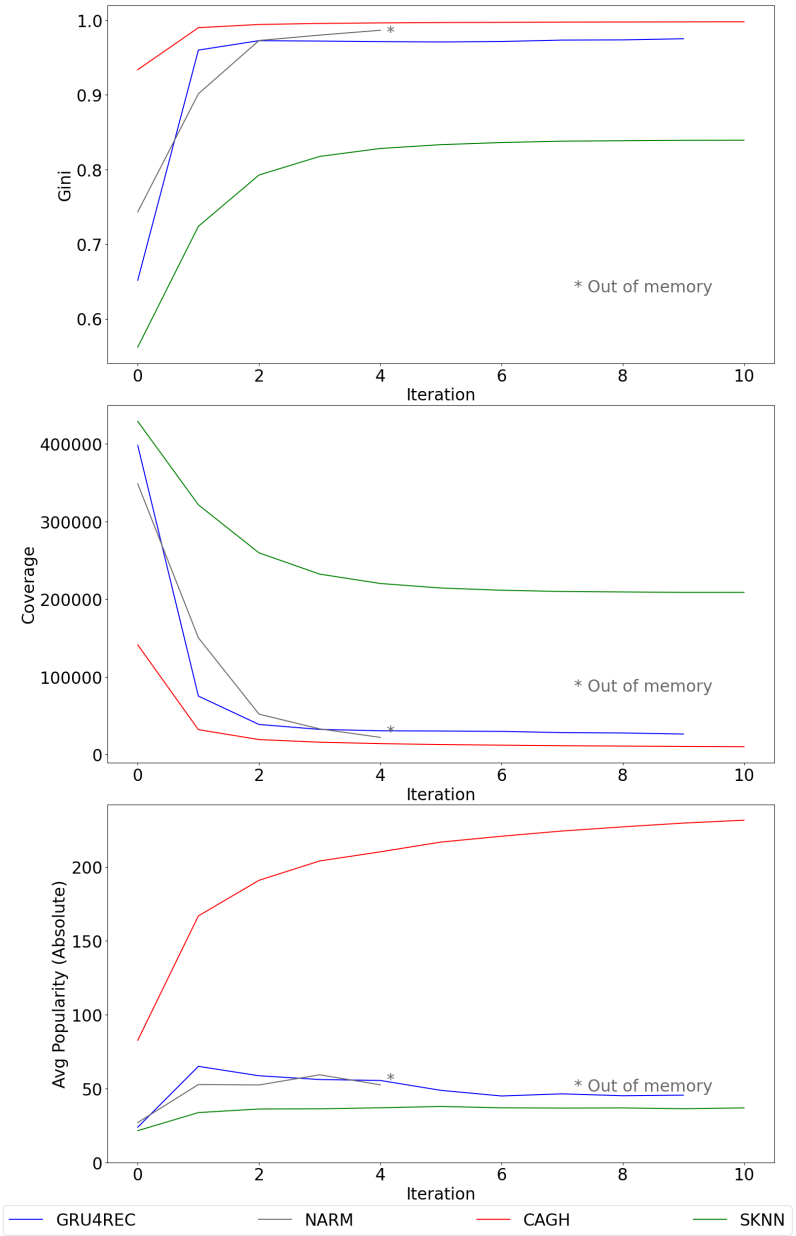

Figure 6.2: Simulation Results for the 30Music Dataset. NARM again ran out of memory after a few iterations.

long-term information is not generally available for all users, this experiment gives us some insights on the possible benefits of personalization. Technically, for each user u, we count the number of times it has consumed an item i in the session, denoted as *previous_consumptions(i,u)*. The penalty p for user u and item i is consumed as

$$p = 10 \times \text{previous\_consumptions}(i, u)$$

For example, if a user has listened to a track two times before, the track will be moved back 20 positions in the new recommendation list.

The results show that both re-ranking strategies are effective, but in slightly different ways. When avoiding to repeatedly recommend the same tracks to everyone (*Reranking Strategy 1*), we can observe that the increase in concentration can be slowed down for all algorithms except again for CAGH (compare Figure 6.3). For the case of individualized reranking *Reranking Strategy 2*, the increase of the concentration bias can be stopped at a certain level for SKNN and GRU4REC. For NARM, however, which exhibited a strong concentration bias already at the beginning, this personalized reranking does not seem to be very effective. This phenomenon can be attributed to the limited level of personalization of NARM, as shown in Table 6.2. Overall, however, the results also indicate that already simple reranking strategies can be effective countermeasures to avoid undesired concentration effects.

In previous works reranking strategies often come with a limited loss of accuracy (Adomavicius & Kwon, 2012) or sometimes even to a slight increase (Jannach et al., 2015a). In our case, the reranking strategies do not lead to a loss in accuracy. Looking a the precision and recall values obtained in our simulation experiment, we see that the accuracy of GRU4REC for both reranking strategies remains almost constant; for SKNN, the performance is even slightly increased, as observed previously for the music domain in Jannach et al. (2015a). These results are consistent for both datasets. Specifically, if we average precision and recall over all iterations for GRU4REC without reranking on the #nowplaying dataset, both precision and recall are at about 0.11. Applying either reranking strategy only leads to changes at the third place after the decimal. For SKNN, precision and recall even go slightly up with both strategies from about 0.13 up to 0.16. This shows that for SKNN we get the highest accuracy and at the same time we can improve coverage.

Finally, additional measurements show that coverage also ceases to go down for SKNN and GRU4REC when a personalized reranking strategy is applied, and that the popularity bias continues to remain stable. The same measurements were furthermore made for the 30Music dataset (except for NARM, again due to high computational costs). The results are again generally in line with what was observed for the #nowplaying dataset.

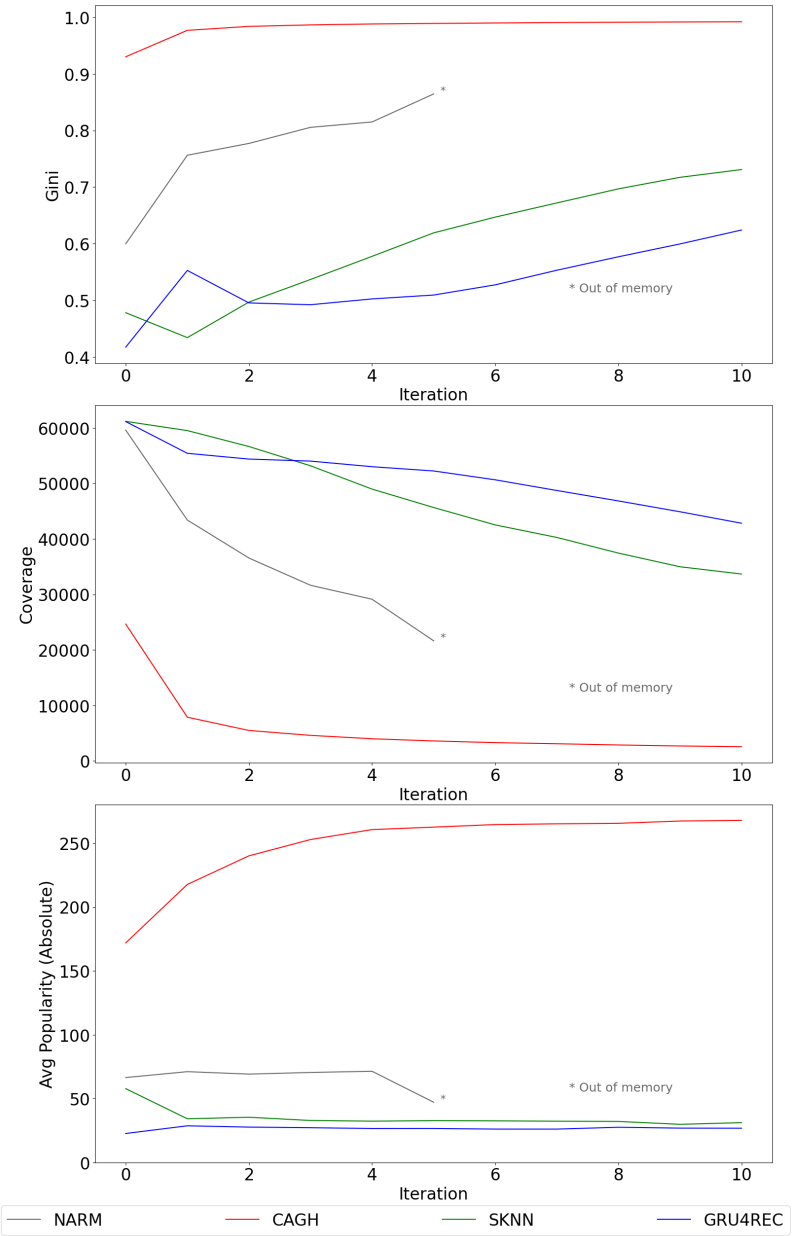

Figure 6.3: Simulation Results (Reranking) for the #nowplaying Dataset. Reranking based on Recommendation Frequency.

6.4 Conclusion

In this chapter, we analyzed through a simulation-based approach to what extent modern approaches to session-based recommendation exhibit certain

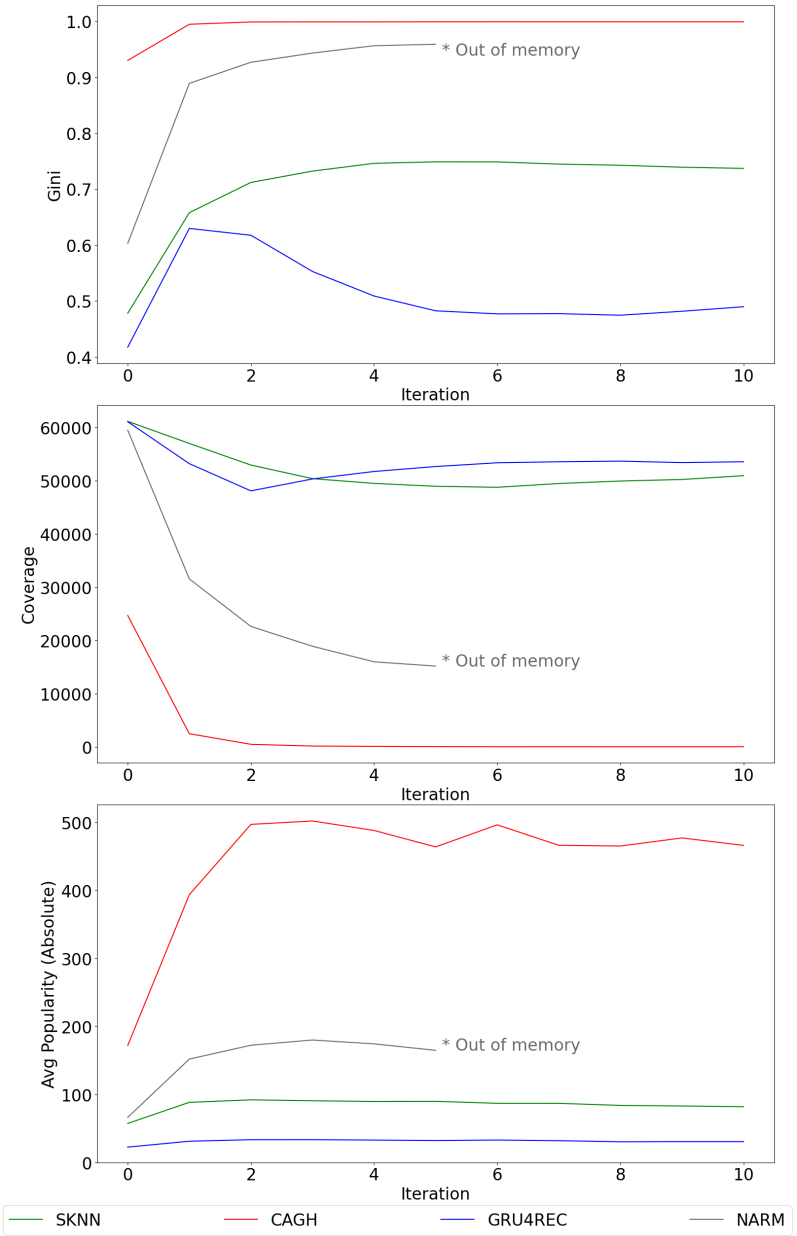

Figure 6.4: Simulation Results (Reranking) for the #nowplaying Dataset. Individualized Reranking based on Consumption Frequency.

potentially undesired biases, e.g., recommending the same set of items to everyone, and if these biases are reinforced over time. On a methodological level, we see our work as another step to move beyond today's "single-snapshot" analysis of recommendation algorithms, which does not allow us to investigate or simulate longitudinal effects of such systems.

Unlike the findings in Jannach et al. (2015b), our experiments show that all investigated types of algorithms, both recent neural ones and heuristic-based ones, may lead to undesired concentration effects over time. Furthermore, we find that even though the prediction accuracy of some algorithms is often comparable, they may exhibit largely different concentration tendencies and, as a result, recommend very different sets of items to users in the end. This observation is particularly important from a practical point of view since such differences might go unnoticed when an algorithmic comparison is solely based on accuracy measures. In practice, we are generally interested in a system that makes highly accurate predictions but does not exhibit undesired reinforcement tendencies. In an additional simulation experiment, therefore, we investigated the effects when applying a re-ranking strategy to avoid too many repeated recommendations. The experiment provides indications that relatively simple strategies can help to counteract the undesired effects without a loss regarding accuracy.

Reducing the popularity bias in recommendations is one of the aspects identified in Chapter 3 as important to make music platforms fair from the artists' perspective. The experiments described in this chapter show that commonly-used session-based recommenders may have such undesired biases in the long term, therefore, showing the importance of measuring such potential negative effects. Finally, the proposed reranking strategies allow mitigating such negative effects while not reducing accuracy, leading to a better recommendation for both artists and users.

CHAPTER 7

Melon Playlist Dataset

7.1 Introduction

Open access to adequately large datasets is one of the main challenges when conducting research in the fields of music information retrieval and music recommender systems due to the limitations of the copyrighted material. However, leveraging content information for music information is fundamental to make music recommendations that can recommend less popular and new content. Moreover, the lack of public datasets makes collaboration between researchers and reproducibility of academic studies more difficult, limiting developments in these fields.

Having representations of songs' audio with reduced information would make it possible to distribute music content avoiding licensing issues. However, reducing the information available from the content would have an impact in the performance of the models. For this reason, in this chapter we first compare the performance of using reduced mel-spectrogram representations as an input for the task of automatic tagging of music. More specifically, our research question is whether it is possible for the state-of-the-art approaches of auto-tagging (Choi et al., 2016; Pons et al., 2018) to operate on reduced data inputs without a significant drop in prediction accuracy.[14]

Based on these findings, in this chapter, we then present a public dataset of information about 148,826 playlists collected by Kakao[15] from Melon,[16] the most popular music platform in Korean. This dataset also contains the mel-spectrogram representations of the audio for 649,091 tracks, covering the music consumed in Korea (i.e., mainly Korean pop, but also Western music).

[14]The findings described in the following sections related with reduced representation for auto-tagging are based on our published work presented in Ferraro et al. (2020b)

[15]https://www.kakaocorp.com

[16]https://www.melon.com

Thus, we provide a large-scale public dataset of playlists that includes audio information for commercial music directly accessible without the need to collect it from different external sources, which is the problem of other existing playlist datasets.[17]

The playlists are collected from Melon users manually verified by moderators for providing quality public playlists. These users add metadata to the playlists, such as tags and titles, which are also included in the dataset. The dataset was originally collected for a challenge related to Automatic Playlist Continuation (APC) and tag prediction. Possible applications go beyond the scope of the original challenge, and the size of the dataset makes it suitable for deep learning approaches that require a large amount of information. New methods can be applied for music, e.g., deep metric learning, representation learning, and semi-supervised learning.

In order to reduce the information from the audio, we consider reducing the size of the input spectrograms in terms of both lesser amount of frequency bands and larger frame rates. We show that by reducing the frequency and time resolution we can train the network for automatic tagging faster with a small decrease in the performance. The results of this study can also help researchers to build faster CNN models as well as to reduce the amount of data to be stored and transferred optimizing resources when handling large collections of music.

This chapter is structured as follows. Section 7.2 discusses previous works that consider different representations of audio for auto-tagging. Then, Section 7.3 describes previous datasets available for auto-tagging and playlist continuation. Section 7.4 describes the experiments we conducted with different architectures for auto-tagging and multiple configurations of mel-spectrograms, we also discuss the performance loss from the results of the experiment when reducing information from the audio. Section 7.5 describes the Melon Playlist Dataset. Section 7.6 describes an application of the dataset. Finally, Section 7.7 gives the conclusions.

7.2 Related Work

In image processing, there are studies that consider simplifications of CNN architectures by means of reducing network width, depth and input resolution (Tan & Le, 2019). However, only few previous studies compared different spectrogram representations for CNN architectures in MIR. Instead, it is common to focus on tuning model hyper-parameters with a fixed chosen input.

[17] The findings described in the following sections related with Melon Playlist Dataset are based on our published work presented in Ferraro et al. (2021b)

Dataset	Tracks	Tags	Playlists	Audio (official)
MTAT	5,405	188	–	30 s previews
MSD	505,216	522,366	–	–
FMA	106,574	161	–	full CC tracks
MTG-J	55,609	195	–	full CC tracks
MPD	2,262,292	–	1,000,000	some previews through API
MPSD	1,993,607	–	74,996	–
Melon Music	649,091	30,652	148,826	20-50 s mel-spectrograms

Table 7.1: Public datasets for automatic playlists continuation and auto-tagging compared to Melon Playlist Dataset. CC stands for audio available under Creative Commons licenses.

The choice of the spectrogram input is done empirically and often follows approaches previously reported in literature. Very few information comparing different inputs is available as the authors tend to only report the most successful approaches. Also, as the existing studies on music auto-tagging focus on optimizing accuracy metrics, there is a lack of works that intend to simplify networks and their inputs for computational efficiency and consider practical aspects of the efficient ways to store spectrogram representations.

To the best of our knowledge, there is no systematic comparison of mel-spectrogram representations for auto-tagging. The only work we are aware of in this direction has been done by Choi et al. (2018b), where the authors compare model performances under different pre-processing strategies such as scaling, log-compression, and frequency weighting. The same authors provide an overview of different inputs that can be used for the auto-tagging task in Choi et al. (2017a). In relation to mel-spectrograms, they suggest that one can optimize the input to the network by changing some of the signal processing parameters such as sampling rate, window size, hop size or mel bins resolution. These optimizations can help to minimize data size and train the networks more efficiently, however, no quantitative evaluations are provided.

7.3 Datasets

Table 7.1 summarizes the existing datasets for the tasks of music auto-tagging and automatic playlist continuation.

7.3.1 Datasets for Automatic Tagging of Audio

MagnaTagATune (Law et al., 2009) is one of the most used datasets for benchmarking music auto-tagging systems. It contains multi-label annotations of genre, mood and instrumentation for 25,877 audio segments. Each segment is 30 seconds long, and the dataset contains multiple segments per song. All the audio is in MP3 format with 32 Kbps bitrate and 16 KHz sample rate. The dataset is split into 16 folders, and researchers commonly use the first 12 folders for training, the 13th for validation and the last three for testing. Also, only 50 most frequent tags are typically used for evaluation. These tags include genre and instrumentation labels, as well as eras (e.g., '80s' and '90s') and moods.

The Million Song Dataset (MSD) (Bertin-Mahieux et al., 2011) is a large dataset of audio features, expanded by the MIR community with additional information including tags, lyrics and other annotations. It was previously possible to download 30-second audio previews from 7digital and collaborative tags from Lastfm, but it is no longer accessible. This subset contains 241,904 annotated track fragments and it is commonly used as another larger scale benchmark for music auto-tagging systems. The tags cover genre, instrumentation, moods and eras. Audio fragments vary in their quality, encoded as MP3 with a bitrate ranging from 64 to 128 Kbps and the sample rates of 22 KHz or 44 KHz. Another limitation of this dataset is the noise in the tags (Choi et al., 2018a).

Researchers in music auto-tagging commonly use the MagnaTagATune dataset (Law et al., 2009) to evaluate multiple settings and then repeat some settings on Million Song Dataset (Bertin-Mahieux et al., 2011) to validate differences in performances on a larger scale (Dieleman & Schrauwen, 2014; Choi et al., 2016; Pons et al., 2018). It is important to note that both datasets contain unbalanced and noisy and/or weakly-labeled annotations (Choi et al., 2018a) and therefore are challenging to work with, as the reliability of conducted evaluations may be affected (Sturm, 2012). Still, these are the two most used datasets for benchmarking due to the availability of audio.

To address the issue of open access to audio, the FMA (Defferrard et al., 2017) and MTG-Jamendo datasets (Bogdanov et al., 2019) were proposed for auto-tagging, both containing audio under Creative Commons licenses. The former is based on poorly structured music archives with inconsistent annotations and low-quality recordings. The latter tries to address this issue, focusing on a free music collection maintained for a commercial use-case, thus containing better quality audio and annotations. Yet, their content is different from commercial music platforms.

7.3.2 Datasets for Automatic Playlist Continuation

Recently the Million Playlist Dataset (Chen et al., 2018) (MPD) was released by Spotify. This dataset contains information about one million playlists created by users located in the United States. However, it does not include the tracks' audio information. Even if it may be possible to download 30-second audio previews with the Spotify API, it is unclear if it is legal to redistribute them. Also, there can be inconsistencies when trying to download audio previews in the future (e.g., due to songs changing their identifier or restricted access to some of the previews in different countries). These limitations significantly affect the reproducibility and complicate the use of MPD for audio research.

The Million Playlists Songs Dataset (Falcao & Mélo, 2017) (MPSD) combines multiple smaller datasets (Art of The Mix (McFee & Lanckriet, 2012), #nowplaying (Pichl et al., 2015), and 30Music (Turrin et al., 2015)). Similar to MPD, this dataset does not provide audio nor its representations for the songs. Since it contains playlists collected from different sources, there can be noise in the data due to song matching inconsistencies between multiple sources. Also, one of the source datasets, 30Music, was originally created for session-based recommendations instead of playlist continuation.

7.4 Auto-tagging with Reduced Mel-spectrograms

In this section, we reproduce two CNN architectures applying them on mel-spectrograms with reduced frequency and time resolution with the goal of understanding the performance loss of the models when reducing information from the audio.

7.4.1 Baseline Auto-tagging Architectures

We decided to apply two architectures that are among the best performing for the task of auto-tagging according to the existing evaluations on the MTAT and MSD datasets:

- **VGG applied for music (VGG-CNN)** (Choi et al., 2016). This architecture contains multiple layers of small-size 2D-filters as it has been adapted from the computer vision field (Simonyan & Zisserman, 2014). It is a fully-convolutional network consisting of four convolutional layers with small 3×3 filters[18] and max pooling (MP) settings presented in

[18] Number of mel bands × number of frames.

Table 7.2. The network operates on 96-bands mel-spectrograms for 29.1s audio segments, 12 KHz sample rate, 512 samples frame size and the hop size of 256 samples.

- **Musically-motivated CNN (MUSICNN)** (Pons et al., 2018). The architecture contains more filters of different shapes designed with an intention to capture musically relevant information such as timbre (38×1, 38×3, 38×7, 86×1, 86×3, 86×7) and temporal patterns (1×32, 1×64, 1×128, 1×165) in the first layer. The convolution results are concatenated and passed to three additional convolutional layers including residual connections. Original network operates on 96-bands mel-spectrograms computed on smaller 15s audio segments with 16 KHz sample rate, 512 samples frame size and 256 samples hop size.[19] It then averages tag activation scores across multiple segments of the same audio input.

For evaluation on MTAT and MSD, we use batch normalization, Adam (Kingma & Ba, 2014) as optimization method with a learning rate of 0.001 and binary cross-entropy as loss function for both architectures following their authors.

Input	**Mel-spectrogram (96×1366 × 1)**
Layer 1	Conv 3×3×128
	MP (2, 4) (output: 48×341×128)
Layer 2	Conv 3×3×384
	MP (4, 5) (output: 24×85×384)
Layer 3	Conv 3×3×768
	MP (3, 8) (output: 12×21×768)
Layer 4	Conv 3×3×2048
	MP (4, 8) (output: 1×1×2048)
Output	50×1 (sigmoid)

Table 7.2: The baseline VGG CNN model architecture.

7.4.2 Mel-spectrograms

We computed mel-spectrograms using typical setting for the MTAT dataset in the state of the art (Choi et al., 2016; Pons et al., 2018). The most common settings are 12 KHz or 16 KHz sample rate, frame and hop size of 512 and 256 samples, respectively, and Hann window function. Commonly, 96 or 128 mel bands are used, covering all frequency range below Nyquist (6 KHz and 8 KHz, respectively) and computed using Slaney's mel scale implementation (Slaney,

[19] Frame and hop size settings are confirmed in personal communication with the author.

sample rate	# mel	hop size	log type
12 KHz	128	$\times1, \times2, \times3, \times4 \times 5, \times10$	log, dB
12 KHz	96	$\times1, \times2, \times3, \times4 \times 5, \times10$	log, dB
12 KHz	48	$\times1, \times2, \times3, \times4 \times 5, \times10$	log, dB
12 KHz	32	$\times1$	log, dB
12 KHz	24	$\times1$	log, dB
12 KHz	16	$\times1$	log, dB
12 KHz	8	$\times1$	log, dB
16 KHz	128	$\times1, \times2, \times3, \times4 \times 5, \times10$	log, dB
16 KHz	96	$\times1, \times2, \times3, \times4 \times 5, \times10$	log, dB
16 KHz	48	$\times1, \times2, \times3, \times4 \times 5, \times10$	log, dB
16 KHz	32	$\times1$	log, dB
16 KHz	24	$\times1$	log, dB
16 KHz	16	$\times1$	log, dB
16 KHz	8	$\times1$	log, dB

Table 7.3: Mel-spectrograms configurations evaluated on the MTAT dataset. Hop sizes are reported relative to the reference hop size of 256 samples (e.g., $\times5$ stands for a 5 times longer hop size).

1998). To normalize the mel-spectrograms we considered two log-compression alternatives denominated as *"dB"* for $10 \cdot log_{10}(x)$ (Choi et al., 2018b) and *"log"* for $log(1 + 10000 \cdot x)$ (Dieleman & Schrauwen, 2014).

Starting with these settings, we then considered different variations in frequency and time resolutions (smaller number of mel bands and larger hop sizes). Table 7.3 shows all different spectrogram configurations that we evaluated on the MTAT dataset. Each configuration results in a different dimension of the resulting feature matrix (the number of mel bands $\times$ the number of frames). An audio segment of 29.1 seconds corresponds to 1366 and 1820 frames in the case of no temporal reduction ($\times1$) and the 12 KHz and 16kHz sample rate, respectively. In turn, the maximum reduction we considered ($\times10$) results in 137 and 182 frames.

All spectrograms were computed using Essentia[20] music audio analysis library (Bogdanov et al., 2013b). It was configured to reproduce mel-spectrograms from another analysis library used by the state of the art, LibROSA,[21] for compatibility. As a matter of interest, to have a better understanding of what information these spectrograms are able to capture, we provide a number of examples sonifying the resulting mel-spectrograms for all considered frequency

[20] https://essentia.upf.edu
[21] https://librosa.github.io

and time resolutions online.[22]

7.4.3 Baseline Architecture Adjustments

In this section we explain the changes introduced to the original model architectures presented in Section 7.4.1.

7.4.4 VGG-CNN

We try to preserve the original architecture defined in Choi et al. (2016) in terms of the size and number of filters in each layer, but we need to adjust max pooling settings since we are reducing the dimensions of the mel-spectrogram input. We report all such modifications for the VGG-CNN architecture in Table 7.4. It reports the sizes of square max-pooling windows in each layer selected accordingly to the number of mel bands and the hop size. We prioritize changes in max pooling in the latter layers when possible. We adjust the pooling size to match the input dimensions when possible, otherwise padding is applied. In the case of 16 KHz sample rate, more adjustments to VGG-CNN are necessary because, having a fixed reference hop size of 256 samples, the higher sample rate implies better temporal resolution and the larger mel-spectrograms (1820 frames).

It is important to note that if we change the resolution of the input, the 3×3 filters in VGG-CNN capture different ranges of frequency and temporal information. For example, they cover twice the mel-frequency range and a doubled time interval when using 48 mel bands and ×2 hop size. This can be an advantage, because it reduces the amount of information that the network needs to learn.

7.4.5 MUSICNN

In the original model, timbre filters' sizes in frequency are computed relative to the number of mel bands (90% and 40%). We preserve the same relation when we change this number. In our implementations we modified the segment size to 3 seconds, as we obtained slightly better results in our preliminary evaluation.[23] We keep the temporal dimension of the filters (the number of frames) intact for all considered mel-spectrograms settings.

[22] https://andrebola.github.io/EUSIPCO2020/demos

[23] Similar to suggestions by other researchers reproducing this model.

hop size	max-pooling size (time)	
	12 KHz	***16 KHz***
×1	**4, 5, 8, 8**	4, 5, 9, 10
×2	4, 5, 8, 4	4, 5, 9, 5
×3	4, 5, 8, 2	4, 5, 9, 3
×4	4, 5, 8, 2	4, 5, 9, 2
×5	4, 5, 8, 1	4, 5, 9, 2
×10	4, 5, 4, 1	4, 5, 9, 1

# mel	max-pooling size (frequency)
128	2, 4, 4, 4
96	**2, 4, 3, 4**
48	2, 4, 3, 2
32	2, 2, 3, 2
24	2, 2, 3, 2
16	2, 2, 2, 2
8	2, 2, 2, 1

Table 7.4: Adjusted sizes for max-pooling windows (time and frequency) in the four consecutive layers of the VGG CNN model with respect to the hop size, sample rate and the number of mel bands. The original sizes are highlighted in bold.

7.4.6 Evaluation Metrics for Auto-tagging

CNN models for auto-tagging output continuous activation values within $[0, 1]$ for each tag, and therefore we can study the performance of binary classifications under different activation thresholds. To this end, following previous works (Oramas et al., 2017a; Pons et al., 2018; Choi et al., 2016) we use Receiver Operating Characteristic Area Under Curve (ROC AUC) averaged across tags as our performance metric. We also report Precision-Recall Area Under Curve (PR AUC), because previous studies (Davis & Goadrich, 2006) have shown that ROC AUC can give over-optimistic scores when the data is unbalanced, which is our case. Both ROC AUC and PR AUC are single value measures characterizing the overall performance, which allows to easily compare multiple systems.

To measure the computational cost of models' training and inference we use an estimate of the number of multiply-accumulate operations required by a network to process one batch (1 GMAC is equal to 1 Giga MAC operations). This metric is related to the time a model requires for training and inference. We use an online tool[24] to compute approximate MAC values for our architectures.

[24] https://dgschwend.github.io/netscope/quickstart.html

7.4.7 Results

We evaluated the considered mel-spectrogram settings on the adjusted CNN models. Full results for all evaluated configurations are available online.[25] In Figure 7.1 we show the results of the evaluation for VGG-CNN on the MTAT dataset, repeated three times for each configuration. The first two plots show the ROC AUC results for the 12 KHz and 16 KHz sample rate using the log and dB scaling. Similarly, the third and forth plots show the PR ROC results under the same conditions. The last plot shows GMAC.

The results show that using some of the settings we can reduce the size of the input in frequency and time without affecting much the performance of VGG-CNN on the MTAT dataset. For example, if we reduce the frequency resolution from 96 to 48 mel bands we can reduce the MAC operations near 50% without affecting the performance in all configurations. Similarly, we can also reduce time resolution by 50% without affecting performance, and in this case we also reduce the MAC operations by 50% in all configurations. We can further reduce the number of operations by the cost of some performance decrease. This can be especially useful for applications requiring lightweight models, as we can get a model $\times 10$ faster by sacrificing between 1.4 and 1,8% of the performance depending on the configuration. Interestingly enough, both ROC AUC and PR AUC slightly improve when using 48 mel bands compared to 96 bands in most of the cases, however no statistically significant difference was found ($P > 0.08$ for all corresponding configurations in an independent samples t-test).

For the MUSICNN model, we have tested some of the configurations reported in Table 7.5. We only considered the frequency resolution reduction to 48 mel bands and no hop size increments due to significantly slower training time (see Section 7.4.1). The results show comparable performance of 96- and 48-band mel-spectrograms and are consistent with the above mentioned findings for the VGG-CNN model. Overall, using 128 mel bands resolution provided the best performance. Also, according to the results, the MUSICNN architecture outperforms VGG-CNN, which is consistent with the reports from the authors.

To check how our findings scale, we selected a number of configurations and re-evaluated the models on the MSD dataset. The results are reported in Table 7.6. In the case of VGG-CNN the performance of the baseline architectures is slightly superior to the ones working with lower-resolution mel-spectrograms, which comes by cost of a significantly larger computational effort. For example, for the 12 KHz sample rate, $\times$ 1 hop size and dB compression settings, reducing the number of mel bands from 96 to 48 results in the decrease is 0.16% in the ROC AUC performance and 50% reduction in GMACs. For

[25] https://andrebola.github.io/EUSIPCO2020/results

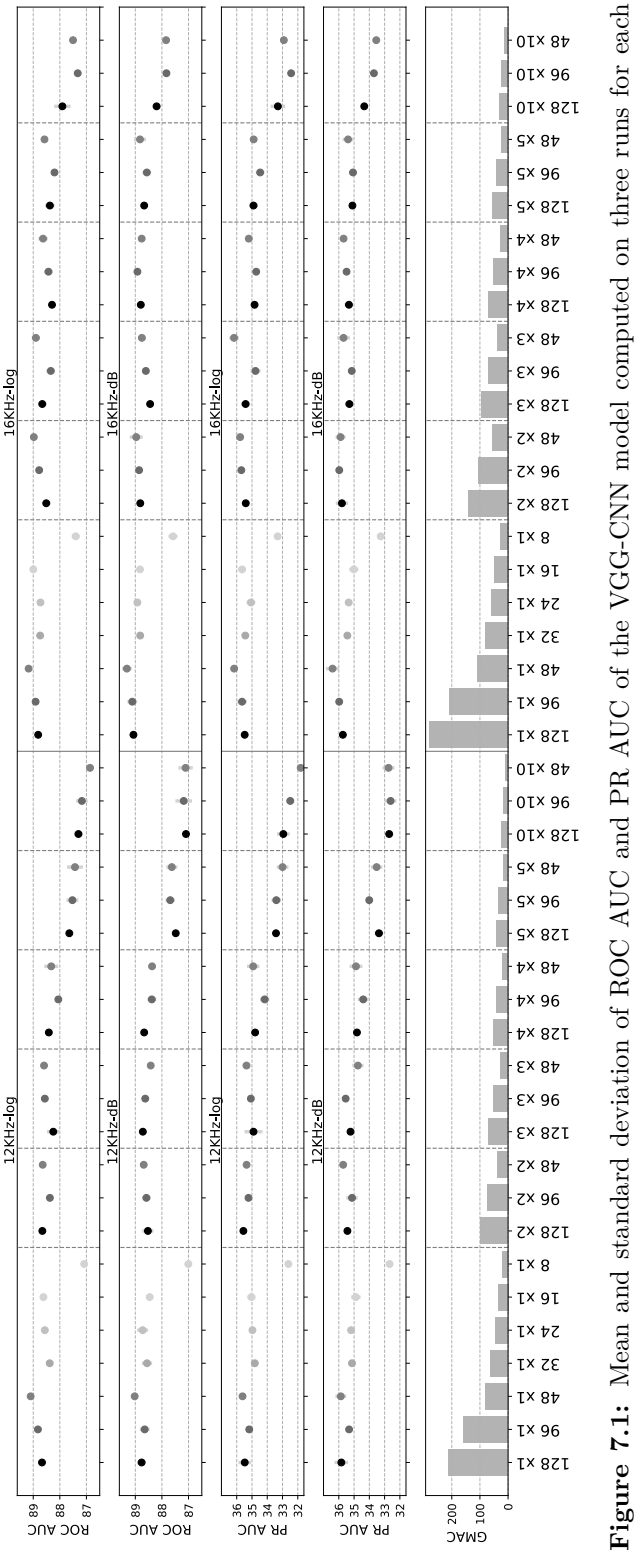

Figure 7.1: Mean and standard deviation of ROC AUC and PR AUC of the VGG-CNN model computed on three runs for each mel-spectrogram configuration (# mel, hop size, sample rate, and log type) and the associated GMAC values.

a similar 16 KHz/*dB* case the reduced model has the same performance with the benefit of twice as low computational speed. In the case of the MUSICNN architecture we see a reduction of the performance of 0.19% if we compare 96 vs 48 mel bands using 12 KHz sample rate and 0.11% for 16 KHz.

# mels	sample rate	ROC AUC	PR AUC
128	12 KHz	90.40	**38.54**
96	12 KHz	**90.50**	37.70
48	12 KHz	90.33	37.80
128	16 KHz	**90.83**	**38.92**
96	16 KHz	90.60	38.09
48	16 KHz	90.50	37.70

Table 7.5: ROC AUC and PR AUC of the MUSICNN model on the MTAT dataset for a selection of configurations using *dB* log-compression and the reference hop size ($\times$ 1).

7.5 Melon Playlist Dataset

Based on the results described in Section 7.4.7 the optimal configuration for audio representations was defined to create Melon Playlist Dataset. In this chapter, we try to overcome the limitations of the existing datasets presented in Section 7.3. Our main contribution is to provide a large research dataset of commercial music with quality playlist and tag information that includes audio representations suitable for audio-based approaches. Furthermore, our dataset is different because it represents music consumption in Korea instead of Western countries, bringing more cultural diversity in MIR research applied to music consumption platforms.

All the data was originally collected from Melon for a playlist continuation challenge that took place on the Kakao Arena[26] platform between April and July 2020 with participation of 786 teams. The dataset consists of 649,091 tracks, represented by their mel-spectrograms, and 148,826 playlists with annotations by 30,652 different tags. The playlists were created and annotated by selected users recognized for the quality of their submissions. These users are named Melon DJs on the platform after Melon moderators verify them for the quality of the playlist metadata (titles, tags, and genres) they provide.

To reduce distributable data size, we computed mel-spectrograms only for a segment of each song (20 to 50 seconds long, not adjacent to the start or the end of the songs). Furthermore, for copyright reasons and based on the results

[26] https://arena.kakao.com/c/8

# mels	hop size	sample rate	ROC AUC	PR AUC
128	×1	12 KHz	86.48	27.56
96	×1	12 KHz	**86.67**	**27.70**
48	×1	12 KHz	86.53	27.27
128	×2	12 KHz	86.28	27.24
96	×2	12 KHz	86.18	26.93
48	×2	12 KHz	85.86	26.42
128	×1	16 KHz	**86.84**	**28.10**
96	×1	16 KHz	86.71	28.06
48	×1	16 KHz	86.73	27.78
128	×2	16 KHz	86.34	27.06
96	×2	16 KHz	86.63	27.70
48	×2	16 KHz	86.41	26.83

(a) VGG-CNN

# mels	hop size	sample rate	ROC AUC	PR AUC
128	×1	12 KHz	87.10	26.97
96	×1	12 KHz	87.16	**27.10**
48	×1	12 KHz	86.99	26.66
128	×1	16 KHz	**87.21**	26.91
96	×1	16 KHz	**87.21**	26.96
48	×1	16 KHz	87.10	26.64

(b) MUSICNN

Table 7.6: ROC AUC and PR AUC of the models on the MSD dataset for a selection of configurations using *dB* log-compression.

Property	Count
Track-playlist relations	5,904,718
Unique tracks	649,091
Tag-playlist relations	516,405
Unique tags	30,652
Playlists	148,826
Playlist titles	121,485
Unique playlist titles	116,536
Artists	107,824
Albums	269,362
Genres	30

Table 7.7: Melon Playlist Dataset statistics.

described in Section 7.4.7 we used a reduced 48 mel-bands resolution (with 16 KHz sample rate, frame and hop size of 512 and 256 samples, and Hann window function). This configuration did not negatively affect the performance of the auto-tagging approaches, while having a significantly lower reconstructed audio quality. These decisions allow saving bandwidth and disk space required to transfer and store the dataset. The dataset is distributed in 40 files, 6 GB each, with a total download size of 240 GB.

The dataset also includes playlist and tracks metadata. Playlist metadata contains tags and titles submitted by playlist creators, the number of users who like the playlist, and the last modification date. Track metadata contains album, title, artists, release date, and genres. The statistics of the dataset are presented in Table 7.7.

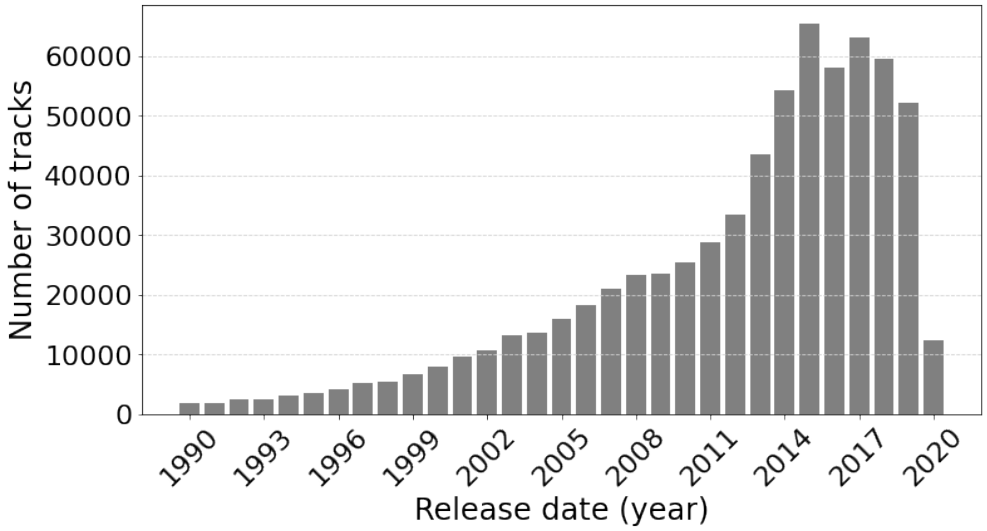

Figure 7.2: The distribution of release year of all tracks.

Figure 7.2 shows the distribution of the tracks concerning their release year. Over 95% of the tracks in the dataset were published after the year 1990. Considering genre annotations, 25.45% tracks in the dataset belong to only Korean music genres, 38.44% tracks to non-Korean music genres, and 27.70% tracks to both Korean and non-Korean genres (8.39% tracks are annotated with music genres of which the origin is unknown).

Playlists contain up to 200 tracks, with 41.46 tracks on average. The average of tags per playlist is 3.91 with a maximum of 11 tags. The number of different genres in a playlists on average is 6.31 with a maximum of 26. Figure 7.3 shows the distribution of number of tags, genres and tracks in the playlists.

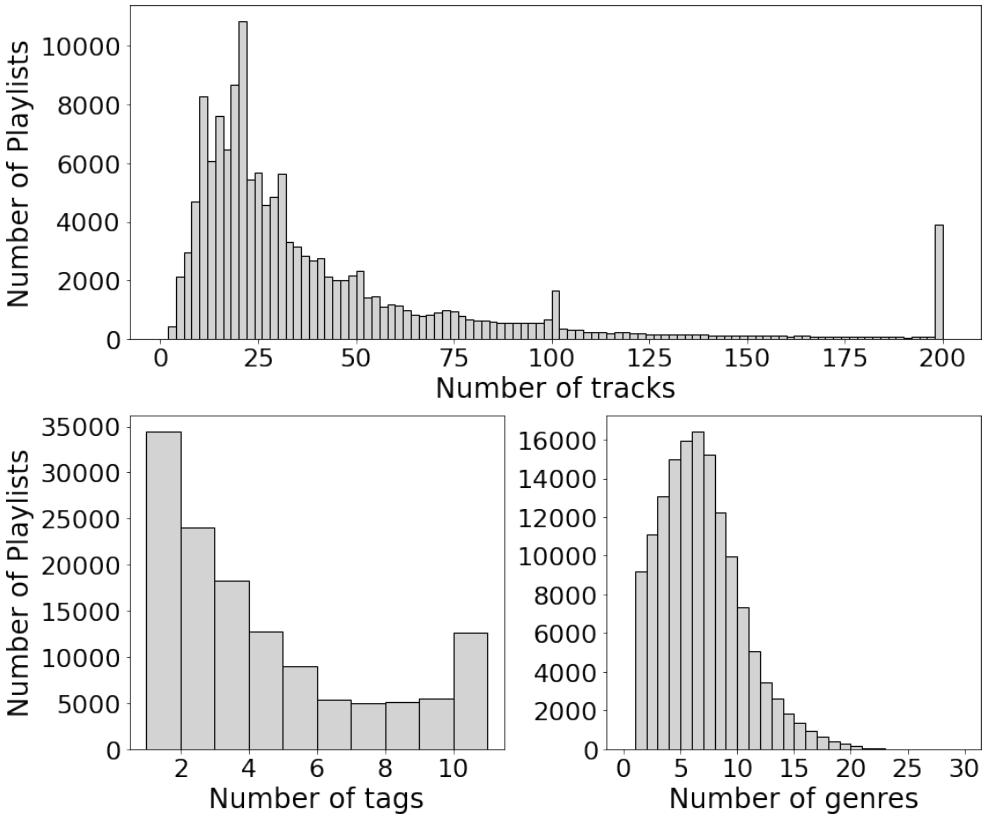

Figure 7.3: Number of tracks, tags, and genres in playlists.

7.5.1 Kakao Arena Challenge and the Dataset Split

In the context of the challenge, the playlists where divided in three groups: 115,071 playlists (77.32%) in the train set, 23,015 playlists (15.46%) in the validation set, and 10,740 playlists (7.22%) in the test set. For the 33,755 playlists in validation and test sets, we considered either fully or partially hiding the tags, titles and tracks metadata. Table 7.8 shows the total number of playlists for each of these problem cases. The goal of the challenge was to predict the missing tracks and tags for the playlists in the test set.

Even though the challenge has finished, the Kakao Arena evaluation platform remains open for submissions of the predicted tracks and tags for the APC and auto-tagging tasks. In this way, it offers the possibility to the research community to benchmark new approaches in a standardized way using the test set with hidden tracks and tags.

Tracks	Tags	Title	Frequency
all	half	half	3860 (11.43%)
half	all	half	0 (0.00%)
half	half	all	13165 (39.00%)
all	all	half	2554 (7.56%)
all	half	all	2 (0.00%)
half	all	all	14168 (41.97%)
all	all	all	6 (0.01%)

Table 7.8: Number of playlists in test and validation sets for which the tracks, tags and title were hidden either entirely ("all") or for the half of the instances ("half").

7.6 Automatic Playlist Continuation

Melon Playlist Dataset offers many research possibilities. The most direct are playlists generation and auto-tagging for which it was originally created.

The task of APC consists on recommending a list of tracks to continue a given playlist. Many approaches had been proposed for this task including collaborative and content-based (Schedl et al., 2018; Zamani et al., 2019). Collaborative filtering approaches usually offer the best performance according to offline metrics in the task of track recommendations to users. Given that it is not possible to recommend items without any previous interaction with these approaches (the cold-start problem), in the last years deep learning approaches have been proposed to overcome this problem by predicting the collaborative representations from audio (Liang et al., 2015; Van den Oord et al., 2013). Melon Playlist Dataset is the first public dataset to contain playlist information together with directly available audio information of the tracks on a large scale, allowing to experiment with such audio-based approaches.

In what follows, we provide an example of an audio-based APC approach, allowing us to expand a playlist with previously unseen tracks. We focus on underrepresented tracks in our evaluation, which is different from the Kakao Arena challenge, where the tracks in the test set had significantly more associated track-playlist interactions available for collaborative filtering. For this reason, and for reproducibility outside the Kakao Arena platform, we create an alternative split.

7.6.1 Method

We created a subset of Melon Playlist Dataset, discarding the playlists with less than 5 tracks. For each playlist we split its track-playlist interactions, using the tracks that appear at least in 10 playlists for our training set (*APC-train*)

and the rest of the tracks (considered cold-start tracks) for testing (*APC-test*). The *APC-train* subset contains interactions for a total of 104,645 playlists and 81,219 tracks.

Similar to Van den Oord et al. (Van den Oord et al., 2013), we train a Matrix Factorization (MF) model on the APC-train track-playlist matrix using WARP loss function Weston et al. (2011) and optimizing the parameters on 10% of the training interactions.

The MF model outputs the latent factors of the tracks and playlists in APC-train, we train an audio model to predict these track factors from mel-spectrograms provided in the streaming-platform. To this end, we split the tracks in APC-train into *APC-train-train* (90%) for training and *APC-train-val* (10%) for validation. We use a fully-convolutional neural network common for auto-tagging, based on VGGish architecture (Choi et al., 2017b) and trained with Mean Squared Error (MSE) as a loss function. We observed reasonable approximation of the CF track factors by the audio model, with the MSE of 0.0098.

Once trained, we apply the model to predict latent factors for the cold-start tracks in APC-test and match those factors to the playlist factors (Zamani et al., 2019) in APC-train to generate rankings of the best tracks to expand those playlists. We evaluate the top-10 and top-200 rankings using MAP and nDCG (Ricci et al., 2010) and the rest of playlist-track interactions kept as ground truth in *APC-test* for the playlists.

Method	MAP@10	nDCG@10	MAP@200	nDCG@200
Random	0.0000	0.0001	0.0001	0.0010
Audio	0.0159	0.0395	0.0135	0.0516
CF	0.0165	0.0414	0.0148	0.0545

Table 7.9: Performance on APC-train-val.

7.6.2 Results

In all evaluations we compare the audio approach to the random baseline and the collaboration filtering approach used as our lower-bound and upper-bound baselines, respectively. Table 7.9 shows the performance on the validation set (APC-train-val). Comparing the performance of latent factors predicted from audio with the ones from the MF model itself, we see that the performance of both is very similar, which shows that the audio-based approach can be used to predict latent factors for unseen tracks.

For the collaborative filtering baseline on APC-test, we use all interactions in

Test subset	Track in # playlist	Tracks	Playlists
APC-test-1	8-9	17,042	27,229
APC-test-2	5-8	46,069	35,910
APC-test-3	2-5	155,688	31,925

Table 7.10: Track frequency based subsets of the APC-test set.

	Method	MAP@10	nDCG@10	MAP@200	nDCG@200
APC-test	Random	0.0000	0.0000	0.0000	0.0002
	Audio	0.0007	0.0014	0.0010	0.0052
	CF	0.0802	0.1338	0.0581	0.1099
APC-test-1	Random	0.0001	0.0003	0.0003	0.0022
	Audio	0.0041	0.0065	0.0063	0.0267
	CF	0.0846	0.1200	0.0979	0.1923
APC-test-2	Random	0.0000	0.0000	0.0001	0.0009
	Audio	0.0022	0.0038	0.0032	0.0136
	CF	0.0490	0.0745	0.0582	0.1291
APC-test-3	Random	0.0000	0.0000	0.0000	0.0002
	Audio	0.0001	0.0001	0.0001	0.0002
	CF	0.0274	0.0416	0.0341	0.0756

Table 7.11: Performance on APC-test.

APC-train together with 70% of the interactions in the APC-test to train the MF model and the other 30% to evaluate. Some test tracks are discarded from evaluation due to this split. For consistency, we use the same set of test tracks for evaluation of the rest of the approaches.

Table 7.11 shows the overall performance using all considered tracks in APC-test for ranking. In addition, we independently evaluated three subsets of APC-test described in Table 7.10, generating separate ranking lists among the tracks with different popularity (or "cold-startness") level in the dataset. The results on these subsets are given as an additional reference, but they aren't directly comparable as the performance is measured on ranking lists of different track sets.

7.7 Conclusions

In this chapter, we first studied how different mel-spectrogram representations affect the performance of CNN architectures for music auto-tagging. We have compared the performances of two state-of-the-art models when reducing the mel-spectrogram resolution in terms of the number of frequency bands and frame rates. We used the MagnaTagaTune dataset for comprehensive performance comparisons and then we compared selected configurations on the larger Million Song Dataset. The results suggest that is possible to preserve a similar performance while reducing the size of the input. They can help researchers and practitioners to make trade-off decision between the accuracy of the models, data storage size and training and inference time, which are crucial in many applications. All the code to reproduce this study is open-source and available online.[27]

Based on the previous findings, we presented Melon Playlist Dataset, the first public large-scale dataset of commercial music including the playlists, audio representation, and tags altogether, submitted by users verified for their quality annotations. We describe how the dataset is composed in terms of playlists, tracks, year, genres and tag annotations. Since the dataset reflects the music consumption in Korea, it offers novel opportunities to diversify MIR research.

The dataset has various applications. As an example, we considered automatic playlist continuation in a cold-start scenario and trained a baseline model to predict the latent factors of collaborative filtering from mel-spectrograms. All the code to reproduce this experiment, including the generation of dataset splits, is available online.[28]

[27] https://andrebola.github.io/EUSIPCO2020/
[28] https://github.com/andrebola/icassp2021

CHAPTER 8

Enriched Music Representation Using Multi-Modal Contrastive Learning

8.1 Introduction

From the conclusions presented in Chapter 3, better methods for recommending new and less-popular music are identified to be important for music artists to make the music platforms fair from their perspective. In addition, artists also express that music platforms should consider and take into account the context of the music when making recommendations. Learning representations of the music is essential for multiple tasks such as music recommendations or automatic tagging. Methods allowing to obtain a representation from multiple types of data related to the music (such as text, audio and image) have the possibility to capture complementary information from the different modalities, which is more aligned with the artists' interest in considering the music context when generating recommendations.

There are multiple sources and types of information related to the music that can be used for different applications, e.g., audio was shown to give better performance to predict the genre (Won et al., 2020c), users' listening data give higher user satisfaction when generating recommendations (Celma & Herrera, 2008) and also to predict the mood of the songs (Korzeniowski et al., 2020). Having a numerical feature representation that combines all the relevant information of a song would allow creating better automatic tools to solve these problems.

Advances in deep learning in the past years enabled improvement in the performance of multiple tasks by combining different types of data. For example, Oramas et al. (2018a) propose a multi-modal approach combining text, audio, and images for music auto-tagging and Surís et al. (2018) propose a method to combine audio-visual embeddings for cross-modal retrieval.

Deep learning allows learning representations mapping from different input data to an embedding space that can be applied in multiple downstream tasks (Radford et al., 2015). With this goal, the most common approach in the music domain is to train a classifier and use the pre-trained model to obtain embeddings that could be used in different tasks. Alonso-Jiménez et al. (2020) compare different pre-trained architectures for predicting multiple aspects of a song such as danceability, mood, gender and timbre, showing the generalization capabilities of these pre-trained models. Recent deep metric learning approaches have shown a better performance across multiple downstream tasks compared to the approach of pre-training the model for a classification task (Zhai & Wu, 2018; Lee et al., 2020a), which suggests that they generalize better for unseen data. Similarly, Cramer et al. (2019b) propose a self-supervised learning approach combining audio and video producing embeddings that show improved performance in multiple downstream tasks.

Contrastive learning has gained popularity in the last years (Le-Khac et al., 2020). These approaches allow learning representation by employing a metric learning objective, contrasting similar and dissimilar items. The similar items are referred to as positive examples and the dissimilar items are referred to as negative examples. Approaches based on triplet loss (Weinberger & Saul, 2009) are popular to learn content features. They require to define triplets composed of an anchor, a positive and a negative example. Triplet loss was recently applied in the music domain for retrieval (Won et al., 2020c) and zero-shot learning (Choi et al., 2019). However, the strategy for sampling the triplets is crucial to the learning process and can require significant effort.

There are other losses that instead of defining triplets rely on the comparison of paired examples such as *infoNCE* (Van den Oord et al., 2018) and *NT-Xent* (Chen et al., 2020). They have the advantage of involving all the data points within a mini-batch when training without requiring to define a specific strategy for sampling the training examples. Employing these contrastive loss functions in a self-supervised way has led to powerful image (Chen et al., 2020), sound (Fonseca et al., 2020) and music (Saeed et al., 2020) representations learned without the need for annotated data.

Contrastive learning was also applied in a supervised way (Favory et al., 2020a; Khosla et al., 2020) with a cross-modal approach using sound information and associated text metadata in order to learn semantically enriched audio features. The learned features achieve competitive performance in urban sound events

identification and musical instrument recognition (Favory et al., 2020a).

From the works mentioned above, it is clear that methods based on contrastive learning show good results in multiple domains and have the potential to exploit heterogeneous data that can lead to improved performance in different tasks. However, we are not aware of any previous work that focuses on the alignment of multiple modalities of information based on contrastive learning to exploit heterogeneous data in the music domain.

Motivated by such promising results, in this chapter we study an approach that takes advantage of different types of music-related information (i.e. audio, genre, and playlist relations of the tracks) to obtain representations from the audio that can perform well in multiple downstream tasks such as music genre classification, automatic playlist continuation, and music automatic tagging. In summary: i) We propose updated audio and text encoders optimized for the music domain based on the architecture proposed by Favory et al. (2020b); ii) We use the alignment of multi-modal data for exploiting the semantic metadata and collaborative filtering information; iii) We evaluate the obtained representations in three downstream tasks using different datasets comparing with other common approaches based on CNNs; iv) We also include an ablation study by comparing the performance of each source of information independently, which allows us to understand the importance of the different parts of our model.[29]

The rest of this chapter is structured as follows. Section 8.2 describes our method based on contrastive learing. Section 8.3 describes the three downstream tasks on which the model was evaluated and the dataset used in each case. Section 8.4 shows the results of each downstream task. Section 8.5 describes a web-based application built to demonstrate the outcome of our contrastive learning method. Finally, Section 8.6 gives the conclusions.

8.2 Proposed Method

As we illustrate in Figure 8.1, our method employs three encoders: $e_{\mathrm{a}}(\cdot)$, $e_{\mathrm{w}}(\cdot)$, and $e_{\mathrm{cf}}(\cdot)$, and a dataset $\mathbb{D} = \{(\mathbf{X}_{\mathrm{a}}, \mathbf{X}_{\mathrm{w}}, \mathbf{x}_{\mathrm{cf}})^m\}_{m=1}^{M}$, of M paired examples. Each of the paired examples consists of a sequence of T_{a} vectors with F_{a} features of music signals, $\mathbf{X}_{\mathrm{a}}^m \in \mathbb{R}^{T_{\mathrm{a}} \times F_{\mathrm{a}}}$, a set of musical genres embeddings with T_{w} vectors of F_{w} features, $\mathbf{X}_{\mathrm{w}}^m \in \mathbb{R}^{T_{\mathrm{w}} \times F_{\mathrm{w}}}$, and a vector of F_{cf} features correlating each music genre in $\mathbf{X}_{\mathrm{w}}^m$ with a human created playlist, $\mathbf{x}_{\mathrm{cf}}^m \in \mathbb{R}_{\geq 0}^{1 \times F_{\mathrm{cf}}}$.

Three latent representations are obtained from each element of the paired examples, and these representations are aligned with each other using three

[29] The findings described in the following sections are based on our published work presented in Ferraro et al. (2021a)

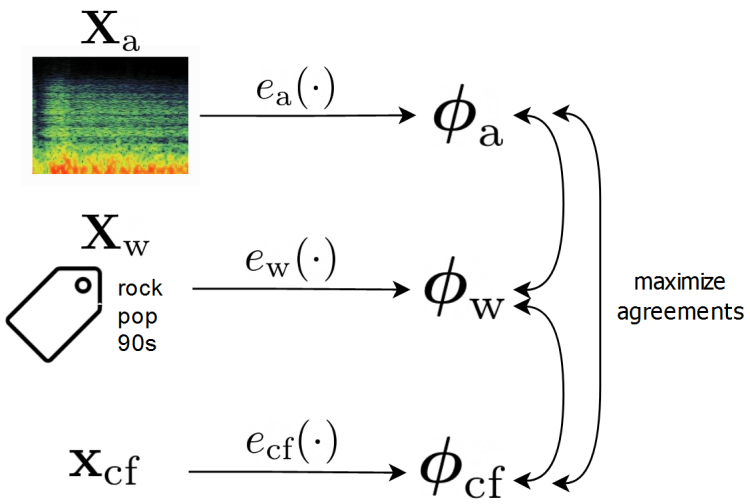

Figure 8.1: Diagram with architecture of the method

contrastive losses between paired and non-paired examples. Through the joint minimization of the contrastive losses, we obtain the optimized audio encoder $e_\mathrm{a}^\star$ which can reflect genre information and can be efficiently used for recommending next songs in a playlist.

8.2.1 Obtaining the Latent Representations

The audio encoder e_a consists of Z cascaded 2D-CNN blocks ($\mathrm{2DCNN}_z$), and a feed-forward block (FFN). Each $\mathrm{2DCNN}_z$ consists of a 2D convolutional neural network (CNN_z) with a square kernel of size K_z and unit stride, a batch normalization process (BN), a rectified linear unit (ReLU), and a pooling operation (PO). The FFN consists of a feed-forward layer ($\mathrm{FF}_{\mathrm{a}1}$), another BN process, a ReLU, a dropout with probability p, another feed-forward layer ($\mathrm{FF}_{\mathrm{a}2}$), and a layer normalization (LN) process. e_a takes as an input $\mathbf{X}_\mathrm{a}^m$ and the Z 2D-CNN blocks and the feed-forward block process the input in a serial way. The output of e_a is the learned representation $\phi_\mathrm{a}^m = e_\mathrm{a}(\mathbf{X}_\mathrm{a}^m)$, computed as

$$\mathbf{H}_z^m = \mathrm{2DCNN}_z(\mathbf{H}_{z-1}^m), \text{ and} \tag{8.1}$$

$$\phi_\mathrm{a}^m = \mathrm{FFN}(\mathbf{H}_Z^m), \text{ where} \tag{8.2}$$

$$\mathrm{2DCNN}_z(u) = (\mathrm{PO} \circ \mathrm{ReLU} \circ \mathrm{BN} \circ \mathrm{CNN}_z)(u), \tag{8.3}$$

$$\mathrm{FFN}(u) = (\mathrm{LN} \circ \mathrm{FF}_{\mathrm{a}2} \circ \mathrm{DP} \circ \mathrm{ReLU} \circ \mathrm{BN} \circ \mathrm{FF}_{\mathrm{a}1})(u), \tag{8.4}$$

$\mathbf{H}_0^m = \mathbf{X}_\mathrm{a}^m$, and $\circ$ is the function composition symbol, i.e. $(f \circ g)(x) = f(g(x))$.

The encoder $e_\mathrm{w}(\cdot)$ is the genre encoder and consists of a self-attention (SA) over the input sequence, a feed-forward layer (FF_w), DP with probability p, an LN process, and a skip connection between the input of the feed-forward

layer and its output. $e_{\mathrm{w}}(\cdot)$ is after the self-attention mechanism employed in the Transformer model (Vaswani et al., 2017), and is used to learn a contextual embedding of its input, similarly to Favory et al. (2020b). Each musical genre associated with $\mathbf{X}_{\mathrm{a}}^{m}$ is first one-hot encoded and then given as an input to a pre-optimized word embeddings model. The output of the word embeddings model is the sequence of embeddings $\mathbf{X}_{\mathrm{w}}^{m}$, which is then given as an input to $e_{\mathrm{w}}(\cdot)$. The output of $e_{\mathrm{w}}(\cdot)$ is the vector $\boldsymbol{\phi}_{\mathrm{w}}^{m} = e_{\mathrm{w}}(\mathbf{X}_{\mathrm{w}}^{m})$, containing the contextual embedding of $\mathbf{X}_{\mathrm{w}}^{m}$ and calculated as

$$\mathbf{V}'^{m} = \mathrm{SF}(\mathbf{X}_{\mathrm{w}}^{m}), \tag{8.5}$$

$$\mathbf{V}^{m} = \mathbf{V}'^{m} + (\mathrm{DP} \circ \mathrm{FF}_{\mathrm{w}})(\mathbf{V}'^{m}), \text{ and} \tag{8.6}$$

$$\boldsymbol{\phi}_{\mathrm{w}}^{m} = \mathrm{LN}(\sum_{i=1}^{T_{\mathrm{w}}} \mathbf{V}_{i}^{m}), \tag{8.7}$$

where $\mathbf{V}'^{m}, \mathbf{V}^{m} \in \mathbb{R}^{T_{\mathrm{w}} \times F'_{\mathrm{w}}}$.

The third encoder, $e_{\mathrm{cf}}(\cdot)$, is the playlist association encoder and consists of a feed-forward block, similar to $e_{\mathrm{a}}(\cdot)$, Specifically, $e_{\mathrm{cf}}(\cdot)$ consists of a feed-forward layer, $\mathrm{FF}_{\mathrm{cf1}}$, a ReLU, a dropout process with probability p, another feed-forward layer, $\mathrm{FF}_{\mathrm{cf2}}$, and a LN process. The input to $e_{\mathrm{cf}}(\cdot)$ is a vector, $\mathbf{x}_{\mathrm{cf}}^{m}$, obtained by a collaborative filtering (CF) process, using M_{pl} playlists created by humans.

The CF process gets as input a binary matrix, $\mathbf{B}_{\mathrm{cf}} \in \{0,1\}^{M \times M_{\mathrm{pl}}}$, that indicates which songs are included on which playlist. Then, by minimizing the WARP loss (Weighted Approximate-Rank Pairwise loss) using SGD and the sampling technique defined in Weston et al. (2011), to approximate ranks between playlists and songs efficiently. CF outputs the matrices $\mathbf{X}_{\mathrm{cf}} \in \mathbb{R}_{\geq 0}^{M \times F_{\mathrm{cf}}}$ and $\mathbf{Q}_{\mathrm{cf}} \in \mathbb{R}_{\geq 0}^{F_{\mathrm{cf}} \times M_{\mathrm{pl}}}$, where $\mathbf{B}_{\mathrm{cf}} \simeq \mathbf{X}_{\mathrm{cf}} \cdot \mathbf{Q}_{\mathrm{cf}}$. It follows that each $\mathbf{x}_{\mathrm{cf}}^{m}$ is a vector from $\mathbf{X}_{\mathrm{cf}}$. We employ $e_{\mathrm{cf}}(\cdot)$ to process the $\mathbf{x}_{\mathrm{cf}}^{m}$, by providing a representation of $\mathbf{x}_{\mathrm{cf}}^{m}$ that is learned specifically for the alignment process that our method tries to achieve. This practice typically employed in many similar tasks where an extra learned projection of learned representations is employed, like Van den Oord et al. (2013) and Favory et al. (2020a). The output of $e_{\mathrm{cf}}(\cdot)$ is the vector $\boldsymbol{\phi}_{\mathrm{cf}}^{m} = e_{\mathrm{cf}}(\mathbf{x}_{\mathrm{cf}}^{m})$, calculated as

$$\boldsymbol{\phi}_{\mathrm{cf}}^{m} = (\mathrm{LN} \circ \mathrm{FF}_{\mathrm{cf2}} \circ \mathrm{DP} \circ \mathrm{ReLU} \circ \mathrm{FF}_{\mathrm{cf1}})(\mathbf{x}_{\mathrm{cf}}^{m}). \tag{8.8}$$

8.2.2 Optimization and Alignment of Latent Representations

We jointly optimize all encoders using $\mathbb{D}$ and three contrastive losses. We expand previous approaches on audio representation learning using multi-modal alignment, by employing multiple cross-modal and single modal alignment processes. Specifically, we align $\boldsymbol{\phi}_{\mathrm{a}}^{m}$ with $\boldsymbol{\phi}_{\mathrm{w}}^{m}$ (audio-to-genre, A2G, alignment), $\boldsymbol{\phi}_{\mathrm{a}}^{m}$

with ϕ_{cf}^m (audio-to-playlist, A2P, alignment), and ϕ_{cf}^m with ϕ_{cf}^m (genre-to-playlist, G2P, alignment).

We use A2G alignment so that ϕ_{a}^m will be able to keep information about musical genre. Additionally, we further enhance the information in ϕ_{a}^m by the A2P alignment, which is targeted to allow ϕ_{a}^m to have information about playlist associations. Finally, we employ G2P alignment, so that we keep genre and playlist related information tied up together and not let them degenerate to some representation that just helps to minimize the employed losses. Specifically, we use the contrastive loss between two paired examples ψ_α and ψ_b

$$\mathcal{L}_{\psi_\alpha,\psi_b} = \sum_{i=1}^{M} -\log \frac{\Xi(\psi_\alpha^i, \psi_b^i, \tau)}{\sum_{k=1}^{2M} \mathbb{1}_{[k\neq i]} \Xi(\psi_\alpha^i, \zeta^k, \tau)}, \text{ where} \tag{8.9}$$

$$\Xi(\mathbf{a}, \mathbf{b}, \tau) = \exp(\text{sim}(\mathbf{a}, \mathbf{b})\tau^{-1}), \tag{8.10}$$

$$\text{sim}(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top \mathbf{b}(||\mathbf{a}||\,||\mathbf{b}||)^{-1}, \tag{8.11}$$

$$\zeta^k = \begin{cases} \psi_a^k, \text{ if } k \leq M \\ \psi_b^{k-M} \text{ else} \end{cases}, \tag{8.12}$$

$\mathbb{1}_{\text{A}}$ is the indicator function with $\mathbb{1}_{\text{A}} = 1$ iff A else 0, and τ is a temperature hyper-parameter.

We identify as the loss for A2G alignment the $\mathcal{L}_{\text{A2G}} = \mathcal{L}_{\phi_{\text{a}},\phi_{\text{w}}} + \mathcal{L}_{\phi_{\text{w}},\phi_{\text{a}}}$, as the loss for A2P alignment the $\mathcal{L}_{\text{A2P}} = \mathcal{L}_{\phi_{\text{a}},\phi_{\text{cf}}} + \mathcal{L}_{\phi_{\text{cf}},\phi_{\text{a}}}$, and as the loss for G2P alignment the $\mathcal{L}_{\text{G2P}} = \mathcal{L}_{\phi_{\text{w}},\phi_{\text{cf}}} + \mathcal{L}_{\phi_{\text{cf}},\phi_{\text{w}}}$. We optimize all of our encoders and we obtain $e_{\text{a}}^\star$ by minimizing the

$$\mathcal{L}_{\text{tot}} = \lambda_{\text{A2G}}\mathcal{L}_{\text{A2G}} + \lambda_{\text{A2P}}\mathcal{L}_{\text{A2P}} + \lambda_{\text{G2P}}\mathcal{L}_{\text{G2P}}, \tag{8.13}$$

where $\lambda_{\cdot}$ are different hyper-parameters used as weighting factors for the losses.

8.3 Evaluation

To evaluate our method, we employ Melon Playlist Dataset as $\mathbb{D}$, in order to obtain e_{a}. Then, we assess the learned representations by e_{a}, employing it in different downstream tasks. Specifically, we focus on genre classification, audio-tagging, and automatic playlist continuation. For each of the tasks, we employ e_{a} as audio encoder, which will provide embeddings to a classifier, trained for the corresponding task.

We compare the performance of the proposed approach described in Section 8.2 using only Genre (Contr_{G}), only CF information (Contr_{CF}) and combining

genre with CF information ($\text{Contr}_{\text{CF-G}}$). In addition, we compare the performance on each task using a baseline architecture that directly predicts the target information from the audio encoder. We refer to these methods as B-line_{G} for the model trained with genre information, $\text{B-line}_{\text{CF}}$ for the model trained to predict CF information and $\text{B-line}_{\text{CF-G}}$ for the model trained to predict both information at the same time.

8.3.1 Melon Playlist Dataset and Audio Featrures

The dataset $\mathbb{D}$ used to train the models is the Melon Playlist Dataset, described in Chapter 7. The dataset consists of M=649,091 tracks, represented by their mel-spectrograms, and M_{pl}=148,826 playlists. In order to train the model we split the songs of the dataset in train (80%), validation (10%) and test (10%). The split was done applying a stratified approach (Sechidis et al., 2011) in order to assure a similar distribution of example in all the sets for the genres associated to the songs.

The pre-computed mel-spectrograms provided in the dataset correspond to a range of 20 to 50 seconds with a resolution of $F_{\text{a}} = 48$ mel-bands. Such reduced mel-bands resolution did not negatively affect the performance of the auto-tagging approaches and have a significantly lower quality of reconstructed audio, as shown in Chapter 7, allowing us to avoid copyright issues. Following the previous work (Won et al., 2020b), we randomly select sections the songs to train the audio encoder, using $T_{\text{a}} = 256$.

8.3.2 Parameters Optimization

Following the best performance in previous work (Won et al., 2020c,b) the audio encoder use Z=7 layers and K=3. We conducted a preliminary evaluation to select the hyper-parameters of the models, comparing the loss in the validation and training set to prevent the models of overfitting. We defined the dimensions for CF representations to F_{cf}= 300 and genres representations F_{w}= 200 with T_{w}= 10 genres per song. From the same preliminary evaluation we defined the temperature τ=0.1, batch size of 128 , learning rate of 1e-4, dropout of 0.5 and number of heads for self-attention of 4. We did not experiment with changing the weights λ for the different losses and we used $\lambda_{\text{A2G}} = \lambda_{\text{A2P}} = \lambda_{\text{G2P}} = 1$.

8.3.3 Downstream Tasks

Once the models are trained with the Melon Playlist Dataset, we use the pre-trained models to generate an embedding from the audio of each song in the

different datasets. Then, we use the generated embeddings and compare the performance for each particular task.

In summary, we consider various downstream tasks to evaluate the performance of the models:

- Genre classification from audio using GTZAN dataset
- Auto-tagging in three different categories of tags (Mood/theme, instruments, and genre) using MTG-Jamendo dataset
- Automatic playlist continuation using Melon Playlist Dataset.

In the following, we describe each downstream task and the dataset used.

8.3.4 Genre Classification

Following recent work on representation learning for music signals (Alonso-Jiménez et al., 2020), we adopt genre classification as one of the downstream tasks.

We use the fault-filtered version of the GTZAN dataset (Tzanetakis & Cook, 2002; Kereliuk et al., 2015) consisting of music excepts of 30 seconds, single-labeled using 10 classes and split in pre-computed sets of 443 songs for training and 290 for testing. The GTZAN dataset is common for benchmark music genre classification algorithms in MIR since defines a train, validation and test set. We train a multilayer perceptron (MLP) of one hidden layer of size 256 with ReLU activations, using the training set and compute its accuracy on the test set. In order to obtain an unbiased evaluation, we repeat this process 10 times and average the accuracies. We consider each embedding frame of a track as a different training instance, and when inferring the genres, we apply a majority voting strategy. We also include the performance of pre-trained embedding models taken from the literature (Cramer et al., 2019b; Pons & Serra, 2019; Gemmeke et al., 2017), using the results reported in Pons & Serra (2019).

8.3.5 Automatic Tagging

We rely on the MTG-Jamendo dataset (Bogdanov et al., 2019) which was built using audio data and metadata from Jamendo and made available under Creative Commons licenses, which is different than typical music present in commercial music platforms. The dataset contains over 55,000 full audio tracks multi-labeled using 195 different tags from *genre*, *instrument*, and *mood/theme*

categories. [30] For this task, we train a MLP that takes our pre-trained audio embeddings as input. We compute the embedding of all the tracks by averaging their embeddings computed on non-overlapping frames with the mean statistic. The model is composed of two hidden layers of size 128 and 64 with ReLU activations, it includes batch normalizations after each layer and a dropout regularization after the penultimate layer. We use the validation sets for early stopping and we finally evaluate the performances on the test sets using ROC AUC. These evaluations are done on the three separated category of tags, each of them uses its own split. We repeat the procedures 10 times and report the mean average.

The tags for this task are divided in three different types: Mood/theme, instruments, and genre.[31]

8.3.6 Playlist Continuation

We make use of the playlists from the Melon Playlist Dataset that contain at least one track in our test set (not used when training our embedding model). This provides 104,410 playlists, for the which we aim at providing 100 continuation tracks. We compute the embedding of all the tracks by averaging their embeddings computed on non-overlapping frames with the mean statistic. Then, for each track in a playlist, we compute the 100 most similar tracks, among the ones from the test set. These tracks are obtained using the cosine similarity in the embedding space computed using Annoy [32] which is based on Approximate Nearest Neighbors (Dasgupta & Freund, 2008). Among all the retrieved similar tracks for a playlist, we finally select the 100 most repeated ones. We compare these to the ground truth using normalized Discounted Cumulative Gain (nDCG) and Mean Average Precision (MAP) (Ricci et al., 2010), which are commonly used to evaluate the performance of music recommendation systems. These ranking metrics evaluate the order of the items for each playlist returned by the prediction. They return a higher score for a given playlist if the predicted ranked list contains items in the test set closer to the top. Their main difference is that nDCG considers the order that is provided in the ground truth, whereas MAP does not.

8.4 Results

In this section we describe the results for the tasks of genre classification, automatic tagging and automatic playlist continuation.

[30] https://mtg.github.io/mtg-jamendo-dataset/
[31] https://mtg.github.io/mtg-jamendo-dataset/
[32] https://github.com/spotify/annoy

Table 8.1: GTZAN results

Model	Mean Accuracy ± STD
B-line$_{\text{G}}$	63.28 ± 1.19
B-line$_{\text{CF}}$	57.12 ± 1.82
B-line$_{\text{CF-G}}$	64.35 ± 1.10
Contr$_{\text{G}}$	**76.78** ± 1.22
Contr$_{\text{CF}}$	67.12 ± 0.94
Contr$_{\text{CF-G}}$	75.29 ± 1.32
COALA freesound (Favory et al., 2020a)	60.70
VGGish audioset (Gemmeke et al., 2017)	77.58
OpenL3 audioset (Cramer et al., 2019b)	74.65
Musicnn MTT (Pons & Serra, 2019)	71.37
Musicnn MSD (Pons & Serra, 2019)	77.24
VGG MTT (Choi et al., 2016)	72.75

Table 8.2: Automatic tagging results

	ROC AUC ± STD		
Model	Genre	Mood	Instrument
B-line$_{\text{G}}$	0.840 ± 0.004	0.722 ± 0.004	0.781 ± 0.005
B-line$_{\text{CF}}$	0.836 ± 0.002	0.722 ± 0.003	0.770 ± 0.008
B-line$_{\text{CF-G}}$	0.845 ± 0.004	0.727 ± 0.006	0.785 ± 0.004
Contr$_{\text{G}}$	**0.847** ± 0.004	0.732 ± 0.005	**0.797** ± 0.005
Contr$_{\text{CF}}$	0.845 ± 0.004	0.732 ± 0.004	0.793 ± 0.007
Contr$_{\text{CF-G}}$	0.843 ± 0.004	**0.733** ± 0.005	0.791 ± 0.006

8.4.1 Genre Classification

The results in Table 8.1 shows that the performance of the audio embedding when trained using the contrastive loss is always higher than using the models trained directly to predict the modality information (B-line). The best performance is obtained with Contr$_{\text{G}}$ with a similar result to when also considering CF information when training the embedding model (Contr$_{\text{CF-G}}$). We also see that the performances of the Contr$_{\text{G}}$ model are comparable with state-of-the-art pre-trained embeddings (VGGish audioset) (Gemmeke et al., 2017; Pons & Serra, 2019). This is particularly interesting since a large percentage of the Melon Playlist Dataset consists of korean music, which can be different from popular western music from the GTZAN collection.

Table 8.3: Playlist generation results

Model	NDCG@100	MAP@100
Random	0.0005	0.0001
B-line$_{\text{G}}$	0.0044	0.0007
B-line$_{\text{CF}}$	0.0035	0.0007
B-line$_{\text{CF-G}}$	0.0042	0.0008
Contr$_{\text{G}}$	0.0074	0.0016
Contr$_{\text{CF}}$	0.0076	0.0017
Contr$_{\text{CF-G}}$	**0.0085**	**0.0020**

8.4.2 Automatic Tagging

The results for the task of automatic tagging follow the same trend of the genre classification task. From the results in Table 8.2 we see that the methods based on contrastive learning outperform the baselines in almost all the cases. The best results for the instrument and genre tags is obtained with the Contr$_{\text{G}}$ model. For the mood tags the best performance is achieved with Contr$_{\text{CF-G}}$, which takes advantage of the information in the playlists and the genre annotations.

8.4.3 Automatic Playlist Continuation

The results for the task of automatic playlist continuation follow the same trend of the other tasks. The results in Table 8.3 show that the performance of the audio embedding when trained using the contrastive loss is always higher than using the models trained directly to predict the genres or the CF representation (B-line). In this case, the best performance is obtained with the Contr$_{\text{CF-G}}$ model, which combines genre and CF information.

8.5 Demo of Automatic Playlist Continuation

In this section we describe a web-based application built to demonstrate the outcome of our contrastive learning method used to train the audio encoder (Contr$_{\text{CF-G}}$).

The demo application is for playlist continuation. For each playlist in Melon Playlist Dataset, we selected 200 random tracks and connect these random tracks with the tracks of the playlist based on the cosine similarity between the embeddings obtained with the audio encoder (Contr$_{\text{CF-G}}$). The random tracks that have more connections with the tracks in the playlist are better candidates than the tracks without connections. Note that for this demos we

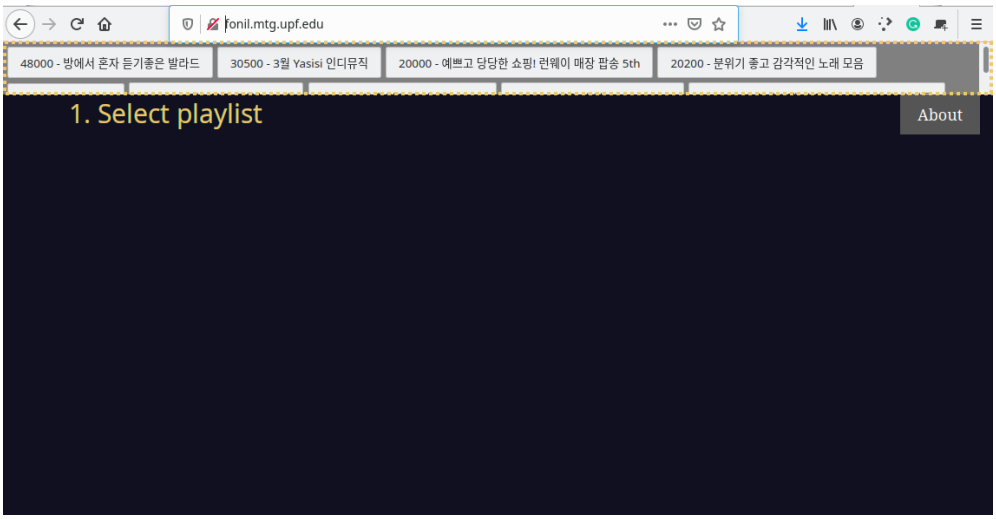

Figure 8.2: Screenshot of demo application (first step).

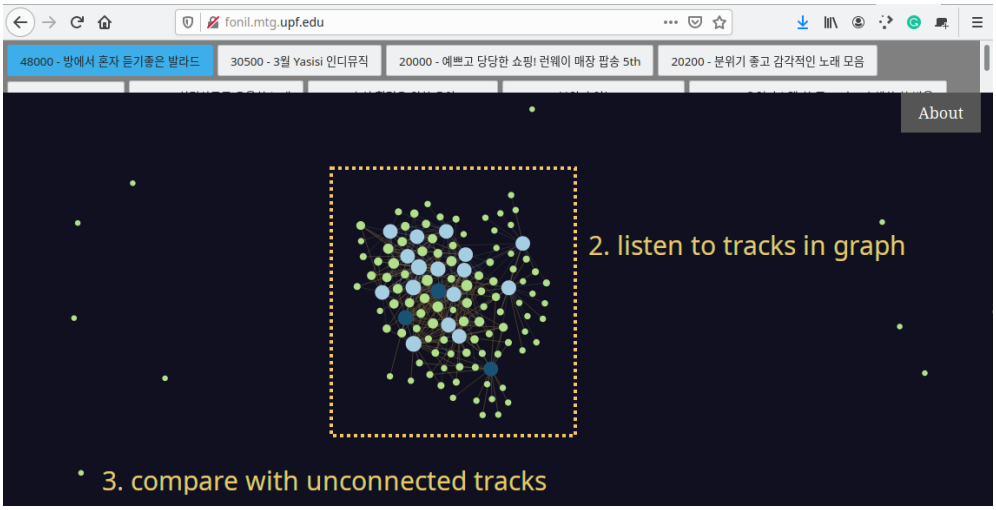

Figure 8.3: Screenshot of demo application (second step).

only use 10% of the tracks in the dataset, therefore, some playlists do not have tracks[33].

Figure 8.2 shows a screenshot of the initial page of the demo. Here the user has to select a playlist in the top menu to start, as highlighted with the yellow in the figure. After selecting a playlist, the tracks are shown as nodes in a graph (Figure 8.3).

While visualizing a playlist, the blue nodes are tracks in the original playlists

[33]The demo can be accessed online using this link: http://fonil.mtg.upf.edu/

Figure 8.4: Screenshot of demo application. Use scroll to zoom in/out and mouse over to listen.

while the green nodes are randomly selected tracks. Light blue indicates the track was used in the training of the model while dark blue indicates that the track was used in the validation set of the model. For each track in the playlist we connect with the 20 most similar tracks from the random group. The green nodes have different size depending on the number of connections they have. The color of the connections can be yellow or grey depending on the value of similarity between the tracks. Yellow connections indicate higher similarity and grey connections indicate lower similarity.

Figure 8.4 shows that once the graph is displayed, the user can zoom in and out using the mouse scroll. When placing the mouse over a node, the information of the track is displayed and the audio will be automatically played.

Using the demo, the user can evaluate the effectiveness of the model to capture similarity between the tracks only using the embeddings obtained with the audio encoder. For example, the effectiveness can be evaluated by comparing how disconnected tracks (Figure 8.5) are worse candidates than connected tracks. However, given that only 200 random tracks are selected for each playlist and from these the most similar 20 tracks are connected to each track of the playlist, it is possible that the connected tracks are not the optimal candidates to continue the playlist from the full catalog.

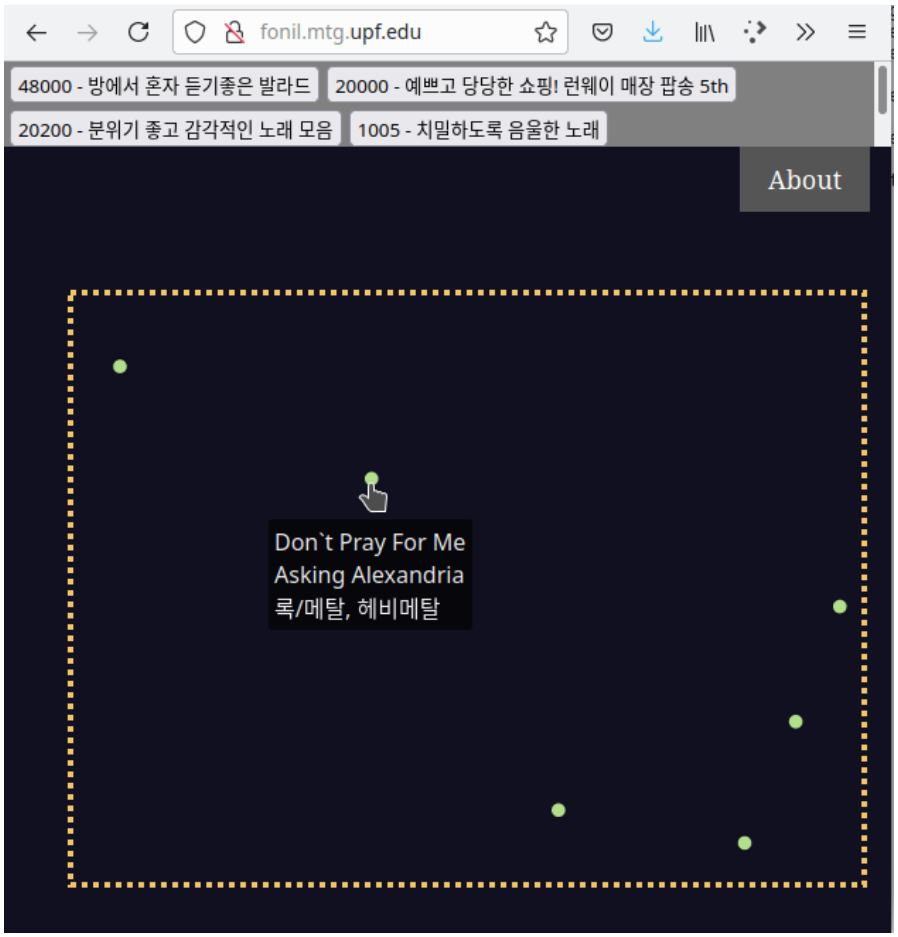

Figure 8.5: Screenshot of demo application. Compare connected nodes and disconnected nodes.

8.6 Conclusions

In this chapter, we propose a method for learning an audio representation, by combining multiple sources of information related to the music using contrastive learning. We evaluate the method by pre-training the model using information from the Melon Playlist Dataset and we compare the performance in three downstream tasks in the music domain (genre classification, automatic tagging, and automatic playlist continuation). We see that using contrastive learning allows us to reach higher performance than using the models trained directly to predict the genre or the collaborative filtering information. This indicates that contrastive learning is effective at learning simultaneously from heterogeneous information, enabling us to improve the overall performance across different tasks. Moreover, the contrastive learning method shows im-

proved performance when combining multiple sources of information compared to using only one source.

From the results in the task of automatic playlist continuation, we can also see that the method is more effective in recommending new tracks that were not in the training data. Therefore, the proposed method is aligned with the conclusions presented in Chapter 3 to generate better recommendations of new and less popular music. In addition, the proposed method allows us to consider and also to learn from information related to the context of the music. Incorporating information from the context of the music is also aligned with the artists' interest and can be further explored with the proposed model.

CHAPTER 9

Summary and Future Perspectives

9.1 Introduction

In this thesis we stress the importance that recommender systems have in music platforms, helping users to decide what to listen to, thus giving more or less exposure to different music artists. Researchers, developers and designers have a big responsibility to make sure the systems are fair, they are not biased, they do not discriminate and they give a fair chance to all the music artists on the platform. However, many groups of people are affected differently by these systems and sometimes it is not possible that the systems affect all the groups of people positively. In some cases, we need to find a trade-off between the interests of the people affected by the systems, deciding who gets more benefit.

Streaming platforms and their recommender systems are user-centered, previous works mostly focus on how these systems can give more satisfaction to the users since they need to retain users to maximize the income. There is no previous work focused on how the recommender system can give more value to the artists. Therefore, it is not clear how the recommender systems affect the artists and how they can be more beneficial for the artists. Understanding the ethical implications of a system is essential to find the optimal solution. However, there is no previous work that involved music artists to understand the implications that streaming platforms and recommender systems have in their work.

In this thesis, in order to understand how the music platforms–and their recommender systems– affect the artists and how they can be more beneficial in the future, we started by interviewing a diverse group of music artists. From these interviews, we identified some problems in how platforms present their music and also we identified some issues that come from the music industry that can be solved with the use of technology.

For example, we identified that artists consider that platforms and recommender systems do not give them the chance of reaching a larger audience. This suggests that there is still a strong popularity bias in the systems and recommender systems are not solving the cold start problem. Another example in which all interviewed artists agree is the gender imbalance in the platforms, an aspect in which they claimed platforms and recommender systems have the chance to improve.

Based on these findings we further investigated some recommender algorithms and identified situations that are aligned with the artists' opinions and show undesired biases. For example, we identified that a commonly-used collaborative filtering algorithm reproduces the bias in the data, recommending fewer female artists and also giving less exposure to female artists generating a feedback loop that is hard to break. We also see that the collaborative filtering algorithm has a lower performance for female artists compared to male artists. In another study, using simulations we see that session-based recommenders can reduce the number of elements that are recommended, focusing on more popular items. We proposed solutions to these problems that show improved performance in offline settings.

In this thesis, we also proposed a method for recommending music that at the same time is aligned with the interests of both artists and users. This method is based on collaborative filtering and leverages the user interactions to capture how engaged a user is with an artist. The goal of this method is to recommend artists that the user will be more engaged with. Therefore, the user would probably attend the artists' concerts, buy the artists' merchandising or listen to their new tracks. This method shows promising results and supports the idea that better systems are possible.

In addition, we created a dataset that can be used in the research community combining content-based methods with collaborative filtering, something that was limited before because of copyright issues. We first identified the optimal setting to compute mel-spectrograms from the audio, showing that they have a good trade-off in the task of automatic tagging using two state-of-the-art architectures. Based on this comparison we built the Melon Playlist Dataset which contains the mel-spectrograms for 650k tracks and 150k playlists with 30k different tags.

Finally, we proposed a method for improving music representation that can be used for recommending new music in a cold-start situation showing comparative performance with the state-of-the-art in tasks such as music genre classification, automatic playlist continuation and music automatic tagging. The method is based on contrastive learning and aligns multiple modalities of information.

In the following section, the contributions of the thesis are summarized. We

also mention the scientific publications that were made during the thesis, the code and datasets that were released with the goal of reproducibility. Finally, we mention some limitations and future work.

9.2 Summary of Contributions

Since this thesis focuses on some ethical aspects of music recommendations, one goal was to contribute to the society, by raising some questions that are important to discuss not only from a technical point of view but also from society in general.

This thesis is the first work focused on understanding the artists' perspective on music streaming platforms and the automatic recommendations in those platforms. Based on a qualitative study, some concrete aspects are identified which could make the streaming platforms more beneficial for the artists.

Based on some of the aspects identified from the qualitative study the following contributions are made in order to understand the effects of different algorithms and to propose concrete solutions to mitigate the negative effects:

- Studied the gender imbalance in a commonly used collaborative filtering method, analyzing the effect that this algorithm could have from the artists' perspective in the long term and proposing actions to reduce the imbalance in a progressive way.
- Proposed a way to combine multiple signals of user interaction in order to produce recommendations that maximize the engagement that users will have with artists.
- Studied the longitudinal effect of multiple algorithms for session-based recommendations, showing that these algorithms could have a negative effect in terms of coverage of the catalog, increasing the popularity of the items. A method to mitigate such negative effects is proposed.

It is well known that one of the main limitations for conducting research in the field of music recommendations is the limited datasets that combine audio information with user-created playlists due to commercial licensing of the content. One of the main contributions of this thesis to the research community is the creation of an open dataset of commercial music that contains commonly used mel-spectrogram representations of the audio and playlist information created by users from Melon, a popular streaming platform from Korea. In order to avoid licensing issues, it has been studied the trade-off between performance on the task of automatic tagging when reducing the information of the spectrograms used.

Finally, using the published dataset, a new method is proposed to take advantage of the alignment of multi-modal music data for exploiting the semantic metadata and collaborative filtering information. The proposed method shows improved performance in multiple tasks related to MIR (automatic playlist generation, genre classification and automatic-tagging from audio) compared to other common approaches that are considered state-of-the-art.

9.3 Publications, Open Research and Reproducibility

Open science has been promoted in the last years from private and public organizations to make available for everyone the advances of science. It brings multiple advantages to the research community, some of these advantages are: 1) facilitating collaborations between researchers by sharing resources 2) making results more rigorous and transparent 3) larger impact of the work in both academia and industry by making the outputs of the research publicly available and reproducible.

One additional reason to follow Open Science is to make public the work that is carried out in part with public resources and with the help of users' data.

The work described in this thesis was published in multiple conferences and journals. Chapter 3 was published in a conference paper (Ferraro et al., 2021d). The research presented in Chapter 4 was presented as a conference paper (Ferraro et al., 2021c). Chapter 5 also presents the work published in a conference (Ferraro et al., 2020e). The work described in Chapter 6 was presented in a conference (Ferraro et al., 2020c). Chapter 7 combines the work presented in two conferences (Ferraro et al., 2020b, 2021b). Finally, Chapter 8 describes the work presented in a journal paper (Ferraro et al., 2021a).

Following the principles of Open Science, as part of the dissemination strategy, we published under an open license all the preprint versions of the papers in open repositories like arxiv[34] or e-Repository[35].

9.3.1 Software

To make a reliable contribution to the field it is important to think about reproducibility. For this purpose, from the beginning of the thesis, we decided to make the code related to all the publications available online together with the datasets needed to reproduce the results.

The code to reproduce the experiments is available in the following repositories:

[34]https://arxiv.org/
[35]https://repositori.upf.edu/

- Chapter 4: https://github.com/andrebola/gender-recs
- Chapter 5: https://github.com/andrebola/artist-engagement
- Chapter 6: https://github.com/andrebola/session-rec-effect
- Chapter 7: https://github.com/andrebola/icassp2021 and https://github.com/andrebola/EUSIPCO2020
- Chapter 8: https://github.com/andrebola/contrastive-mir-learning

9.3.2 Datasets

Two new datasets were created for this thesis and are made public as a contribution for the research community:

- The data collected from MusicBrainz to carry out the experiments in Chapter 4 are available in Zenodo (Ferraro et al., 2020a). The dataset contains for *LFM-1b* information for 112k users and 465k tracks by 33k artists, and for *LFM-360k* contains information for 220k users and 12k artists.
- The Melon Playlist Dataset described in Chapter 7 contains 150k playlists and 650k songs, it also contains genre information for the songs and tag information for the playlists. The number of unique tags is 30k, the number of unique genres is 30 and the sub-genres are 219. For all the songs, it is provided the mel-spectrogram representation of a segment (20-50s) of the audio which enables the possibility of applying content-based approaches. The dataset can be accessed on the following page: https://mtg.github.io/melon-playlist-dataset/

9.3.3 Media Coverage

Some of the research carried out during this thesis was covered in the media which suggests that the topic is timely and relevant for society in general. We first published two articles for general audience (Bauer & Ferraro, 2021; Ferraro, 2021), which multiple journals and news portals used as a source to publish their own articles referring to our work.

9.4 Limitations and Future Work

In this section we describe the limitations faced in this thesis and we propose future work that can extend the work done.

9.4.1 Involving Artists for Building Fair Music Platforms

Future research should reach out to a wider set of artists and investigate the identified aspects more deeply. While we did look for some aspects of diversity in our sample of music artists, reaching out to a larger sample will allow for even more aspects of diversity (e.g., the scope of countries and continents, including artist of non-binary gender, considering solo artists, mixed-gender bands, mono-gender bands, various ethnic groups, artists dedicated to only a small set of niche music styles). Although we reached thematic saturation in our sample, reaching out to a wider set of diverse artists may reveal additional important topics or different viewpoints. For example, using surveys could allow confirming if the opinion of the interviewed artists is more or less shared by different groups of artists.

However, our work gives direction towards relevant topics for fairness on music platforms and their integrated music recommender systems. The findings indicate pathways towards fairer music platforms, whereas the concrete operationalization is subject to further research. In doing so, we can build on stronger foundations of prior research. For instance, from the interviews, we understand that artists see the need to promote new and less popular artists. While collaborative filtering is the vastly adopted approach in music recommender systems, it is an approach that is prone to popularity bias. As the music information retrieval research community had been fundamental in improving content-based recommendation approaches for the music domain (for an overview, see Murthy & Koolagudi (2018)), such approaches could be especially apt to promote new and less popular artists. Furthermore, different topics such as promoting local music and ensuring gender balance in recommendations exhibit similarities for their operationalization. First, meta-data about both, the artists' regional or cultural affiliation as well as gender information, are available for popular artists (e.g., using sources such as Wikipedia or MusicBrainz), but scarce for new and less popular artists. Thus, while existing meta-data may be used, other approaches have to be leveraged to gather missing data; This may be challenging for new and less popular artists in particular. Second, while multiple works investigated the diversity and coverage of computed recommendations (for an overview, see Kunaver & Požrl (2017)), little is known about how to ensure a ratio of attributes (such as region, culture and gender). In addition, the finding that artists perceive their profile presentation on the music platform as being fragmented and misplaced from its context, are an inspiration and rationale to leverage more approaches related to the field of music information retrieval in order to consolidate and structure information from dispersed sources, so that the music presentation can be enriched with this information and put into context.

For new and less popular artists, but also for new music by established artists,

it will be challenging to retrieve such information. Besides challenges with respect to the operationalization for information retrieval and consolidation, it is subject to future research to investigate how such contextualized information should be presented so that it (i) puts the music into context as it is meant by the respective artist and (ii) is understandable and appealing to the user.

9.4.2 Gender Imbalance in Music Recommendations

The work done related with gender imbalance in music recommendations opens multiple research possibilities. The analysis conducted included only 'solo artists' and binary genders, this is due to limitations regarding the data that was collected. Ideally, artists should indicate the gender that they feel more identified with and also bands should be considered in the analysis. In addition, further collaborations with researchers from the field of social science and ethics is needed to define how female artists can be better represented and promoted by music recommender systems. Improving female artists' exposure also brings the questions of how users perceive this and can be studied with mixed methods either in an independent environment or in collaboration with streaming platforms.

Future research should investigate how alternative algorithms behave. Also, a crucial future path of research will be to study how real users perceive the changes introduced by the re-ranking strategy in a real-world setting with an online study and how it impacts the users' listening behavior in the long term.

9.4.3 Maximizing Users' Engagement With Artists

As future work, following the multi-stakeholders literature, it would be important to get a trade-off between the interests of both users and artists. This can be achieved by optimizing at the same time for the users' and the artists' metrics. However, to properly assess that, a more comprehensive online evaluation with real users for a long period would be required to understand the impact of the different inputs, involving many users and for a very long period.

For the evaluation of the recommendations, there are strong limitations in the offline evaluation of recommender systems that we also face. One of the strongest limitations is that recommending something outside of the ground truth items does not mean that the user would not like it. In addition, offline evaluation usually has a popularity bias, favoring the algorithm that recommends more popular items (Bellogin et al., 2011; Steck, 2011; Park & Tuzhilin, 2008; Ferraro et al., 2019b), which can also vary depending on demographic aspects of the users (Ekstrand et al., 2018a), and can have a disparate bias for different user groups (Lin et al., 2019). Given the challenge of performing

this type of evaluation, simulation-based techniques have shown to be effective to study the impact that recommender systems can have on users' behaviour (Wall, 2016; Zhang et al., 2020b; Jannach et al., 2015b). In future work, these simulation techniques can be used to evaluate the impact that recommendations may have on both users and artists.

9.4.4 Algorithmic Influence in Session-Based Recommendation

This work so far is limited to a specific set of assumptions used in the simulation, e.g., that the observed interactions all come from the recommendations, which leads to an amplification of the observed effects. A user study could allow measuring the impact of these session-based algorithms considering how the behavior of the users changes when interacting with the recommendations. Our future works also include the investigation of other scenarios where long-term preference information about the users can be leveraged to diversify and de-bias the recommendations.

9.4.5 Melon Playlists Dataset

Our dataset's main limitation is that it provides mel-spectrograms instead of audio, making it impossible to apply methods based on other audio representations (e.g., raw waveforms). Nevertheless, the provided mel-spectrograms are suitable for the tasks of auto-tagging and automatic playlist continuation, which are the main focus of the proposed dataset. They offer a good trade-off considering the common limitations of re-using copyrighted commercial music in the field of MIR and audio signal processing. Besides, due to the large scale of the dataset, the reduced audio representations lower its distributable size offering advantages in transfer and storage.

Future work should be done to understand how the spectrograms provided in Melon Playlist Dataset can be adapted to be used in publicly available models that were trained with other datasets.

9.4.6 Enriched Music Representation Using Multi-Model Contrastive Learning

The results of the proposed method based on contrastive learning show that multiple modalities of information can be aligned. However, the dataset used for training our embedding model offers additional types of information that we did not use. They include title, playlist tags and authors, as well as other metadata of the tracks. We propose as future work to incorporate this playlist-

level information since it requires an additional level of abstraction to our architecture. Learning playlist embeddings have additional advantages since can be used for example to recommend playlists using a similarity measure.

In addition, the proposed method has the advantage that allows doing cross-modal retrieval, which is something that we did not explore in this thesis. With cross-modal retrieval we could, for example, obtain the tracks of a given combination of genres or given the collaborative filtering representation obtain the audio of a song.

Finally, understanding the implications of these content-based models can be explored more in deep. For example, measuring the robustness to biases in the datasets when using this content-based method for recommending new songs.

Appendix A

Glossary

A.1 Acronyms

ACM	Association of Computer Machinery
ALS	Alternating Least Squares
APC	Automatic Playlist Continuation
BPR	Bayesian Personalized Ranking
CB	Content-based
CF	Collaborative Filtering
MIR	Music Information Retrieval

Bibliography

Abdollahpouri, H., Adomavicius, G., Burke, R., Guy, I., Jannach, D., Kamishima, T., Krasnodebski, J., & Pizzato, L. (2020a). Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, *30*, 127–158. [Cited on pages 2, 23, 26, 62, and 71.]

Abdollahpouri, H., Burke, R., & Mobasher, B. (2017). Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the 11th ACM Conference on Recommender Systems (RecSys 2017)*, pp. 42–46. [Cited on pages 23 and 83.]

Abdollahpouri, H., Mansoury, M., Burke, R., & Mobasher, B. (2019). The impact of popularity bias on fairness and calibration in recommendation. *arXiv preprint arXiv:1910.05755*. [Cited on page 62.]

Abdollahpouri, H., Mansoury, M., Burke, R., & Mobasher, B. (2020b). The connection between popularity bias, calibration, and fairness in recommendation. In *Fourteenth ACM Conference on Recommender Systems*, RecSys '20, pp. 726–731. New York, NY, USA: ACM. [Cited on page 62.]

ACM (1972). ACM code of ethics. https://ethics.acm.org/code-of-ethics/previous-versions/1972-acm-code/. [Cited on page 2.]

Adamopoulos, P. & Tuzhilin, A. (2015). On unexpectedness in recommender systems: Or how to better expect the unexpected. *ACM Transactions on Intelligent Systems and Technology*, *5*(4), 54. [Cited on page 22.]

Adomavicius, G. & Kwon, Y. (2012). Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Trans. on Knowl. and Data Eng.*, *24*(5), 896–911. [Cited on pages 23, 90, and 92.]

Aguiar, L., Waldfogel, J., & Waldfogel, S. (2018). Playlisting favorites: Is Spotify gender-biased? JRC Technical Reports JRC113503, European Commission, Seville, Spain. JRC Digital Economy Working Paper 2018-07. [Cited on pages 2 and 30.]

Akimchuk, D., Clerico, T., & Turnbull, D. (2019). Evaluating recommender system algorithms for generating local music playlists. [Cited on page 26.]

Alonso-Jiménez, P., Bogdanov, D., Pons, J., & Serra, X. (2020). Tensorflow audio models in Essentia. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 266–270. [Cited on pages 118 and 124.]

Anderson, A., Maystre, L., Anderson, I., Mehrotra, R., & Lalmas, M. (2020). Algorithmic effects on the diversity of consumption on Spotify. In *Proc. of The Web Conference 2020*, WWW 2020, pp. 2155–2165. [Cited on page 25.]

Anderson, C. (2004). The long tail. https://www.wired.com/2004/10/tail/. [Cited on page 24.]

Anderson, C. (2006). *The long tail: Why the future of business is selling less of more*. New York, NY, USA: Hyperion. [Cited on page 24.]

Arcand, R. (2017). Is Lo-fi House the first genre of the algorithm age? https://www.vice.com/en/article/yp9e5j/lo-fi-house-youtube-related-video-algorithm-essay. [Cited on page 5.]

Bauer, C. (2019). Allowing for equal opportunities for artists in music recommendation: A position paper. In *Proc. of the 1st Workshop on Designing Human-Centric Music Information Research Systems*, wsHCMIR '19, pp. 16–18. [Cited on page 55.]

Bauer, C. & Ferraro, A. (2021). Music recommendation algorithms are unfair to female artists, but we can change that. https://theconversation.com/music-recommendation-algorithms-are-unfair-to-female-artists-but-we-can-change-that-158016. [Cited on page 137.]

Bauer, C., Kholodylo, M., & Strauss, C. (2017). Music recommender systems challenges and opportunities for non-superstar artists. In *Proc. of the 30th Bled eConference*, pp. 21–32. [Cited on page 24.]

Bauer, C. & Schedl, M. (2019). Global and country-specific mainstreaminess measures: Definitions, analysis, and usage for improving personalized music recommendation systems. *PLOS ONE*, *14*(6), 1–36. [Cited on page 48.]

Beckwith, L., Burnett, M., Grigoreanu, V., & Wiedenbeck, S. (2006). Gender HCI: What about the software? *Computer*, *39*(11), 97–101. [Cited on page 28.]

Beckwith, L., Burnett, M., Wiedenbeck, S., Cook, C., Sorte, S., & Hastings, M. (2005). Effectiveness of end-user debugging software features: Are there gender issues? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pp. 869–878. New York, NY, USA: ACM. [Cited on page 28.]

Bellogin, A., Castells, P., & Cantador, I. (2011). Precision-oriented evaluation of recommender systems: An algorithmic comparison. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pp. 333–336. New York, NY, USA: ACM. [Cited on pages 25 and 139.]

Benghozi, P.-J. & Benhamou, F. (2010). The long tail: Myth or reality? *International Journal of Arts Management*, *12*(3), 43–53. [Cited on page 24.]

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*. [Cited on pages 18 and 100.]

Biega, A. J., Gummadi, K. P., & Weikum, G. (2018). Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM Conference on Research & Development in Information Retrieval*, SIGIR '18, pp. 405–414. [Cited on page 27.]

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pp. 149–159. PMLR. [Cited on page 3.]

Binns, R. (2020). On the apparent conflict between individual and group fairness. In *Proc. of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 514–524. [Cited on page 27.]

Bogdanov, D., Haro, M., Fuhrmann, F., Xambó, A., Gómez, E., & Herrera, P. (2013a). Semantic audio content-based music recommendation and visualization based on user preference examples. *Information Processing & Management*, *49*(1), 13–33. [Cited on page 22.]

Bogdanov, D., Serrà, J., Wack, N., Herrera, P., & Serra, X. (2011). Unifying low-level and high-level music similarity measures. *IEEE Transactions on Multimedia*, *13*(4), 687–701. [Cited on page 15.]

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J. R., & Serra, X. (2013b). Essentia: an audio analysis library for music information retrieval. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 493–498. [Cited on page 103.]

Bogdanov, D., Won, M., Tovstogan, P., Porter, A., & Serra, X. (2019). The MTG-Jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*. Long Beach, CA, United States. [Cited on pages 18, 100, and 124.]

Bonnin, G. & Jannach, D. (2014). Automated generation of music playlists: Survey and experiments. *ACM Computing Surveys*, *47*(2), 1–35. [Cited on pages 14 and 87.]

Born, G., Morris, J., Diaz, F., & Anderson, A. (2021). Artificial intelligence, music recommendation, and the curation of culture. *Schwartz Reisman Institute White Paper*. [Cited on page 2.]

Bostrom, N. (2017). *Superintelligence*. Dunod. [Cited on pages 1 and 3.]

Brey, P. (2000). Method in computer ethics: Towards a multi-level interdisciplinary approach. *Ethics and information technology*, *2*(2), 125–129. [Cited on page 5.]

Burke, R. (2017). Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093*. [Cited on page 26.]

Burnett, M. M., Beckwith, L., Wiedenbeck, S., Fleming, S. D., Cao, J., Park, T. H., Grigoreanu, V., & Rector, K. (2011). Gender pluralism in problem-solving software. *Interacting with Computers*, *23*(5), 450–460. [Cited on page 28.]

Castells, P., Hurley, N. J., & Vargas, S. (2015). Novelty and diversity in recommender systems. In *Recommender systems handbook*, pp. 881–918. Springer. [Cited on page 23.]

Celma, Ò. (2009). *Music recommendation and discovery in the long tail*. Ph.D. thesis, Universitat Pompeu Fabra. [Cited on pages 2, 14, and 83.]

Celma, Ò. (2010). *Music recommendation and discovery: The long tail, long fail, and long play in the dgital music space*. Berlin, Heidelberg, Germany: Springer. [Cited on pages 24 and 61.]

Celma, Ò. & Herrera, P. (2008). A new approach to evaluating novel recommendations. In *Proceedings of the 2008 ACM conference on Recommender systems*, pp. 179–186. [Cited on pages 14, 21, and 117.]

Chaney, A. J. B., Stewart, B. M., & Engelhardt, B. E. (2018). How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, pp. 224–232. New York, NY, USA: ACM. [Cited on page 55.]

Chayka, K. (2019). Can monoculture survive the algorithm? https://www.vox.com/the-goods/2019/12/17/21024439/monoculture-algorithm-netflix-spotify. [Cited on page 4.]

Chen, C.-W., Lamere, P., Schedl, M., & Zamani, H. (2018). Recsys challenge 2018: Automatic music playlist continuation. In *Proc. of the 12th ACM Conference on Recommender Systems*, p. 527–528. [Cited on pages 16, 18, and 101.]

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*. [Cited on page 118.]

Cheng, C., Yang, H., Lyu, M. R., & King, I. (2013). Where you like to go next: Successive point-of-interest recommendation. In *Twenty-Third international joint conference on Artificial Intelligence*. [Cited on page 14.]

Choi, J., Lee, J., Park, J., & Nam, J. (2019). Zero-shot learning for audio-based music classification and tagging. *arXiv preprint arXiv:1907.02670*. [Cited on page 118.]

Choi, K., Fazekas, G., Cho, K., & Sandler, M. (2017a). A tutorial on deep learning for music information retrieval. *arXiv preprint arXiv:1709.04396*. [Cited on pages 16, 17, and 99.]

Choi, K., Fazekas, G., Cho, K., & Sandler, M. (2018a). The effects of noisy labels on deep convolutional neural networks for music tagging. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *2*(2), 139–149. [Cited on pages 18 and 100.]

Choi, K., Fazekas, G., & Sandler, M. (2016). Automatic tagging using deep convolutional neural networks. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 805–811. [Cited on pages 16, 17, 97, 100, 101, 102, 104, 105, and 126.]

Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017b). Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2392–2396. IEEE. [Cited on pages 16, 17, and 113.]

Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2018b). A comparison of audio signal preprocessing methods for deep neural networks on music tagging. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 1870–1874. IEEE. [Cited on pages 17, 99, and 103.]

Coelho, M. P. & Mendes, J. Z. (2019). Digital music and the "death of the long tail". *Journal of Business Research*, *101*, 454–460. [Cited on page 24.]

Collins, A., Tkaczyk, D., Aizawa, A., & Beel, J. (2018). Position bias in recommender systems for digital libraries. In G. Chowdhury, J. McLeod,

V. Gillet, & P. Willett (Eds.) *Transforming Digital Worlds. iConference 2018*, iConference 2018, pp. 335–344. Cham, Germany: Springer. [Cited on page 62.]

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 797–806. [Cited on page 3.]

Corrêa, D. C. & Rodrigues, F. A. (2016). A survey on symbolic data-based music genre classification. *Expert Systems with Applications*, *60*. [Cited on page 17.]

Cramer, H., Garcia-Gathright, J., Reddy, S., Springer, A., & Takeo Bouyer, R. (2019a). Translation, tracks & data: An algorithmic bias effort in practice. In *Extended Abstracts of the 2019 Conference on Human Factors in Computing Systems*, CHI EA '19, pp. 1–8. [Cited on pages 4 and 33.]

Cramer, H., Garcia-Gathright, J., Springer, A., & Reddy, S. (2018). Assessing and addressing algorithmic bias in practice. *Interactions*, *25*(6), 58–63. [Cited on pages 4 and 33.]

Cramer, J., Wu, H.-H., Salamon, J., & Bello, J. P. (2019b). Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3852–3856. IEEE. [Cited on pages 118, 124, and 126.]

Crawford, K. (2016). Artificial intelligence's white guy problem. https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html. [Cited on page 1.]

Crawford, K. & Calo, R. (2016). There is a blind spot in ai research. *Nature News*, *538*(7625), 311. [Cited on page 1.]

Creswell, J. W. & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA, USA: Sage Publications. [Cited on pages 35 and 56.]

Dacrema, M. F., Boglio, S., Cremonesi, P., & Jannach, D. (2021). A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems (TOIS)*, *39*(2), 1–49. [Cited on page 13.]

D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., & Halpern, Y. (2020). Fairness is not static: Deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness,*

Accountability, and Transparency, FAT* '20, pp. 525–534. New York, NY, USA: ACM. [Cited on page 55.]

Dasgupta, S. & Freund, Y. (2008). Random projection trees and low dimensional manifolds. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pp. 537–546. [Cited on page 125.]

Davis, J. & Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240. ACM. [Cited on page 105.]

de Revere, P. (2015). A bechdel test for music. *Pitchfork*. [Cited on page 56.]

Defferrard, M., Benzi, K., Vandergheynst, P., & Bresson, X. (2017). Fma: A dataset for music analysis. In *Proc. 18th International Society for Music Information Retrieval Conference (ISMIR)*, CONF. [Cited on pages 18 and 100.]

Dervakos, E., Kotsani, N., & Stamou, G. (2021). Genre recognition from symbolic music with cnns. In *Artificial Intelligence in Music, Sound, Art and Design: 10th International Conference, EvoMUSART 2021, Held as Part of EvoStar 2021, Virtual Event, April 7–9, 2021, Proceedings 10*, pp. 98–114. Springer International Publishing. [Cited on page 18.]

Dieleman, S. & Schrauwen, B. (2014). End-to-end learning for music audio. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6964–6968. IEEE. [Cited on pages 16, 100, and 103.]

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proc. of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pp. 214–226. [Cited on page 27.]

Ebesu, T., Shen, B., & Fang, Y. (2018). Collaborative memory network for recommendation systems. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 515–524. [Cited on page 13.]

Ekstrand, M. D., Burke, R., & Diaz, F. (2019). Fairness and discrimination in recommendation and retrieval. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, pp. 576–577. New York, NY, USA: ACM. [Cited on page 29.]

Ekstrand, M. D. & Mahant, V. (2017). Sturgeon and the cool kids: Problems with random decoys for top-n recommender evaluation. In *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference*, FLAIRS '17, pp. 639–644. Palo Alto, CA, USA: AAAI. [Cited on page 29.]

Ekstrand, M. D., Tian, M., Azpiazu, I. M., Ekstrand, J. D., Anuyah, O., McNeill, D., & Pera, M. S. (2018a). All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Proc. of the 1st Conference on Fairness, Accountability and Transparency*, FAT* '18, pp. 172–186. [Cited on pages 25, 26, 29, and 139.]

Ekstrand, M. D., Tian, M., Kazi, M. R. I., Mehrpouyan, H., & Kluver, D. (2018b). Exploring author gender in book rating and recommendation. In *Proc. of the 12th ACM Conference on Recommender Systems*, RecSys '18, pp. 242–250. [Cited on pages 26 and 29.]

Epps-Darling, A., Bouyer, R. T., & Cramer, H. (2020). Artist gender representation in music streaming. In *Proceedings of the 21st International Society for Music Information Retrieval Conference*, ISMIR 2020, pp. 248–254. ISMIR. [Cited on pages 30 and 61.]

European Commission (2021). Europe fit for the digital age: Commission proposes new rules and actions for excellence and trust in artificial intelligence. https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1682. [Cited on page 5.]

Falcao, F. & Mélo, D. (2017). The million playlists songs dataset: a descriptive study over multiple sources of user-curated playlists. In *16th Brazilian Symposium on Computer Music*. [Cited on pages 19 and 101.]

Farnadi, G., Kouki, P., Thompson, S. K., Srinivasan, S., & Getoor, L. (2018). A fairness-aware hybrid recommender system. *arXiv preprint arXiv:1809.09030*. [Cited on page 26.]

Farrahi, K., Schedl, M., Vall, A., Hauger, D., & Tkalcic, M. (2014). Impact of listening behavior on music recommendation. In *Proceedings of the 15th International Society for Music Information Retrieval Conference*, pp. 483–488. [Cited on page 14.]

Favory, X., Drossos, K., Virtanen, T., & Serra, X. (2020a). Coala: Co-aligned autoencoders for learning semantically enriched audio representations. In *International Conference on Machine Learning (ICML), Workshop on Self-supervised learning in Audio and Speech*. [Cited on pages 118, 119, 121, and 126.]

Favory, X., Drossos, K., Virtanen, T., & Serra, X. (2020b). Learning contextual tag embeddings for cross-modal alignment of audio and tags. *arXiv preprint arXiv:2010.14171*. [Cited on pages 119 and 121.]

Ferraro, A. (2019). Music cold-start and long-tail recommendation: Bias in deep representations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, pp. 586–590. New York, NY, USA: ACM. [Cited on page 71.]

Ferraro, A. (2021). A study finds gender bias in music recommendation algorithms. https://www.upf.edu/web/focus/noticies/-/asset_publisher/qOocsyZZDGHL/content/id/244709236/maximized. [Cited on page 137.]

Ferraro, A., Bauer, C., & Serra, X. (2020a). Last.fm artists gender information. https://doi.org/10.5281/zenodo.3748787. [Cited on pages 61 and 137.]

Ferraro, A., Bogdanov, D., Jay, X. S., Jeon, H., & Yoon, J. (2020b). How low can you go? Reducing frequency and time resolution in current CNN architectures for music auto-tagging. In *28th European Signal Processing Conference (EUSIPCO)*, pp. 131–135. IEEE. [Cited on pages 97 and 136.]

Ferraro, A., Bogdanov, D., & Serra, X. (2019a). Skip prediction using boosting trees based on acoustic features of tracks in sessions. *arXiv preprint arXiv:1903.11833*. [Cited on page 15.]

Ferraro, A., Bogdanov, D., Serra, X., & Yoon, J. (2019b). Artist and style exposure bias in collaborative filtering based music recommendations. In *Proc. of the 1st Workshop on Designing Human-Centric Music Information Research Systems*, wsHCMIR '19, pp. 8–10. [Cited on pages 25, 60, 71, and 139.]

Ferraro, A., Bogdanov, D., Yoon, J., Kim, K., & Serra, X. (2018). Automatic playlist continuation using a hybrid recommender system combining features from text and audio. In *Proceedings of the ACM Recommender Systems Challenge 2018*, p. 2. [Cited on page 16.]

Ferraro, A., Favory, X., Drossos, K., Kim, Y., & Bogdanov, D. (2021a). Enriched music representations with multiple cross-modal contrastive learning. *IEEE Signal Processing Letters*. [Cited on pages 119 and 136.]

Ferraro, A., Jannach, D., & Serra, X. (2020c). Exploring longitudinal effects of session-based recommendations. In *14th ACM Conference on Recommender Systems*, RecSys '20, pp. 474–479. [Cited on pages 25, 55, 84, and 136.]

Ferraro, A., Jeon, J. H., Kim, B., Serra, X., & Bogdanov, D. (2020d). Artist biases in collaborative filtering for music recommendation. In *Machine Learning for Media Discovery Workshop at International Conference on Machine Learning (ICML)*. [Cited on page 30.]

Ferraro, A., Kim, Y., Lee, S., Kim, B., Jo, N., Lim, S., Lim, S., Jang, J., Kim, S., Serra, X. et al. (2021b). Melon playlist dataset: a public dataset for audio-based playlist generation and music tagging. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 536–540. IEEE. [Cited on pages 98 and 136.]

Ferraro, A. & Lemström, K. (2018). On large-scale genre classification in symbolically encoded music by automatic identification of repeating patterns. In *Proceedings of the 5th International Conference on Digital Libraries for Musicology*, pp. 34–37. [Cited on page 17.]

Ferraro, A., Oramas, S., Quadrana, M., & Serra, X. (2020e). Maximizing the engagement: Exploring new signals of implicit feedback in music recommendations. In *Bogers T, Koolen M, Petersen C, Mobasher, Tuzhilin A, Sar Shalom O, Jannach D, Konstan JA, editors. Proceedings of the Workshops on Recommendation in Complex Scenarios and the Impact of Recommender Systems co-located with 14th ACM Conference on Recommender Systems (RecSys 2020); 25 Sep 2020; Brazil. Aachen: CEUR Workshop Proceedings; 2020.* CEUR Workshop Proceedings. [Cited on pages 72 and 136.]

Ferraro, A., Serra, X., & Bauer, C. (2021c). Break the loop: Gender imbalance in music recommenders. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pp. 249–254. [Cited on pages 56 and 136.]

Ferraro, A., Serra, X., & Bauer, C. (2021d). What is fair? exploring the artists' perspective on the fairness of music streaming platforms. In *IFIP TC13 International Conference on Human-Computer Interaction (INTERACT 2021).* [Cited on pages 34 and 136.]

Fleder, D. & Hosanagar, K. (2009). Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management science*, *55*(5), 697–712. [Cited on pages 24 and 83.]

Fonseca, E., Ortego, D., McGuinness, K., O'Connor, N. E., & Serra, X. (2020). Unsupervised contrastive learning of sound event representations. *arXiv preprint arXiv:2011.07616.* [Cited on page 118.]

Forsblom, A., Nurmi, P., Åman, P., & Liikkanen, L. (2012). Out of the bubble: Serendipitous even recommendations at an urban music festival. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pp. 253–256. [Cited on page 16.]

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780. IEEE. [Cited on pages 124 and 126.]

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, *35*(12), 61–71. [Cited on page 12.]

Graves, W. (2021). Spotify secures horrifying patent to monitor users' speech. https://consequence.net/2021/01/spotify-patent-monitor-users-speech/. [Cited on page 4.]

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough?: An experiment with data saturation and variability. *Field Methods*, *18*(1), 59–82. [Cited on pages 35 and 56.]

Harrison, G., Hanson, J., Jacinto, C., Ramirez, J., & Ur, B. (2020). An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, p. 392–402. New York, NY, USA: Association for Computing Machinery. [Cited on page 27.]

He, X., Du, X., Wang, X., Tian, F., Tang, J., & Chua, T.-S. (2018). Outer product-based neural collaborative filtering. *arXiv preprint arXiv:1808.03912*. [Cited on page 13.]

Helberger, N., Araujo, T., & de Vreese, C. H. (2020). Who is the fairest of them all? public attitudes and expectations regarding automated decision-making. *Computer Law & Security Review*, *39*, 105456. [Cited on page 27.]

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, *22*(1), 5–53. [Cited on pages 22 and 84.]

Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2016). Session-based recommendations with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016*. [Cited on pages 13, 83, and 87.]

Hogan, M. (2020). This is how much more money artists earn from bandcamp compared to streaming services. https://pitchfork.com/thepitch/how-much-more-money-artists-earn-from-bandcamp-compared-to-spotify-apple-music-youtube/. [Cited on page 4.]

Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In *Proc. of the Conference on Human Factors in Computing Systems*, CHI '19. New York, NY, USA: ACM. [Cited on page 34.]

Holzapfel, A., Sturm, B., & Coeckelbergh, M. (2018). Ethical dimensions of music information retrieval technology. *Transactions of the International Society for Music Information Retrieval*, *1*(1), 44–55. [Cited on page 2.]

Hsuan-Tien Lin, Maria Florina Balcan, R. H. & Ranzato, M. (2020). What we learned from neurips 2020 reviewing process.

https://neuripsconf.medium.com/what-we-learned-from-neurips-2020-reviewing-process-e24549eea38f. [Cited on page 4.]

Hu, R. & Pu, P. (2011). Helping users perceive recommendation diversity. In *Proceedings of the Workshop on Novelty and Diversity in Recommender Systems (DiveRS), at the 5th ACM International Conference on Recommender Systems*, pp. 43–50. [Cited on page 22.]

Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, ICDM 2008, pp. 263–272. New York, NY, USA: IEEE. [Cited on pages 12, 60, 72, and 78.]

Hutchinson, B. & Mitchell, M. (2019). 50 years of test (un)fairness: Lessons for machine learning. In *Proc. of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 49–58. [Cited on pages 27 and 33.]

Hutson, M. (2021). Who should stop unethical a.i.? https://www.newyorker.com/tech/annals-of-technology/who-should-stop-unethical-ai. [Cited on page 3.]

IFPI (2019). IFPI music listening 2019. https://www.ifpi.org/wp-content/uploads/2020/07/Music-Listening-2019-1.pdf. [Cited on page 1.]

Jannach, D. & Jugovac, M. (2019). Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems*, *10*(4). [Cited on page 69.]

Jannach, D., Lerche, L., & Kamehkhosh, I. (2015a). Beyond "hitting the hits" - generating coherent music playlist continuations with the right tracks. In *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys 2015)*, pp. 187–194. [Cited on page 92.]

Jannach, D., Lerche, L., Kamehkhosh, I., & Jugovac, M. (2015b). What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction*, *25*(5), 427–491. [Cited on pages 23, 25, 26, 60, 83, 85, 86, 95, and 140.]

Jannach, D., Lerche, L., & Zanker, M. (2018). Recommending based on implicit feedback. In *Social Information Access*, pp. 510–569. Springer. [Cited on page 71.]

Johnson, D. G. (1985). Computer ethics. *Englewood Cliffs (NJ)*. [Cited on page 2.]

Johnson, D. G. & Verdicchio, M. (2017). Reframing ai discourse. *Minds and Machines*, *27*(4), 575–590. [Cited on page 4.]

Kamehkhosh, I., Bonnin, G., & Jannach, D. (2020). Effects of recommendations on the playlist creation behavior of users. *User Modeling and User-Adapted Interaction*, *30*, 285–322. [Cited on page 86.]

Kereliuk, C., Sturm, B. L., & Larsen, J. (2015). Deep learning and music adversaries. *IEEE Transactions on Multimedia*, *17*(11), 2059–2071. [Cited on page 124.]

Keyes, O. (2018). The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–22. [Cited on page 29.]

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2020). Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*. [Cited on page 118.]

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. [Cited on page 102.]

Kirn, P. (2019). No, sharing your Spotify year-end artist stats is not a good idea – here's why not. https://cdm.link/2019/12/no-sharing-your-spotify-for-artists-wrapped/. [Cited on page 5.]

Knees, P. & Schedl, M. (2013). A survey of music similarity and recommendation from music context data. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *10*(1), 2. [Cited on page 15.]

Knees, P. & Schedl, M. (2016). Collaborative music similarity and recommendation. In *Music Similarity and Retrieval*, pp. 179–211. Springer. [Cited on page 15.]

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, *22*(4-5), 441–504. [Cited on page 21.]

Konstan, J. A. & Riedl, J. (2012). Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, *22*(1), 101–123. [Cited on page 25.]

Korzeniowski, F., Nieto, O., McCallum, M., Won, M., Oramas, S., & Schmidt, E. (2020). Mood classification using listening data. *arXiv preprint arXiv:2010.11512*. [Cited on page 117.]

Kunaver, M. & Požrl, T. (2017). Diversity in recommender systems – a survey. *Knowledge-Based Systems*, *123*, 154–162. [Cited on page 138.]

Law, E., West, K., Mandel, M. I., Bay, M., & Downie, J. S. (2009). Evaluation of algorithms using games: The case of music tagging. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 387–392. [Cited on pages 18 and 100.]

Le-Khac, P. H., Healy, G., & Smeaton, A. F. (2020). Contrastive representation learning: A framework and review. *IEEE Access*. [Cited on page 118.]

Lee, D. & Hosanagar, K. (2019). How do recommender systems affect sales diversity? a cross-category investigation via randomized field experiment. *Information Systems Research*, *30*(1), 239–259. [Cited on pages 23, 83, and 85.]

Lee, J., Bryan, N. J., Salamon, J., Jin, Z., & Nam, J. (2020a). Metric learning vs classification for disentangled music representation learning. In *Proc. of the 21st International Society for Music Information Retrieval Conference (ISMIR)*. [Cited on page 118.]

Lee, J. & Nam, J. (2017). Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging. *IEEE Signal Processing Letters*, *24*(8), 1208–1212. [Cited on page 16.]

Lee, J., Park, J., Kim, K. L., & Nam, J. (2017). Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms. In *Proceedings of the 14th Sound and Music Computing Conference (SMC)*. [Cited on pages 16 and 17.]

Lee, M. K., Grgić-Hlača, N., Tschantz, M. C., Binns, R., Weller, A., Carney, M., & Inkpen, K. (2020b). Human-centered approaches to fair and responsible ai. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pp. 1–8. New York, NY, USA: ACM. [Cited on page 27.]

Li, J., Ren, P., Chen, Z., Ren, Z., Lian, T., & Ma, J. (2017). Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pp. 1419–1428. [Cited on pages 13 and 87.]

Liang, D., Zhan, M., & Ellis, D. P. (2015). Content-aware collaborative music recommendation using pre-trained neural networks. In *Proc. of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 295–301. [Cited on page 112.]

Lin, K., Sonboli, N., Mobasher, B., & Burke, R. (2019). Crank up the volume: Preference bias amplification in collaborative recommendation. In *RMSE workshop held in conjunction with the 13th ACM Conference on Recommender Systems*. [Cited on pages 25, 29, and 139.]

Ludewig, M. & Jannach, D. (2018). Evaluation of session-based recommendation algorithms. *User-Modeling and User-Adapted Interaction*, *28*(4–5), 331–390. [Cited on pages 13 and 87.]

Ludewig, M., Mauro, N., Latifi, S., & Jannach, D. (2019). Performance comparison of neural and non-neural approaches to session-based recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, pp. 462–466. [Cited on pages 13, 14, 75, and 87.]

Maicki, S. (2020). Spotify to offer artists and labels the option to promote their music in your recommendations. https://www.thefader.com/2020/11/02/spotify-promotional-tool-discovery-weekly-royalty-rate-union-letter. [Cited on page 5.]

Maner, W. (1980). Starter kit in computer ethics. *Hyde Park, NY: Helvetia Press and the National Information and Resource Center for Teaching Philosophy*, *3*. [Cited on page 2.]

Mayring, P. (2004). Qualitative content analysis. In U. Flick, E. von Kardoff, & I. Steinke (Eds.) *A companion to qualitative research*, chap. 5.12, pp. 159–176. London, United Kingdom: Sage Publications. [Cited on pages 6, 34, 38, and 57.]

McDermott, M. (2020). Anti-algorithmic music: How bandcamp is helping artists beat the odds. https://ra.co/features/3703. [Cited on page 4.]

McFee, B. & Lanckriet, G. R. (2012). Hypergraph models of playlist dialects. In *Proc. 13th International Society for Music Information Retrieval Conference (ISMIR)*, vol. 12, pp. 343–348. Citeseer. [Cited on pages 19 and 101.]

McNee, S. M., Riedl, J., & Konstan, J. A. (2006). Making recommendations better: An analytic model for human-recommender interaction. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, p. 1103–1108. New York, NY, USA: ACM. [Cited on page 25.]

Mehrotra, R., Lalmas, M., Kenney, D., Lim-Meng, T., & Hashemian, G. (2019). Jointly leveraging intent and interaction signals to predict user satisfaction with slate recommendations. In *The World Wide Web Conference*, WWW '19, p. 1256–1267. New York, NY, USA: ACM. [Cited on page 71.]

Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., & Diaz, F. (2018). Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proc. of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pp. 2243–2251. [Cited on page 26.]

Meredith, D., Lemström, K., & Wiggins, G. A. (2002). Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, *31*(4), 321–345. [Cited on page 17.]

Metaxa-Kakavouli, D., Wang, K., Landay, J. A., & Hancock, J. (2018). Gender-inclusive design: Sense of belonging and bias in web interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 1–6. New York, NY, USA: ACM. [Cited on page 28.]

Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. *AI & Society*. [Cited on pages 2, 33, and 69.]

Minsker, E. (2021). Over 180 musicians protest Spotify's speech monitoring patent in open letter. https://pitchfork.com/news/over-180-musicians-protest-spotify-speech-monitoring-patent-in-open-letter/. [Cited on page 4.]

Mitchum, R. & Garcia-Olano, D. (2018). Tracking the gender balance of this year's music festival lineups. *Pitchfork*. [Cited on page 56.]

Moor, J. H. (1985). What is computer ethics? *Metaphilosophy*, *16*(4), 266–275. [Cited on page 2.]

Morse, J. M. (1994). Designing funded qualitative research. In *Handbook of qualitative research*, pp. 220–235. Thousand Oaks, CA, USA: Sage Publications. [Cited on pages 35 and 56.]

Mulligan, M. (2021). Equitable remuneration, artist income and unintended consequences. https://www.midiaresearch.com/blog/equitable-remuneration-artist-income-and-unintended-consequences. [Cited on page 4.]

Murthy, Y. V. S. & Koolagudi, S. G. (2018). Content-based music information retrieval (cb-mir) and its applications toward the music industry: A review. *ACM Computing Survey*, *51*(3). [Cited on page 138.]

Narayanan, A. (2018). Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, vol. 1170. [Cited on page 3.]

North, A. C. & Hargreaves, D. J. (2007a). Lifestyle correlates of musical preference: 1. relationships, living arrangements, beliefs, and crime. *Psychology of music*, *35*(1), 58–87. [Cited on page 4.]

North, A. C. & Hargreaves, D. J. (2007b). Lifestyle correlates of musical preference: 2. media, leisure time and music. *Psychology of music*, *35*(2), 179–200. [Cited on page 4.]

North, A. C. & Hargreaves, D. J. (2007c). Lifestyle correlates of musical preference: 3. travel, money, education, employment and health. *Psychology of music*, *35*(3), 473–497. [Cited on page 4.]

Oliveira, R. S., Nóbrega, C., Marinho, L. B., & Andrade, N. (2017). A multiobjective music recommendation approach for aspect-based diversification. In *Proceedings of the 18th International Society for Music Information Retrieval Conference*, ISMIR 2017, pp. 414–420. ISMIR. [Cited on pages 14 and 29.]

O'neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown. [Cited on pages 1 and 3.]

Oramas, S., Barbieri, F., Nieto, O., & Serra, X. (2018a). Multimodal deep learning for music genre classification. *Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21.* [Cited on pages 17 and 118.]

Oramas, S., Bogdanov, D., & Porter, A. (2018b). Mediaeval 2018 acousticbrainz genre task: A baseline combining deep feature embeddings across datasets. In *MediaEval 2018 Workshop*. Sophia Antipolis, France. [Cited on page 16.]

Oramas, S., Nieto, O., Barbieri, F., & Serra, X. (2017a). Multi-label music genre classification from audio, text and images using deep features. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 23–30. [Cited on pages 16 and 105.]

Oramas, S., Nieto, O., Sordo, M., & Serra, X. (2017b). A deep multimodal approach for cold-start music recommendation. In *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems*, pp. 32–37. [Cited on pages xix, 7, 15, and 16.]

Oramas, S., Ostuni, V. C., Noia, T. D., Serra, X., & Sciascio, E. D. (2016). Sound and music recommendation with knowledge graphs. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *8*(2), 1–21. [Cited on page 74.]

Pariser, E. (2011). *The filter bubble: How the new personalized web is changing what we read and how we think*. East Rutherford, NJ, USA: Penguin. [Cited on page 55.]

Pariser, E. (2012). *The filter bubble: What the internet is hiding from you*. London: Penguin Books. [Cited on page 83.]

Park, Y.-J. & Tuzhilin, A. (2008). The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM Conference on*

Recommender Systems, RecSys '08, pp. 11–18. New York, NY, USA: ACM. [Cited on pages 25 and 139.]

Pelly, L. (2018). Streambait pop. https://thebaffler.com/downstream/streambait-pop-pelly. [Cited on page 5.]

Pichl, M., Zangerle, E., & Specht, G. (2015). Towards a context-aware music recommendation approach: What is hidden in the playlist name? In *IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 1360–1365. IEEE. [Cited on pages 19 and 101.]

Pierson, E. (2017). Demographics and discussion influence views on algorithmic fairness. [Cited on page 28.]

Pons, J., Nieto, O., Prockup, M., Schmidt, E. M., Ehmann, A. F., & Serra, X. (2018). End-to-end learning for music audio tagging at scale. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 637–644. [Cited on pages 16, 17, 97, 100, 102, and 105.]

Pons, J. & Serra, X. (2017). Designing efficient architectures for modeling temporal features with convolutional neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2472–2476. IEEE. [Cited on page 17.]

Pons, J. & Serra, X. (2019). Musicnn: pre-trained convolutional neural networks for music audio tagging. In *Late-breaking/demo session in 20th International Society for Music Information Retrieval Conference (LBD-ISMIR2019)*. [Cited on pages 124 and 126.]

Pons, J., Slizovskaia, O., Gong, R., Gómez, E., & Serra, X. (2017). Timbre analysis of music audio signals with convolutional neural networks. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 2744–2748. IEEE. [Cited on page 17.]

Qiu, L., Li, S., & Sung, Y. (2021). Dbtmpe: Deep bidirectional transformers-based masked predictive encoder approach for music genre classification. *Mathematics*, *9*(5), 530. [Cited on page 18.]

Quadrana, M., Cremonesi, P., & Jannach, D. (2018). Sequence-aware recommender systems. *ACM Computing Surveys*, *54*, 1–36. [Cited on pages 16 and 83.]

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*. [Cited on page 118.]

Raffel, C. (2016). *Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching*. Ph.D. thesis, Columbia University. [Cited on page 17.]

Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2009). Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 452–461. [Cited on page 12.]

Rendle, S., Freudenthaler, C., & Schmidt-Thieme, L. (2010). Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 811–820. [Cited on page 14.]

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (2010). *Recommender systems handbook*. Berlin, Heidelberg: Springer-Verlag, 1st edn. [Cited on pages 11, 12, 19, 62, 73, 74, 113, and 125.]

Rosen, S. (1981). The economics of superstars. *The American Economic Review*, *71*(5), 845–858. [Cited on page 24.]

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin. [Cited on pages 1 and 3.]

Saeed, A., Grangier, D., & Zeghidour, N. (2020). Contrastive learning of general-purpose audio representations. *arXiv preprint arXiv:2010.10915*. [Cited on page 118.]

Sapiezynski, P., Zeng, W., E Robertson, R., Mislove, A., & Wilson, C. (2019). Quantifying the impact of user attention on fair group representation in ranked lists. In *Proc. of The 2019 World Wide Web Conference*, WWW '19, pp. 553–562. [Cited on page 27.]

Schedl, M. (2016). The LFM-1b dataset for music retrieval and recommendation. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, pp. 103–110. New York, NY, USA: ACM. [Cited on pages 61 and 75.]

Schedl, M., Gómez, E., Urbano, J. et al. (2014). Music information retrieval: Recent developments and applications. *Foundations and Trends® in Information Retrieval*, *8*(2-3), 127–261. [Cited on page 16.]

Schedl, M. & Hauger, D. (2015). 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*, pp. 947–950. [Cited on page 22.]

Schedl, M., Hauger, D., & Schnitzer, D. (2012). A model for serendipitous music retrieval. In *Proceedings of the 2nd Workshop on Context-awareness in Retrieval and Recommendation*, pp. 10–13. [Cited on page 16.]

Schedl, M. & Schnitzer, D. (2013). Hybrid retrieval approaches to geospatial music recommendation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 793–796. [Cited on page 16.]

Schedl, M., Zamani, H., Chen, C.-W., Deldjoo, Y., & Elahi, M. (2018). Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, *7*(2), 95–116. [Cited on pages 19, 20, 22, and 112.]

Schmutz, V. & Faupel, A. (2010). Gender and cultural consecration in popular music. *Social Forces*, *89*(2), 685–707. [Cited on pages 29 and 55.]

Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 145–158. Springer. [Cited on page 123.]

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proc. of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 59–68. [Cited on page 34.]

Serrano, N. (2020). Rosalía criticizes the discrimination that women continue to suffer in music when they collaborate with men. https://www.newsylist.com/rosalia-criticizes-the-discrimination-that-women-continue-to-suffer-in-music-when-they-collaborate-with-men/. [Cited on page 4.]

Shah, N. (2020). Music streaming makes major labels rich, while musicians like me go broke. https://www.theguardian.com/commentisfree/2020/dec/03/music-streaming-major-labels-musicians-uk-government. [Cited on page 4.]

Shakespeare, D., Porcaro, L., Gómez, E., & Castillo, C. (2020). Exploring artist gender bias in music recommendation. In *Proceedings of the Workshops on Recommendation in Complex Scenarios and the Impact of Recommender Systems co-located with 14th ACM Conference on Recommender Systems (RecSys 2020)*, *ComplexRec-ImpactRS 2020*, vol. 2697. CEUR-WS.org. [Cited on page 29.]

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. [Cited on pages 17 and 101.]

Singh, A. & Joachims, T. (2018). Fairness of exposure in rankings. In *Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, SIGKDD '18, pp. 2219–2228. [Cited on pages 27 and 29.]

Slaney, M. (1998). Auditory toolbox. *Interval Research Corporation, Tech. Rep*, *10*(1998). [Cited on page 102.]

Smith, J., Sonboli, N., Fiesler, C., & Burke, R. (2020). Exploring user opinions of fairness in recommender systems. [Cited on page 28.]

Smith, S. L., Choueiti, M., & Pieper, K. (2018). Inclusion in the recording studio?: Gender and race/ethnicity of artists, songwriters & producers across 600 popular songs from 2012–2017. Report, Annenberg Inclusion Initiative. Http://assets.uscannenberg.org/docs/inclusion-in-the-recording-studio.pdf. [Cited on pages 2, 29, and 56.]

Spiel, K., Haimson, O. L., & Lottridge, D. (2019). How to do better with gender on surveys: A guide for hci researchers. *Interactions*, *26*(4), 62–65. [Cited on page 61.]

Srivastava, M., Heidari, H., & Krause, A. (2019). Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2459–2468. [Cited on page 28.]

Steck, H. (2011). Item popularity and recommendation accuracy. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pp. 125–132. New York, NY, USA: ACM. [Cited on pages 25 and 139.]

Sturm, B. L. (2012). A survey of evaluation in music genre recognition. In *International Workshop on Adaptive Multimedia Retrieval*, pp. 29–66. Springer. [Cited on page 100.]

Sun, W., Khenissi, S., Nasraoui, O., & Shafto, P. (2019). Debiasing the human-recommender system feedback loop in collaborative filtering. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pp. 645–651. New York, NY, USA: ACM. [Cited on page 55.]

Surís, D., Duarte, A., Salvador, A., Torres, J., & Giró-i Nieto, X. (2018). Cross-modal embeddings for video and audio retrieval. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 0–0. [Cited on page 118.]

Tan, M. & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*. [Cited on page 98.]

Tavani, H. T. (2002). The uniqueness debate in computer ethics: What exactly is at issue, and why does it matter? *Ethics and Information Technology*, *4*(1), 37–54. [Cited on page 3.]

Turk, V. (2021). How to break out of your Spotify feedback loop and find new music. https://www.wired.co.uk/article/spotify-feedback-loop-new-music. [Cited on page 5.]

Turnbull, D. & Waldner, L. (2018). Local music event recommendation with long tail artists. [Cited on page 26.]

Turrin, R., Quadrana, M., Condorelli, A., Pagano, R., & Cremonesi, P. (2015). 30music listening and playlists dataset. In *Poster Proc. ACM Conference on Recommender Systems*, vol. 15. [Cited on pages 19, 75, 85, and 101.]

Tzanetakis, G. & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, *10*(5), 293–302. [Cited on page 124.]

Ukkonen, E., Lemström, K., & Mäkinen, V. (2003). Sweepline the music. In *Computer Science in Perspective*, pp. 330–342. Springer. [Cited on page 17.]

Vall, A., Dorfer, M., Eghbal-zadeh, H., Schedl, M., Burjorjee, K., & Widmer, G. (2019a). Feature-combination hybrid recommender systems for automated music playlist continuation. *User Modeling and User-Adapted Interaction*. [Cited on page 16.]

Vall, A., Quadrana, M., Schedl, M., & Widmer, G. (2019b). Order, context and popularity bias in next-song recommendations. *International Journal of Multimedia Information Retrieval*, *8*(2), 101–113. [Cited on pages 16 and 46.]

Van den Oord, A., Dieleman, S., & Schrauwen, B. (2013). Deep content-based music recommendation. In *Advances in neural information processing systems*, pp. 2643–2651. [Cited on pages 7, 16, 112, 113, and 121.]

Van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*. [Cited on page 118.]

Vargas, S. & Castells, P. (2011). Rank and relevance in novelty and diversity metrics for recommender systems. In *Proc. of the 5th ACM Conference on Recommender Systems*, RecSys '12, pp. 109–116. [Cited on page 22.]

Vargas, S. & Castells, P. (2013). Exploiting the diversity of user preferences for recommendation. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, pp. 129–136. [Cited on page 22.]

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*, 5998–6008. [Cited on page 121.]

Vigliensoni, G. & Fujinaga, I. (2016). Automatic music recommendation systems: Do demographic, profiling, and contextual features improve their performance? In *Proceedings of the 17th International Society for Music Information Retrieval Conference*, ISMIR '16, pp. 94–100. ISMIR. [Cited on page 14.]

Vorvoreanu, M., Zhang, L., Huang, Y.-H., Hilderbrand, C., Steine-Hanson, Z., & Burnett, M. (2019). From gender biases to gender-inclusive design: An empirical investigation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 1–14. New York, NY, USA: ACM. [Cited on page 28.]

Wall, F. (2016). Agent-based modeling in managerial science: an illustrative survey and study. *Review of Managerial Science*, *10*(1), 135–193. [Cited on pages 25 and 140.]

Wang, Q., Yu, J., & Deng, W. (2019a). An adjustable re-ranking approach for improving the individual and aggregate diversities of product recommendations. *Electron. Commer. Res.*, *19*(1), 59–79. [Cited on page 23.]

Wang, R., Harper, F. M., & Zhu, H. (2020). Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, p. 1–14. New York, NY, USA: Association for Computing Machinery. [Cited on page 27.]

Wang, S., Cao, L., Wang, Y., Sheng, Q. Z., Orgun, M., & Lian, D. (2019b). A survey on session-based recommender systems. *arXiv preprint arXiv:1902.04864*. [Cited on page 13.]

Wang, Y. & Horvát, E.-Á. (2019). Gender differences in the global music industry: Evidence from MusicBrainz and the Echo Nest. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*(01), 517–526. [Cited on page 56.]

Watson, J. (2020). Programming inequality: Gender representation on canadian country radio (2005–2019). In *Proceedings of the 21st International Society for Music Information Retrieval Conference*, ISMIR 2020, pp. 392–399. ISMIR. [Cited on page 29.]

Way, S. F., Garcia-Gathright, J., & Cramer, H. (2020). Local trends in global music streaming. In *Proc. of the International AAAI Conference on Web and Social Media*, vol. 14, pp. 705–714. [Cited on pages 26 and 48.]

Weinberger, K. Q. & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, *10*(2). [Cited on page 118.]

Weston, J., Bengio, S., & Usunier, N. (2011). Wsabie: scaling up to large vocabulary image annotation. In *Proc. of the 22nd international joint conference on Artificial Intelligence-Volume Volume Three*, pp. 2764–2770. [Cited on pages 113 and 121.]

Wiener, N. (2019). *Cybernetics or control and communication in the animal and the machine*. MIT press. [Cited on page 2.]

Won, M., Chun, S., Nieto, O., & Serrc, X. (2020a). Data-driven harmonic filters for audio representation learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 536–540. IEEE. [Cited on page 16.]

Won, M., Ferraro, A., Bogdanov, D., & Serra, X. (2020b). Evaluation of CNN-based automatic music tagging models. In *Proceedings of the SMC2020 - 17th Sound and Music Computing Conference*. [Cited on pages 16 and 123.]

Won, M., Oramas, S., Nieto, O., Gouyon, F., & Serra, X. (2020c). Multimodal metric learning for tag-based music retrieval. *arXiv preprint arXiv:2010.16030*. [Cited on pages 117, 118, and 123.]

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–14. New York, NY, USA: Association for Computing Machinery. [Cited on page 28.]

Yao, S. & Huang, B. (2017). Beyond parity: Fairness objectives for collaborative filtering. In *Proc. of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 2925–2934. [Cited on page 26.]

Youngs, I. (2019). Pop music's growing gender gap revealed in the collaboration age. *BBC*. [Cited on pages 56 and 61.]

Zamani, H., Schedl, M., Lamere, P., & Chen, C.-W. (2019). An analysis of approaches taken in the acm recsys challenge 2018 for automatic music playlist continuation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(5), 1–21. [Cited on pages 16, 112, and 113.]

Zangerle, E., Pichl, M., Gassler, W., & Specht, G. (2014). # nowplaying music dataset: Extracting listening behavior from twitter. In *Proceedings of the First International Workshop on Internet-Scale Multimedia Management*, pp. 21–26. [Cited on pages 75 and 85.]

Zhai, A. & Wu, H.-Y. (2018). Classification is a strong baseline for deep metric learning. *arXiv preprint arXiv:1811.12649*. [Cited on page 118.]

Zhang, J., Adomavicius, G., Gupta, A., & Ketter, W. (2020a). Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework. *Information Systems Research*, *31*(1), 76–101. [Cited on pages 25 and 61.]

Zhang, J., Adomavicius, G., Gupta, A., & Ketter, W. (2020b). Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework. *Information Systems Research*, *forthcoming*. [Cited on pages 85, 86, and 140.]

Zhang, Q., Cao, L., Zhu, C., Li, Z., & Sun, J. (2018). Coupledcf: Learning explicit and implicit user-item couplings in recommendation for deep collaborative filtering. In *IJCAI International Joint Conference on Artificial Intelligence*. [Cited on page 13.]

Zheng, L., Lu, C.-T., Jiang, F., Zhang, J., & Yu, P. S. (2018). Spectral collaborative filtering. In *Proceedings of the 12th ACM conference on recommender systems*, pp. 311–319. [Cited on page 13.]

Ziegler, C.-N., McNee, S. M., Konstan, J. A., & Lausen, G. (2005). Improving recommendation lists through topic diversification. In *Proc. of the 14th International Conference on World Wide Web*, WWW '14, pp. 22–32. [Cited on pages 21 and 22.]