Characterization and Mitigation of Algorithmic Bias in Recommender Systems

Elizabeth Gómez Yepes

Characterization and Mitigation of Algorithmic Bias in Recommender Systems

A dissertation submitted to the Department of Mathematics and Computer Science of the University of Barcelona in fulfillment of the requirements for the degree of

PhD in Mathematics and Computer Science

by

ELIZABETH GÓMEZ YEPES

supervised by

Dr. MARIA SALAMÓ
Dr. LUDOVICO BORATTO

UNIVERSITAT DE
BARCELONA

September, 2024

Declaration

I here by declare that the thesis entitled **"Characterization and Mitigation of Algorithmic Bias in Recommender Systems"** submitted by me, for the award of the degree of *Ph.D. in Mathematics and Computer Science* to University of Barcelona is a record of bonafide work carried out by me under the supervision of Dr. Maria Salamó Llorente, Associate Professor and Researcher, Faculty of Mathematics and Computer Science, University of Barcelona, Barcelona Spain, and Dr. Ludovico Boratto, Associate Professor and Researcher, Department of Mathematics and Computer Science, University of Cagliari, Italy.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Barcelona

Date: 20/09/2024

Elizabeth G.Y.

Signature of the Candidate

Certificate

This is to certify that the thesis entitled **"Characterization and Mitigation of Algorithmic Bias in Recommender Systems"** submitted by Ms. ELIZABETH GÓMEZ YEPES, Faculty of Mathematics and Computer Science, University of Barcelona, Spain for the award of the degree of *Ph.D. in Mathematics and Computer Science*, is a record of bonafide work carried out by her under our supervision, as per the UB code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The thesis fulfills the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place: Barcelona

Date: 20/09/2024

Signature of the Supervisors

(Dr. Maria Salamó Llorente)

(Dr. Ludovico Boratto)

Abstract

Recommender Systems are critical in helping users navigate large amounts of information by providing personalized suggestions. However, these systems can exhibit biases, especially when data imbalances exist, leading to unfair recommendations that favor more popular or majority items over those from minority groups. This thesis explores the identification, characterization, and mitigation of algorithmic bias within Recommender Systems. This research focuses on addressing biases that arise from data imbalances and how these biases can lead to unfair treatment of certain groups, particularly in terms of visibility and exposure in recommendations. The primary goal of the thesis is to mitigate algorithmic bias in Recommender Systems to produce fairer and more equitable recommendation lists, through techniques of post-processing bias mitigation (e.g., re-ranking recommendation results to ensure fairness). This includes identifying and categorizing biases in datasets, designing strategies to mitigate these biases, and developing techniques to optimize recommendation algorithms to reduce bias.

The main contributions of this thesis are five, divided into two thematic parts. The first thematic part focuses on *Provider Fairness* and the second thematic part on *Fairness from Multiple Perspectives*.

Regarding the first thematic part, **two** contributions have been made. In the first, a *Binary Approach* was adopted, by categorizing geographic bias or imbalance associated with the country of production of the items and identifying two groups of providers (majority versus rest), and based on the distribution observed in the original training set, the recommendations are adjusted to align with these groups, with the aim of mitigating disparity bias. In the second contribution, we explain the process of categorization and bias mitigation using a *Multi-Class Approach*. We explore how recommendation algorithms can exacerbate biases by promoting items from certain regions, which could disadvantage underrepresented geographic groups.

Concerning the second thematic part, **three** contributions have been made. The first contribution introduces *CONFIGRE*, a novel methodology designed to ensure fairness in Recommender Systems by balancing visibility between coarse- and fine-grained demographic groups. In second contribution we present *MOReGIn*, a new approach for managing multiple objectives in Recommender Systems. This method specifically addresses the challenge of achieving both global balance and individual fairness in recommendations. Finally, in an additional contribution, we develop a new dataset (*AMBAR*, in the music domain) that includes sensitive attributes at various levels of granularity. Furthermore, we extend two real-world datasets (MovieLens-1M and Book-Crossing) with geographic information to study the link between geographic imbalance and disparate impact.

This thesis advances on the identification, characterization, mitigation and evaluation of biases in collaborative Recommender Systems. It addresses existing gaps in the analysis of geographical biases in different group settings: from binary groups, multi-class groups to different levels of granularity of groups. The outlined contributions establish a basis for further advancements and effective mitigation of biases without significantly compromising accuracy. Our findings, developed software, and resources presented in this dissertation are available to the community to facilitate further research and knowledge transfer.

Keywords: *Thesis*, *Recommender Systems*, *Bias*, *Data Imbalance*, *Disparate Impact*, *Geographic Groups*, *Fairness*, *Calibration*.

Resumen

Los Sistemas de Recomendación son fundamentales para ayudar a los usuarios a navegar por grandes cantidades de información al ofrecer sugerencias personalizadas. Sin embargo, estos sistemas pueden presentar sesgos, especialmente cuando existen desequilibrios en los datos, lo que lleva a recomendaciones injustas que favorecen los elementos más populares o mayoritarios sobre los de los grupos minoritarios. Esta tesis explora la identificación, caracterización y mitigación del sesgo algorítmico dentro de los Sistemas de Recomendación. Esta investigación se centra en abordar los sesgos que surgen de los desequilibrios de datos y cómo estos sesgos pueden llevar a un tratamiento injusto de ciertos grupos, particularmente en términos de visibilidad y exposición en las recomendaciones. El objetivo principal de la tesis es mitigar el sesgo algorítmico en los Sistemas de Recomendación para producir listas de recomendaciones más justas y equitativas, a través de técnicas de mitigación de sesgo de posprocesamiento (por ejemplo, reclasificar los resultados de las recomendaciones para garantizar la imparcialidad). Esto incluye la identificación y categorización de sesgos en los conjuntos de datos, el diseño de estrategias para mitigar estos sesgos y el desarrollo de técnicas para optimizar los algoritmos de recomendación para reducir el sesgo.

Las principales contribuciones de esta tesis son cinco, divididas en dos partes temáticas. La primera parte temática se centra en la *Equidad del Proveedor* y la segunda parte temática en la *Equidad desde Múltiples Perspectivas*.

En relación con la primera parte temática, se han realizado **dos** contribuciones. En la primera, se adoptó un *Enfoque Binario*, categorizando el sesgo geográfico o desequilibrio asociado al país de producción de los artículos e identificando dos grupos de proveedores (mayoría versus resto), y en función de la distribución observada en el conjunto de entrenamiento original, se ajustan las recomendaciones para alinearse con estos grupos, con el objetivo de mitigar el sesgo de disparidad. En la segunda contribución, explicamos el proceso de categorización y mitigación de sesgos utilizando un *Enfoque Multi-Clase*. Exploramos cómo los algoritmos de recomendación pueden exacerbar los sesgos al promover artículos de ciertas regiones, lo que podría perjudicar a grupos geográficos subrepresentados.

En relación con la segunda parte temática, se han realizado **tres** contribuciones. La primera contribución presenta *CONFIGRE*, una nueva metodología diseñada para garantizar la equidad en los Sistemas de Recomendación al equilibrar la visibilidad entre grupos demográficos de grano grueso y fino. En la segunda contribución presentamos *MOReGIn*, un nuevo enfoque para gestionar múltiples objetivos en Sistemas de Recomendación. Este método aborda específicamente el desafío de lograr tanto el equilibrio global como la equidad individual en las recomendaciones. Finalmente, en una contribución adicional, desarrollamos un nuevo conjunto de datos (*AMBAR* para música) que incluye atributos sensibles en varios niveles de granularidad. Además, ampliamos dos conjuntos de datos del mundo real (MovieLens-1M y Book-Crossing) con información geográfica para estudiar el vínculo entre el desequilibrio geográfico y el impacto dispar.

Esta tesis avanza en la identificación, caracterización, mitigación y evaluación de sesgos en Sistemas de Recomendación colaborativos. Aborda las brechas existentes en el análisis de sesgos geográficos en diferentes configuraciones de grupos: desde grupos binarios, grupos multiclase hasta diferentes niveles de granularidad de grupos. Las contribuciones descritas establecen una base para futuros avances y una mitigación eficaz de los sesgos sin comprometer significativamente la precisión. Nuestros hallazgos, el software desarrollado y los recursos presentados en esta tesis están disponibles para la comunidad para facilitar la investigación y la transferencia de conocimientos.

Palabras clave: *Tesis*, *Sistemas de Recomendación*, *Sesgo*, *Desbalance de datos*, *Impacto dispar*, *Grupos geográficos*, *Equidad*, *Calibración*.

Acknowledgements

With immense pleasure and a deep sense of gratitude, I wish to express my sincere gratitude to my supervisors **Dr. Maria Salamó** and **Dr. Ludovico Boratto** for their unwavering guidance, patience, and dedication. Their knowledge, guidance, and encouragement allowed me to grow as a researcher and face every challenge until the completion of this process. Thank you for not giving up on me and continuing to support me until the end. It has been an honor and a privilege to work under your supervision.

I would like to acknowledge the support provided by **Mgtr. Carlos Shui Zhang** and **Dr. David Contreras** in various ways throughout my research work, enriching the work done.

To my beloved mother **Libia Yepes**, whose recent passing has left an immense void in my life. Although she is no longer physically with me, her love, teachings, and sacrifices were fundamental for me to achieve this achievement. This work is dedicated to his memory, with all my love and gratitude.

To my loving husband **Alexander Bautista**, for his unconditional support, patience, and love. You were my rock in the most difficult moments, and your faith in me pushed me to keep going when my strength was failing. This achievement is as much yours as it is mine.

To my dear friend **Dr. Deiner Mena**, thank you for always being there, giving me encouragement. Your support was essential in this process, and I cannot thank you enough for everything you did for me.

To my sister, my brother, my nephew, and my nieces, whom I love most in life. Their love and company have given me the strength to continue, and I feel deeply fortunate to have them in my life. This achievement is also yours, Family.

Last but not least, to **Minciencias** and **Colfuturo**, for giving me the opportunity to pursue these studies through their scholarship. Without your financial support, this project would not have been possible. I am deeply grateful for your trust.

Research supported by projects MISMIS-LANGUAGE (grant No. PGC2018-096212-B-C33) from the Spanish Ministry of Science and Innovation, NanoMoocs (grant No. COMRDI18-1-0010) from ACCIÓ, FairTransNLP-Language Project (MCIN/AEI/10.13039/501100011033/FEDER,UE), and by the Generalitat de Catalunya (2017 SGR 341, 2021 SGR 00313).

Place: Barcelona

Date: 20/09/2024 **ELIZABETH GÓMEZ YEPES**

Table of Contents

Appendices

List of Figures

List of Tables

Part I

PROLOGUE

CHAPTER 1

Introduction

This introductory chapter provides background information in the context of Recommender Systems that support the theoretical basis of this PhD thesis. In this, we expose the problems of bias and equity that exist in datasets and that are amplified by Recommender Systems. Specifically, we explore the open challenges in handling algorithmic bias and unfairness, that guide the main objectives of this research. Furthermore, we present our achieved contributions to these objectives and methodological strategies to tackle the issues raised. Finally, it provides an overview of the thesis structure, presented as a compendium of published articles, each contributing to the overarching aims of this doctoral thesis.

1.1 Recommender Systems

Currently, users around the world can easily and quickly access a large number of items and services through online platforms. It is precisely this quantity and variety that makes Recommender Systems not only useful but necessary [152]. Recommender Systems are software tools designed to suggest products, services, or content to users based on their preferences, previous behavior, or similar characteristics to other users [16]. These systems are widely used in platforms such as online stores, streaming services, social networks, and content websites, with the aim of personalizing the user experience and helping them discover relevant options among a large number of available options [113].

To facilitate user decision-making in the face of this information overload, the Recommender Systems, using techniques such as *collaborative filtering* [162] or *content-based filtering* [140], select items to recommend to users, thereby helping them in their decision-making by reducing the number of items to choose from. However, for these recommendations to align with needs and preferences of users, recommendation algorithms must generate personalized lists and meet the particular preferences of each user [108]. To ensure the quality of the recommendation lists, various factors are considered, such as the characteristics or content of the items, as well as the similarities between items and users [6].

In our research we focus on the use of **Collaborative Filtering** Recommender Systems. This type is based on the behavior and preferences of a group of users. That is, *it assumes that if user A has similar preferences to user B, then A may enjoy the items that B has enjoyed.* So, this method recommends items (products, movies, music, etc.) to a user based on the preferences of other similar users. For example, if two users have given similar ratings to several movies, a collaborative filtering system can recommend to one user a movie that the other has already seen and enjoyed [91]. The most used Collaborative Filtering Recommender Systems methods are: user-, item-, and factorization-based. Specifically, in *User-Based Collaborative Filtering* content is recommended to a user based on their similarity to other users. So if several similar users enjoy a certain product, that product is likely to be recommended to a user with similar tastes [151]. In particular, the *Item-Based Collaborative Filtering* focuses on the similarity between items. So, if a user has liked an item, they are recommended other similar items [159]. Finally, *Matrix Factorization Collaborative Filtering* method, which reduces the dimensionality of the user-item interaction matrix, identifying latent factors that capture the underlying relationships between users and items [108].

1.2 The Bias and Fairness Issues

Unfortunately, during the recommendation process we may encounter biases in the recommendation lists produced for users. Bias in computer systems is defined as *systematic and unfair discrimination against certain individuals or groups of individuals in favor of others* [65]. Thus, a system unfairly discriminates if it denies an opportunity or a good or assigns an undesirable outcome to an individual or group of individuals for unreasonable or inappropriate reasons [65]. Specifically, when we deal with **Algorithmic Bias**, we refer to the biases that can arise in the results of algorithms, especially in artificial intelligence and machine learning systems. These biases occur when an algorithm produces results that are systematically unfair or partial towards certain groups or individuals due to errors, incomplete or poorly represented data, or assumptions implicit in its design [74].

In the literature, according to its origin, several types of biases can be distinguished: *Pre-existing*, *Technical*, and *Emergent*. These categories describe how biases can manifest at different stages of the life cycle of an algorithmic system. *Pre-existing bias* refers to biases that are present before the system is developed and are subsequently incorporated into the system. These biases may originate in society at large, in subcultures, and in formal or informal organizations and institutions, whether private or public. They may also reflect the personal prejudices of individuals who have significant input in the the design of the system, such as the client or the system designer. This type of bias can be introduced into the system either through the conscious efforts of individuals or institutions or unconsciously, even despite the best intentions [8]. On the other hand, *Technical bias* arises from various aspects of the design process, including the limitations of computer tools like hardware, software, and peripherals; the process of attributing social meaning to algorithms developed out of context; imperfections in generating pseudo-random numbers; and the attempt to make human constructs susceptible to computers when we quantify the qualitative, discretize the continuous, or formalize the non-formal [25].

Finally, *Emergent bias* can be distinguished, which arises not at the stage of implementation of the system but during its use, as a result of changes in social knowledge, population or cultural values [132].

Regardless of the origin of the bias, when dealing with Recommender Systems, we can encounter different types of **Unfairness**, such as *Underrepresentation of Minority Groups*, this consists of certain items, services, or content are underrepresented in recommendations, it can lead to a lack of visibility for minority or niche interests. For example, if a music streaming service primarily recommends popular tracks (*Popularity Bias*), it may neglect lesser-known artists, reducing the chances for emerging or minority artists to be discovered [50]. Unluckily, unfairness tends to be magnified by the very process of recommendation, as recommending popular items to users makes them increasingly popular, which can perpetuate existing biases, while items from minority categories are seldom or never suggested to users. This situation is disadvantageous not only for users who cannot view items that offer some degree of diversity and novelty but also for individuals or companies that provide less popular products or services [67, 33].

Another type of unfairness is *Disparate Impact*, this happens when the outcomes of an algorithm disproportionately affect a particular group even if the algorithm does not explicitly target that group. This type of unfairness often arises from unintended consequences of how an algorithm processes data. For example, a credit scoring system might have a disparate impact on minority groups if it uses criteria that correlate with socio-economic disadvantages [8]. Finally, we can also mention the *Disparate Treatment*, which occurs when an algorithm treats different groups differently in a way that is not justified. This form of unfairness is often more direct and can arise from design decisions that intentionally or unintentionally lead to unequal treatment. For instance, if a loan approval algorithm has different criteria for different racial groups, this constitutes disparate treatment [37].

Another problem in recommendations is *Representation Bias*, which occurs when certain groups are underrepresented in the training data. This lack of representation can lead to poorer performance of the algorithm for the underrepresented groups. For instance, a medical diagnosis algorithm that is trained on data from predominantly one ethnic group may not perform well for other ethnic groups [26]. The representation bias produces that *Data imbalances* can naturally arise from the composition of an industry, such as when certain item categories are predominantly offered by providers of a specific gender or produced in particular regions. Addressing these imbalances in data distribution is crucial, as these patterns can become embedded in Recommender Systems, exacerbating inequalities and generating biases. If these imbalances are linked to sensitive attributes like gender or race, they can have significant societal implications, leading to unfairness. This unfairness can impact various stakeholders in a recommender system, such as *users* (when minority groups consistently receive inferior recommendations) or content *providers* (when items from certain groups of providers receive less exposure compared to others) [52, 187].

The issue of bias mitigation in Recommender Systems is crucial to ensure that the recommendations are fair, diverse, and equitable, especially in the case of data imbalance, societal biases, or algorithmic tendencies that might lead to discrimination [49]. Various techniques have been developed to address and mitigate bias in Recommender Systems such as *Pre-processing*, *In-processing* and *Post-processing*. The *Pre-processing* techniques adjust the data before training the model to reduce bias. This could involve reweighting samples or modifying features, through data rebalancing, data augmentation or fair representation learning. Other techniques are *In-processing*, which modify the algorithm during training to mitigate bias, this is done through regularization techniques, fairness constraints, adversarial debiasing or fairness-aware optimization. Finally, we have *Post-processing* techniques, these adjust the outcomes after the model has made predictions to ensure fairness. For example, applying fairness constraints to the output to correct for biases. Some post-processing techniques include result re-ranking, exposure adjustments, fairness-aware Recommender Systems (FairRec), and disparate impact remediation. In this work, the ”re-ranking post-processing technique” is the one that interests us most.

Another important factor in our research is evaluation metrics. In Recommender Systems, *visibility and exposure in rankings* are essential for ensuring equity and fairness for content providers. In our research we focused on these two metrics to assess geographic provider fairness. Since, providers depend on Recommender Systems to be seen and engaged by users, when fairness is not considered, popular providers often dominate the rankings, limiting exposure for others, especially smaller or less established ones. In some cases, fairness involves ensuring equitable treatment of providers based on certain characteristics or groupings, such as the size of the provider, their business model, or demographic attributes. The system should treat each individual provider fairly based on the quality of their content, ensuring that they are not unduly disadvantaged by factors unrelated to content relevance. Beyond individual fairness, the system should ensure that all *groups of providers* are treated fairly, with each group receiving exposure proportional to their relevance and quality. This is particularly relevant in cases where underrepresented groups of providers are at a disadvantage.

Due to the algorithmic bias issues that may arise, during this research, several databases were analyzed in order to study and categorize the biases found and how they affected the different stakeholders. Subsequently, *post-processing* techniques were proposed as solution methods to mitigate the *disparate impact* when dealing with imbalanced data by *re-ranking* the lists generated by the different recommendation models.

1.3 Open Challenges and Research Directions

Considering the bias and fairness issues mentioned above, some challenges arise to be addressed, which guide this research.

As we have explained, Recommender Systems can amplify biases, especially when dealing with unbalanced data, which is detrimental and harmful to the different stakeholders within the system. That is why the purpose addressed in this thesis was to identify algorithmic bias and propose some approaches to mitigate it, thereby obtaining fairer recommendations and providing less discriminatory treatment towards items from minority categories within a dataset and towards users with less popular needs or preferences.

One of the main parties affected by this problem is providers, because when the system disproportionately favours certain elements or providers, it causes unequal exposure among participants. Recommender Systems can be geographically biased, giving preference to content or providers from certain regions or countries, thereby disadvantaging providers from underrepresented areas. This can exacerbate global disparities, limiting the access of certain groups to global markets and opportunities for international visibility. Besides, geographic bias in Recommender Systems can contribute to reinforcing existing economic inequalities by favouring providers from developed regions while marginalising those from emerging or less developed economies. Hence, in our research we have focused on the challenge of **manage geographic bias towards providers from different geographic locations to ensure provider fairness**.

We have identified that fairness mechanisms tend to group providers into broad categories based on attributes such as gender (male or female) or age (old or young). However, it is possible to distinguish between general (coarse-grained) and specific (fine-grained) groupings. Coarse-grained groups cover broad demographic groups, while fine-grained groups focus on specific details such as age or geographic location. This granular approach allows for a more nuanced consideration of equity and fairness. The challenge is **balancing visibility between fine-grained groups and broader categories** For example, favoring one country over others within a continent can marginalize underrepresented providers.

On the other hand, Multi-Objective Recommender Systems (MORSs) aim to balance multiple goals, such as diversity, fairness, and calibration, at both global and individual levels. Global optimization ensures fairness across the system, like equitable exposure for providers, while individual optimization personalizes recommendations for users, such as tailored diversity. Most systems, however, focus on either global or individual objectives, not both. In scenarios where both are needed, focusing on one often neglects the other, making it difficult to achieve a balance between system-wide goals and personalized user needs. That is why, in this thesis we address the challenge of **addressing multiple objectives while achieving both global balance and individual equity in recommendations**.

To address these challenges, this thesis aims to analyze how data imbalances impact various stakeholder groups in Recommender Systems, particularly consumers and providers, and to mitigate the disparate impact resulting from how these systems handle such imbalances. From the providers perspective, disparate impact is primarily evaluated and mitigated by focusing on visibility and exposure. And, from the consumer perspective, the focus is on assessing and improving the effectiveness of recommendations. As a use case, we examine how geographic imbalances related to the origin of item production can lead to disparate impact on providers in recommendations. However, the attributes used to classify groups can be adapted to meet the needs or preferences of interested parties.

1.4 Objectives

This thesis concentrates on the analysis of biases and its mitigation in Recommender Systems. It investigates the impact of biases at different levels of granularities and at different stakeholders in the recommendation process. Below, we summarize the main objectives and the sub-objectives of this thesis.

O1 Design, develop and implement novel approaches to mitigate algorithmic bias in Recommender Systems to produce recommendation lists that ensure geographic provider fairness.

O1.1 Categorize and mitigate geographic disparity bias associated with the country of origin of providers groups through a Binary Approach (majority versus rest).

O1.2 Categorize and mitigate geographic disparity bias associated with the continent of origin of providers groups through a Multi-class Approach.

O2 Design, develop and implement new methodologies to mitigate algorithmic bias in Recommender Systems to promote fairness from multiple perspectives.

O2.1 Design a methodology to ensure fairness in Recommender Systems by balancing visibility between coarse- and fine-grained demographic groups.

O2.2 Design a methodology to manage multiple objectives by achieving both global balance and individual equity in recommendations.

1.5 Contributions

This section presents our contributions. First, we present our contributions concerning the **O1** and second, our contributions regarding the **O2**.

1.5.1 Contributions Regarding Provider Fairness

To address our **O1**, we made two contributions aimed at ensuring geographic provider fairness through new approaches that handle algorithmic bias in Recommender Systems. Considering the **O1.1**, the geographic bias or imbalance associated with the country of production of the items was categorized, identifying two groups of providers (majority versus rest), so we adopt a **Binary Approach**, and according to the original distribution evidenced in a training set, it is intended to adjust the recommendations to these.

One contribution has been made, which was presented in the paper *"Disparate Impact in Item Recommendation: a Case of Geographic Imbalance"*. In this, we present our research on the issue of geographic imbalance in item Recommender Systems, specifically focusing on how the country of production affects the visibility and exposure of items in recommendation algorithms. We study the impact of this imbalance in two domains, movie and book recommendations. The study highlights that items from certain countries (geographic imbalance), particularly the United States, dominate in the lists produced by the Recommender Systems. This dominance leads to a disparity in how items from other countries are recommended. We introduce two key metrics to measure the disparate impact on Recommender Systems: a) Disparate Visibility: The share of recommendations that a group (e.g., items from a particular country) receives compared to its representation in the data. b) Disparate Exposure: The position at which items from a group are recommended, influencing how likely they are to be seen and selected by users. In the study we evaluate several state-of-the-art recommendation algorithms. The analysis shows that these algorithms tend to favor items from majority countries (like the U.S.), leading to a disparate impact on items from other countries.

To address this disparity, we propose a re-ranking algorithm that adjusts the visibility and exposure of items to better reflect their representation in the dataset. This approach seeks to achieve fairness with minimal loss in recommendation effectiveness. We conclude that geographic imbalance in data can lead to significant disparities in how items are recommended. By implementing fairness-driven re-ranking strategies, these disparities can be mitigated, ensuring that items from minority regions receive fairer treatment in recommendation lists without sacrificing the quality of the recommendations. This work has been published at the European Conference on Information Retrieval Research (ECIR 2021).

- Elizabeth Gómez, Ludovico Boratto, and Maria Salamó. 2021. **Disparate Impact in Item Recommendation: A Case of Geographic Imbalance**. In: Hiemstra, D., Moens, MF., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds) Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science(), vol 12656. Springer, Cham. https://doi.org/10.1007/978-3-030-72113-8_13 — **Rank: A in CORE**.

A second article on this new approach, applied in the domain of education, was published, this was titled *"The Winner Takes it All: Geographic Imbalance and Provider (Un)fairness in Educational Recommender Systems"*. This addresses the issue of geographic imbalance and inequity for educators in educational Recommender Systems, specifically in massive open online course (MOOC) platforms. Recommender Systems in educational platforms have been primarily designed to improve the student experience, but their impact on educators has been understudied. This study focuses on how these systems can generate inequalities, particularly in terms of the visibility and exposure that courses receive based on the geographic origin of educators. The study uses data from the COCO platform, which shows a marked concentration of courses and ratings from the United States. This geographic imbalance in the data is amplified in Recommender Systems, overexposing courses from the United States and underexposing those from other regions of the world.

The paper analyzes five collaborative filtering algorithms and finds that these algorithms tend to exacerbate pre-existing disparities, offering less visibility and exposure to courses from non-US faculty. To mitigate these disparities, a recommendation re-ranking algorithm is proposed that redistributes course visibility and exposure more equitably, without compromising the effectiveness of recommendations. We conclude in this study that it is possible to improve equity in course visibility and exposure in educational Recommender Systems through adjustments to the algorithm. This work has been published in the Conference on Research and Development in Information Retrieval (SIGIR 2021).

- Elizabeth Gómez, Carlos Shui Zhang, Ludovico Boratto, Maria Salamó, and Mirko Marras. 2021. **The Winner Takes It All: Geographic Imbalance and Provider (Un)fairness in Educational Recommender Systems**. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21). Association for Computing Machinery, New York, NY, USA, 1808–1812. https://doi.org/10.1145/3404835.3463235 — **Rank: A* in CORE**.

In the context the **O1.2**, the geographic bias or imbalance associated with the continent of production of the items was categorized, identifying several groups of providers (multigroup), and according to the original distribution evidenced in a training set, it is intended to adjust the recommendations to these through a **Multi-class Approach**.

One contribution has been made. It was introduced in the paper *"Provider Fairness Across Continents in Collaborative Recommender Systems"*. In this paper, we focus on addressing fairness in Recommender Systems, particularly concerning the geographic origin of content providers. We investigate how recommendation algorithms may create biases by favoring content from certain regions over others, potentially disadvantaging smaller or less represented geographic groups, such as content from continents other than North America or Europe. The study examines the impact of geographic imbalances in two domains, movies and books. It highlights how state-of-the-art Recommender Systems tend to favor content from more represented regions (e.g., North America), leading to reduced visibility and exposure for content from less represented regions. We use visibility and exposure metrics to assess how equitably recommendations are distributed among content from different continents. Disparate visibility measures how often content from a specific region appears in recommendations, while disparate exposure assesses the position of such content within the recommendation lists.

In the paper, we contrast previous work that addressed fairness between a binary majority-minority setup (e.g., content from the US vs. the rest of the world) with a more complex multi-group setting that considers multiple continents. The findings suggest that while binary mitigation strategies can reduce some disparities, they are insufficient to ensure fairness across multiple groups. A re-ranking algorithm is proposed to adjust recommendation lists, ensuring that content visibility and exposure are proportional to the representation of each continent in the input data. Results show that the algorithm can achieve a fairer distribution of recommendations across different geographic groups, although the effectiveness of mitigation varies depending on the dataset and the specific recommendation algorithm used. We conclude that geographic imbalances in Recommender Systems can create significant disparities in the visibility and exposure of content from less represented regions. The proposed re-ranking algorithm offers a promising approach to mitigating these disparities, providing a more equitable distribution of recommendations across different geographic groups, while maintaining overall recommendation effectiveness. This work has been published in the Journal Information Processing and Management (IPM 2022).

- Elizabeth Gómez, Ludovico Boratto, Maria Salamó. **Provider fairness across continents in collaborative Recommender Systems**, Information Processing and Management, Volume 59, Issue 1, 2022, 102719, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2021.102719 — **Quartile: Q1 - Impact Factor (JCR): 8.6**.

We present a second article on this new approach, applied in the domain of education, this was titled *"Enabling Cross-continent Provider Fairness in Educational Recommender Systems"*. Here, we address the issue of fairness between teachers from different continents in educational Recommender Systems, specifically in massive open online course (MOOC) platforms. With the rise in the use of MOOCs, Recommender Systems have become key tools to support students in their learning process. However, most research has focused on students rather than the teachers who teach the courses. The paper identifies that teachers from certain geographic regions, particularly from less-represented continents, receive less visibility and exposure in these systems, limiting their opportunities.

We analyze the visibility and exposure of courses offered by teachers from different continents, using visibility and exposure metrics to assess fairness. It is observed that Recommender Systems tend to favor teachers from more represented regions, such as North America, to the detriment of those from less-represented continents. To mitigate these inequalities, in the paper we propose a recommendation re-ranking approach that adjusts the visibility and exposure of courses proportionally to their representation in the data. This technique seeks to more equitably distribute recommendations among teachers from different continents, without affecting the overall effectiveness of the recommendation system. Experiments demonstrate that the proposed approach can achieve greater equity among teachers from different continents without compromising the quality of recommendations for students. This work has been published in the Journal Future Generation Computer Systems (FGCS 2022).

- Elizabeth Gómez, Carlos Shui Zhang, Ludovico Boratto, Maria Salamó, Guilherme Ramos. **Enabling cross-continent provider fairness in educational Recommender Systems**, Future Generation Computer Systems, Volume 127, 2022, Pages 435-447, ISSN 0167-739X, https://doi.org/10.1016/j.future.2021.08.025 — **Quartile: Q1 - Impact Factor (JCR): 7.5**.

1.5.2 Contributions Regarding Fairness from Multiple Perspectives

To address our **O2**, we made three contributions to provide balance and fairness from multiple perspectives in Recommender Systems. Regarding **O2.1**, we introduce the **CONFIGRE** (COarse aNd FIne GRained Equity) Approach, which takes into account the geographical origin of the providers of the items, grouping the providers according to their continent (coarse-grained) and their country (fine-grained), and according to the original distribution shown in a training set, it is intended to adjust the recommendations to these.

We have made one contribution, this was titled *"Bringing Equity to Coarse and Fine-Grained Provider Groups in Recommender Systems"*. This paper presents a new approach to ensuring fairness in Recommender Systems by addressing the visibility of content providers across both coarse-grained (e.g., continents) and fine-grained (e.g., countries) demographic groups. Traditional fairness mechanisms in Recommender Systems often focus on coarse-grained demographic groups (like continents), which can overlook disparities at finer levels (like specific countries). This approach can lead to underrepresentation of smaller or less prominent provider groups within broader categories. We introduce CONFIGRE, a methodology designed to balance visibility across both coarse- and fine-grained provider groups. CONFIGRE operates through a re-ranking process that ensures fair exposure to items from underrepresented regions, considering both their continental and national origins. In the study, we use two key metrics, group representation and disparate visibility. Group representation measures how well different provider groups are represented in user ratings, while disparate visibility assesses the fairness of their presence in recommendation outputs.

The results show that CONFIGRE effectively reduces disparities at both the continental and national levels, outperforming existing fairness-aware algorithms in terms of providing balanced visibility. While CONFIGRE focuses on improving fairness, the study also evaluates its impact on recommendation quality, the results indicate that CONFIGRE maintains high recommendation quality while significantly enhancing equity across provider groups. Our method provides a novel solution to the challenge of ensuring fairness in Recommender Systems by addressing both broad and specific demographic categories. The approach successfully mitigates disparities in visibility at multiple levels of granularity, ensuring that even smaller provider groups receive fair representation in recommendation outputs. This work has been published at the Conference on User Modeling, Adaptation and Personalization (UMAP 2024).

- Elizabeth Gómez, David Contreras, Maria Salamo, and Ludovico Boratto. 2024. **Bringing Equity to Coarse and Fine-Grained Provider Groups in Recommender Systems**. In Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24). Association for Computing Machinery (ACM), New York, NY, USA, 18–23. https://doi.org/10.1145/3627043.3659552 — **Rank: B in CORE**.

With respect to **O2.2**, we introduce the **MOReGIn** (Multi-Objective Recommendation at the Global and Individual Levels) Approach, which takes into account the gender of the items to be recommended and the origin of the providers of these items, and according to the original distribution shown in a training set, a calibration process is carried out at the individual level for the users, so that they receive recommendations according to their gender preference of the item, and at the group level for the continents of origin of the providers of the items.

We have made one more contribution, this is *"Multi-Objective Recommendation at the Global and Individual Levels"*. The document presents a novel approach to handling multiple objectives in Recommender Systems, specifically addressing the need to balance global and individual fairness in recommendations. MORSs aim to optimize several goals simultaneously, often balancing conflicting objectives like accuracy, diversity, and fairness. Traditionally, MORSs have focused either on global objectives (affecting all users) or individual objectives (tailored to each user). The novelty of this work lies in its attempt to address both global and individual objectives concurrently. The paper highlights a gap in existing MORS approaches, they typically optimize for either global objectives, such as provider fairness, or individual objectives, such as genre calibration, but not both at the same time. Global objectives might include ensuring fair exposure of items from different providers, while individual objectives might involve aligning recommendations with each specific preferences of user. We propose the MOReGIn algorithm, which seeks to balance global and individual objectives by adjusting the recommendation lists post-processing. MOReGIn operates by categorizing items into "buckets" based on the continent of the provider (a global objective) and the genre preferences of users (an individual objective). The algorithm re-ranks the recommendation lists to ensure that the visibility of items from different providers is proportional to their representation while also aligning with user preferences.

The approach was validated using two datasets, one for movies and another for songs. The MOReGIn algorithm was tested against existing methods that focus solely on global or individual fairness. Results showed that MOReGIn outperformed baseline approaches in terms of both reducing disparity (global fairness) and minimizing miscalibration (individual fairness). While optimizing for fairness and calibration did impact recommendation accuracy slightly, the trade-off was minimal compared to the benefits gained in fairness and user satisfaction. Our approach demonstrates that it is possible to create a recommendation system that balances both global and individual objectives without significantly compromising accuracy. MOReGIn provides a more holistic solution to fairness in Recommender Systems, particularly in contexts where the provenance of content and user preferences vary widely. This work has been published at the Conference on User Modeling, Adaptation and Personalization (ECIR 2024).

- Elizabeth Gómez, David Contreras, Ludovico Boratto, Maria Salamó. 2024. **MOReGIn:**

Multi-Objective Recommendation at the Global and Individual Levels. In: Goharian, N., et al. Advances in Information Retrieval. ECIR 2024. Lecture Notes in Computer Science, vol 14608. Springer, Cham. https://doi.org/10.1007/978-3-031-56027-9_2 — **Rank: A in CORE**.

Finally, we made a one more contribution, we propose a new dataset in the music domain, named **AMBAR**. As far as we know, this is the first dataset that provides several sensitive attributes –with different levels of granularity from several perspectives: the user, the item, and the subject side. We present our new dataset in a paper titled ”AMBAR: A dataset for Assessing Multiple Beyond-Accuracy Recommenders”, which has been accepted at the 18th ACM Conference on Recommender Systems (RecSys '24), Bari, Italy. Additionally, as part of our research, we also extended two real-world datasets (MovieLens-1M and Book-Crossing) with the country and continent of production of each item and characterize the link between geographic imbalance and disparate impact, uncovering the factors that lead a group to be under-/over-exposed.

- Elizabeth Gómez, David Contreras, Ludovico Boratto, Maria Salamó. (in press). **AMBAR: A dataset for Assessing Multiple Beyond-Accuracy Recommenders**. In: 18th ACM Conference on Recommender Systems. RecSys '24. Bari, Italy, October 14–18, 2024. https://doi.org/10.1145/3640457.3688067 — **Rank: B in CORE**.

In general terms, we propose four post-processing approaches to re-rank recommendation lists that lead to mitigating disparities while causing minimal possible effectiveness losses and a new dataset, named AMBAR. During this research, an initial comparative study was proposed in which different the state-of-the-art Recommender Systems, covering both model- and memory-based approaches, and point- and pair-wise algorithms were analyzed, empirically testing the performance and categorizing the biases in the data sets used and the biases infiltrated in the generated recommendation lists. Subsequently, analyzing and delimiting the limitations of current Recommender Systems, in order to design and develop new techniques, proposing functional improvements to mitigate said biases and provide more appropriate solutions than those currently available.

1.6 Thesis Outline

This section describes the outline of the dissertation. The thesis is divided into Parts. Parts I and Part IV refer to the thesis Introduction and Conclusion respectively. The main body is divided two parts. Part II, entitled ”PROVIDER FAIRNESS” contains Chapters 2, 3, 4 and 5 and Part III, named ”FAIRNESS FROM MULTIPLE PERSPECTIVES” contains Chapters 6, 7 and 8. Below there is a description of the contents of each chapter.

Chapter 2 presents a published article on the proposed Binary Approach titled ”Disparate Impact in Item Recommendation: a Case of Geographic Imbalance”. This study focuses on how Recommender Systems can generate inequalities, particularly in terms of the visibility and exposure that items receive based on the geographic origin of providers.

Chapter 3 introduces a second article exploring Binary Approach, focused on the field of education, was published under the title “The Winner Takes it All: Geographic Imbalance and Provider (Un)fairness in Educational Recommender Systems.” The article examines the challenges of geographic disparities and inequities faced by educators in educational Recommender Systems, with a specific emphasis on massive open online course (MOOC) platforms.

Chapter 4 presents a published article on the proposed Multi-class Approach titled ”Provider Fairness Across Continents in Collaborative Recommender Systems”. This article addresses the issue of fairness between providers from different continents in Recommender Systems, the paper identifies that providers from certain geographic regions, particularly from less-represented continents, receive less visibility and exposure in these systems, limiting their opportunities.

Chapter 5 introduces a second article on the new Multi-class Approach, applied to the field of education, titled “Enabling Cross-continent Provider Fairness in Educational Recommender Systems.” In this work, we tackle the issue of fairness among teachers from different continents within educational Recommender Systems, particularly on massive open online course (MOOC) platforms.

Chapter 6 presents a published article on the proposed CONFIGRE Approach titled ”Bringing Equity to Coarse and Fine-Grained Provider Groups in Recommender Systems”. This study introduces a novel approach to promoting fairness in Recommender Systems by focusing on the visibility of content providers across different demographic levels, including broad categories such as continents and more specific ones like countries.

Chapter 7 presents a published article on the proposed MOReGIn Approach titled ”MOReGIn: Multi-Objective Recommendation at the Global and Individual Levels”. The paper introduces a new method for managing multiple objectives in Recommender Systems, with a particular focus on balancing global fairness and individual fairness in the recommendations.

Chapter 8 presents an accepted paper about the newly collected dataset titled ”AMBAR: A dataset for Assessing Multiple Beyond-Accuracy Recommenders”. We introduce a new dataset in the music domain, that provides several sensitive attributes with different levels of granularity from several perspectives: the user, the item, and the subject side.

Chapter 9 discusses the conclusions of the thesis and provides future research directions.

Appendix A contains supplementary information for Chapter 4.

Part II

PROVIDER FAIRNESS

CHAPTER 2

Disparate Impact in Item Recommendation: a Case of Geographic Imbalance

This chapter contains the paper entitled ”Disparate Impact in Item Recommendation: a Case of Geographic Imbalance”, which presents the Binary approach published at the European Conference on Information Retrieval Research (ECIR 2021). This research examines how Recommender Systems can create inequalities, specifically regarding the visibility and exposure of items, which can be influenced by the geographic origin of the providers. The published research manuscript included in this chapter is the following:

- Elizabeth Gómez, Ludovico Boratto, and Maria Salamó. 2021. **Disparate Impact in Item Recommendation: A Case of Geographic Imbalance**. In: Hiemstra, D., Moens, MF., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds) Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science(), vol 12656. Springer, Cham. https://doi.org/10.1007/978-3-030-72113-8_13 — **Rank: A en CORE**.

Disparate Impact in Item Recommendation: a Case of Geographic Imbalance

Recommender Systems are key tools to push items' consumption. Imbalances in the data distribution can affect the exposure given to providers, thus affecting their experience in online platforms. To study this phenomenon, we enrich two datasets and characterize data imbalance w.r.t. the country of production of an item (*geographic imbalance*). We focus on movie and book recommendation, and divide items into two classes based on their country of production, in a majority-versus-rest setting. To assess if Recommender Systems generate a disparate impact and (dis)advantage a group, we introduce metrics to characterize the visibility and exposure a group receives in the recommendations. Then, we run state-of-the-art Recommender Systems and measure the visibility and exposure given to each group. Results show the presence of a disparate impact that mostly favors the majority; however, factorization approaches are still capable of capturing the preferences for the minority items, thus creating a positive impact for the group. To mitigate disparities, we propose an approach to reach the target visibility and exposure for the disadvantaged group, with a negligible loss in effectiveness.

Keywords: Recommender Systems, Bias, Disparate Impact.

2.1 Introduction

Recommender Systems learn patterns from users' behavior, to understand what might be of interest to them [153]. Natural imbalances in the data (e.g., in the amount of observations for popular items) might be embedded in the patterns. The produced recommendations can amplify these imbalances and create biases [19]. When a bias is associated to sensitive attributes of the users (e.g., gender or race), negative societal consequences can emerge, such as unfairness [86, 63, 126, 146]. Unfairness can affect all the stakeholders of a system [3, 11].

Data imbalances might be inherently connected to the way an industry is composed, e.g., with certain items mainly produced in certain parts of the world, and with consumption patterns that differ based on the country of the users [10]. In this paper, we focus on geographic imbalance and study the problem of how the country of production of an item can create a disparate impact to providers in the recommendations. We assess disparate impact by considering both the *visibility* received by the providers of a group (i.e., the percentage of recommendations having them as providers) and their *exposure*, which accounts for the position in which items are recommended [167]. Hence, with these two metrics we measure respectively, (i) the share of recommendations of a group and (ii) the relevance that is given to that group. Both metrics are important to assess disparate impact in this context. Visibility alone might lead a group of providers not being reached by users in case they appear only at the bottom of the list, and exposure alone might not guarantee providers enough sales (a single item at the top of the list would mean these providers are recommended only once).

We assess disparate impact by comparing the visibility and exposure given to a group of providers with the representation of the group in the data. We study two forms of representation, based on (i) the amount of items a group offers, or (ii) the amount of ratings given to the items of a group.

We consider two of the main domains in which Recommender Systems operate, namely movies and books. We show, by extending two real-world datasets with the country of production of the items, that both movie and book data is imbalanced towards the United States. To understand the impact of this imbalance, we divide items into two groups, in a majority-versus-rest setting, and study how this imbalance is reflected in the visibility and exposure given to providers of the two groups when producing recommendations.

We consider state-of-the-art Recommender Systems, covering both model- and memory-based approaches, and point- and pair-wise algorithms. While commonly studied sensitive attributes, such as gender, show a disparate impact effect at the expense of the minority group, our use-case presents several peculiarities. Indeed, user preferences do not reflect these imbalances and users equally like items coming from the majority (the United States) and the minority (the rest of the countries) groups. This leads to disparity scenarios that affect either the majority or the minority group, according to patterns we present in this study.

To mitigate disparities, we propose a re-ranking that optimizes both the visibility and exposure given to providers, based on their representation in the data. Hence, we consider a distributive norm based on *equity* [176]. Our approach introduces in the recommendations items that increase the visibility and exposure of a group, causing the minimum possible loss in user relevance.

Our contributions can be summarized as follows:

- We study, for the first time, the impact of geographic imbalance in the data on the visibility and exposure given to different provider groups;
- We extend two real-world datasets with the country of production of each item and characterize the link between geographic imbalance and disparate impact, uncovering the factors that lead a group to be under-/over-exposed;
- We propose a re-ranking mitigation strategy that can lead to the target visibility and exposure with the minimum possible losses in effectiveness;
- We evaluate our approach, showing we can mitigate disparities with a negligible loss in effectiveness.

The rest of the paper details in §2.2 related work, while in §2.3 the scenario, metrics, recommenders, and datasets. §2.4 assesses disparate impact phenomena. §2.5 contains our mitigation algorithm and results. §2.6 concludes the paper.

2.2 Related Work

This section covers related studies, starting from the concepts of visibility and exposure in ranking, and continuing with the impact of recommendation for providers. We conclude by contextualizing our work with the existing studies.

Visibility and exposure in rankings. Given a ranking, visibility and exposure metrics respectively assess the amount of times an item is present in the rankings [59, 192] and *where* an item is ranked [15, 191]. They were introduced in the context of non-personalized rankings, where the objects being ranked are individual users (e.g., job candidates). These metrics can operate at the *individual* level, thus guaranteeing that similar individuals are treated similarly [15, 45], or at *group* level, by making sure that users belonging to different groups are given adequate visibility or exposure [45, 192, 191]. Under the group setting, the visibility/exposure of a group is proportional to its representation in the data [139, 158, 185, 147].

Impact of recommendations for providers. The impact of the generated recommendations on the item providers is a concept known as *provider fairness (P-fairness).* It guarantees that the providers of the recommended objects that belong to different groups or are similar at the individual level, will get recommended according to their representation in the data. In this domain, Ekstrand et al. [51] assessed that collaborative filtering methods recommend books of authors of a given gender with a distribution that differs from that of the original user profiles. Liu and Burke [120] propose a re-ranking function, which balances recommendation accuracy and fairness, by dynamically adding a bonus to the items of the uncovered providers. Sonboli and Burke [168] define the concept of local fairness, to equalize access to capital across all types of businesses. Mehrotra et al. [130] assess unfairness based on the popularity of the providers. Several policies are defined to study the trade-offs between user-relevance and fairness. Kamishima et al. [102] introduce recommendation independence, which leads to recommendations that are statistically independent of sensitive features.

Contextualizing our work. While our study draws from metrics derived from fairness, *this work does not directly mitigate fairness for the individual providers*. We study a broader phenomenon, i.e., *if an industry of a country is affected by how recommendations are produced in presence of data imbalance*. Considering our use-cases, both cinema and literature are powerful vehicles for culture, education, leisure, and propaganda, as highlighted by the UNESCO[1]. Moreover, both domains have an impact on the economy of a country, with (sometimes public) investments for the production of movies/books that are expected to generate a return. Hence, considering how Recommender Systems can push the consumption of items of a country is a related but different problem w.r.t. provider fairness.

2.3 Preliminaries

Here, we present the preliminaries, to provide foundations to our work.

2.3.1 Recommendation Scenario

Let $U = \{u_1, u_2, ..., u_n\}$ be a set of users, $I = \{i_1, i_2, ..., i_j\}$ be a set of items, and V be a totally ordered set of values that can be used to express a preference. The set of ratings is a ternary relation $R \subseteq U \times I \times V$; each rating is denoted by r_{ui}. These ratings can directly feed an algorithm in the form of triplets (point-wise approaches) or shape user-item observations (pair-wise approaches).

To assess the real impact of the recommendations, we consider a temporal split of the data, where a fixed percentage of the ratings of the users (ordered by timestamp) goes to the training and the rest goes to the test set [12].

The recommendation goal is to learn a function f that estimates the relevance ($\hat{r}_{ui}$) of the user-item pairs that do not appear in the training data. We denote as $\hat{R}$ the set of recommendations, and as $\hat{R}_G$ those involving items of a group G.

Let C_i be the set of production countries of an item i. We use it to shape two groups, a majority $M = \{i \in I : 1 \in C_i\}$, and a minority $m = \{i \in I : 1 \notin C_i\}$. Note that 1 identifies

[1] `https://publications.parliament.uk/pa/cm200203/cmselect/cmcumeds/667/667.pdf`

the country associated to the majority group.

2.3.2 Metrics

Representation. We consider as representation of a group ias the percentage of the input data that involves that group. Concretely, given a group G, we define two forms of representation, based on (i) the percentage of items offered by a group and (ii) the percentage of ratings collected for that group. We denote as $\mathcal{R}$ the *representation* of a group G ($G \in \{M, m\}$) ($\mathcal{R}_I$ denotes an item-based representation, while $\mathcal{R}_R$ a rating-based representation):

$$\mathcal{R}_I(G) = |G|/|I| \tag{2.1}$$

$$\mathcal{R}_R(G) = |\{r_{ui} : i \in G\}|/|R| \tag{2.2}$$

Eq. (2.1) accounts for the proportion of items of a group, while Eq. (2.2) for the proportion of ratings associated to a group. Both metrics are between 0 and 1.

The representation of a group is measured by considering only the training set. It is trivial to notice that, given a group G, the representation of the other, $\overline{G}$, can be computed as $\mathcal{R}_*(\overline{G}) = 1 - \mathcal{R}_*(G)$ (where '*' refers to I or R).

Disparate Impact. We assess disparate impact with two metrics.

Definition 2.3.1 (Disparate visibility)**.** *The* disparate visibility *of a group is computed as the difference between the share of recommendations for items of that group and the representation of that group:*

$$\Delta\mathcal{V}(G) = \frac{1}{|U|}\sum_{u \in U} \frac{|\{\hat{r}_{ui} : i \in \hat{R}_G\}|}{|\hat{R}|} - \mathcal{R}_*(G) \tag{2.3}$$

Its range is in $[-\mathcal{R}_*(G), 1 - \mathcal{R}_*(G)]$; it is 0 when there is no disparate visibility, while negative/positive values indicate that the group received a share of recommendations lower/higher than its representation. This metric is based on that considered by Fabbri et al. [59].

Definition 2.3.2 (Disparate exposure)**.** *The* disparate exposure *of a group is the difference between the exposure obtained by the group in the recommendation lists [167] and the representation of that group:*

$$\Delta\mathcal{E}(G) = \frac{1}{|U|}\sum_{u \in U} \frac{\sum_{pos=1}^{k} \frac{1}{log_2(pos+1)}, \forall i \in \hat{R}_G}{\sum_{pos=1}^{k} \frac{1}{log_2(pos+1)}} - \mathcal{R}_*(G) \tag{2.4}$$

where pos is the position of an item in the top-k recommendations.

This metric also ranges in $[-\mathcal{R}_*(G), 1-\mathcal{R}_*(G)]$; it is 0 when there is no disparate exposure, while negative/positive values indicate that the exposure given to the group in the recommendations is lower/higher than its representation.

Notice that the disparate visibility/exposure of one group can be computed as the opposite of the value obtained for the other group.

Remark. *We do not define a unique "disparate impact" metric, to control both visibility and exposure, so that* providers are recommended enough times and with enough exposure. *A unique metric would not allow us to balance both, by compressing everything in a unique number.*

2.3.3 Recommendation Algorithms

We consider five state-of-the-art Collaborative Filtering algorithms. As memory-based approaches, we consider the UserKNN [92] and ItemKNN [159] algorithms. For the class of matrix factorization based approaches, we consider the BPR [149], BiasedMF [108], and SVD++ [107] algorithms. To contextualize our results, we also consider two non-personalized algorithms (MostPopular and RandomGuess).

2.3.4 Datasets

MovieLens-1M (Movies). The dataset provides 1M ratings (range 1-5), provided by 6,040 users, to 3,662 movies. It contains the IMDb ID of each movie, which allowed us to associate it to its country of production thanks to the OMDB APIs[2] (note that *each movie may have more than one country of production*).

Book Crossing (Books). The dataset contains 356k ratings (in the range 1-10), given by 10,409 users, to 14,137 books. The dataset contained the ISBN code of each book, which was used to add information about its countries of production thanks to the APIs offered by the Global Register of Publishers[3].

For both datasets, we encoded the country of production with an integer, with the United States (which represents the majority group in both datasets) having ID 1, and the rest of the countries having subsequent IDs.

2.4 Disparate Impact Assessment

In this section, we run the algorithms presented in Section 2.3.3 to assess their effectiveness and the disparate impact they generate.

[2] http://www.omdbapi.com/

[3] https://grp.isbn-international.org/search/piid_cineca_solr

2.4.1 Experimental Setting

For both datasets presented in Section 2.3.4, the test set was composed by the most recent 20% of the ratings of each user. To run the recommendation algorithms presented in Section 2.3.3, we considered the LibRec library (version 2). For each user, we generate 150 recommendations (denoted in the paper as the top-n) so that we can mitigate disparate impact through a re-ranking algorithm. The final recommendation list for each user is composed by 20 items (denoted as top-k).

Each algorithm was run with the following hyper-parameters:

- **UserKNN.** similarity: Pearson; neighbors: 50; similarity shrinkage: 10;
- **ItemKNN.** similarity: Cosine for Movies and Pearson for Books; neighbors: 200 (Movies), 50 (Books); similarity shrinkage: 10;
- **BPR.** iterator learnrate: 0.1; iterator learnrate maximum: 0.01; iterator maximum: 150; user regularization: 0.01; item regularization: 0.01; factor number: 10; learnrate bolddriver: false; learnrate decay=1.0;
- **BiasedMF.** iterator learnrate: 0.01; iterator learnrate maximum: 0.01; iterator maximum: 20 (Movies), 1 (Books); user regularization: 0.01; item regularization: 0.01; bias regularization: 0.01; number of factors: 10; learnrate bolddriver: false; learnrate decay: 1.0;
- **SVD++.** iterator learnrate: 0.01; iterator learnrate maximum: 0.01; iterator maximum: 10 (Movies), 1 (Books); user regularization: 0.01; item regularization: 0.01; impItem regularization: 0.001; number of factors: 10; learnrate bolddriver: false; learnrate decay: 1.0.

To evaluate recommendation effectiveness, we measure the ranking quality of the lists by measuring the *Normalized Discounted Cumulative Gain* (NDCG) [99].

$$DCG@k = \sum_{u \in U} \hat{r}_{ui}^{pos} + \sum_{pos=2}^{k} \frac{\hat{r}_{ui}^{pos}}{log_2(pos)} \qquad NDCG@k = \frac{DCG@k}{IDCG@k} \tag{2.5}$$

where $\hat{r}_{ui}^{pos}$ is relevance of item i recommended to user u at position pos. The ideal DCG is calculated by sorting items based on decreasing true relevance (true relevance is 1 if the user interacted with the item in the test set, 0 otherwise).

2.4.2 Characterizing User Behavior

This section characterizes the group representation and users' rating behavior.

Group representation. In the Movies dataset, $\mathcal{R}_I(m) = 0.3$ and $\mathcal{R}_R(m) = 0.23$. In the Books dataset, instead, $\mathcal{R}_I(m) = 0.12$ and $\mathcal{R}_R(m) = 0.08$. Both datasets show a strong geographic imbalance, with the majority group covering 70% of the items in the first dataset and 88% in the second. This imbalance is worsened when we consider the ratings, since in the movie context the ratings associated to the majority are 77%, while in the book content the rating representation for the majority is 92%. It becomes natural to ask ourselves if the majority group also attracts better ratings, to assess if this exacerbated imbalance is because majority items are perceived as of higher quality.

Rating behavior. We considered the average rating associated to the items of each group. In the Movies dataset, the average rating for the majority group is 3.56, while that of the minority group is 3.61. In the Books dataset, we observed an average rating of 4.38 for the majority, and of 4.43 for the minority. This shows that the preference of the users for the two groups does not differ.

> **Observation 1**. *Both datasets expose a big geographic imbalance in the representation of each group, in terms of offered items. The majority group usually attracts more ratings, thus increasing the existing imbalance. However, the minority items are not considered as of lower quality for the users, since the average rating for both groups is the same in both datasets.*

Table 2.1 Results of state-of-the-art Recommender Systems. Normalized Discounted Cumulative Gain (NDCG); Disparate Visibility for the minority group when considering the item representation as a reference ($\Delta\mathcal{V}_I$) ; Disparate Exposure for the minority group when considering the item representation as a reference ($\Delta\mathcal{E}_I$); Disparate Visibility for the minority group when considering the rating- representation as a reference ($\Delta\mathcal{V}_R$) ; Disparate Exposure for the minority group when considering the rating representation as a reference ($\Delta\mathcal{E}_R$). The values in bold indicate the best result.

	Movies					Books				
Algorithm	**NDCG**	$\Delta\mathcal{V}_I$	$\Delta\mathcal{E}_I$	$\Delta\mathcal{V}_R$	$\Delta\mathcal{E}_R$	**NDCG**	$\Delta\mathcal{V}_I$	$\Delta\mathcal{E}_I$	$\Delta\mathcal{V}_R$	$\Delta\mathcal{E}_R$
MostPop	0.1109	-0.1802	-0.2016	-0.1089	-0.1302	0.0089	-0.1239	-0.1239	-0.0839	-0.0840
RandomG	0.0105	**0.0020**	**0.0027**	0.0733	0.0740	8.91E+11	**0.0013**	**0.0015**	0.0412	0.0415
UserKNN	0.1247	-0.1544	-0.1668	-0.0831	-0.0955	0.0053	-0.0438	-0.0360	**-0.0039**	**0.0039**
ItemKNN	0.1199	-0.1744	-0.1926	-0.1031	-0.1212	0.0075	-0.0799	-0.0790	-0.0400	-0.0390
BPR	**0.1395**	-0.1054	-0.1087	**-0.0340**	**-0.0373**	0.0054	-0.0257	-0.0259	0.0142	0.0141
BiasedMF	0.0588	0.0901	0.0954	0.1614	0.1668	**0.0103**	-0.1239	-0.1239	-0.0840	-0.0840
SVD++	0.0684	0.0742	0.0762	0.1455	0.1475	**0.0103**	-0.1239	-0.1239	-0.0840	-0.0840

2.4.3 Assessing Effectiveness and Disparate Impact

We assess disparate impact in terms of visibility and exposure. Table 2.1 presents the results obtained when generating a top-20 ranking for each user, considering as a reference the minority group. The first phenomenon that emerges is that both groups can be affected by disparate impact and that, when one group receives more visibility, it also receives more exposure; hence, when a group is favored in the amount of recommendations, it is also ranked higher.

Considering the Movies dataset, MostPop, UserKNN, ItemKNN, and BPR present a disparate visibility and exposure that disadvantage the minority, for both forms of representation.

The point-wise Matrix Factorization algorithms (BiasedMF and SVD++) and RandomGuess, instead, advantage the minority. This goes in contrast with the literature on algorithmic bias and fairness, where the minority is usually disadvantaged. We conjecture that, since Recommender Systems do not receive any information about the geographic groups and since users equally prefer the items of the two groups, the point-wise Matrix Factorization approaches create factors that capture user preferences as a whole. Our results align with those of Cremonesi et al. [38], who showed the capability of factorization approaches to recommend long-tail items. Interestingly, when considering disparate visibility and exposure, the best results for the item-based representation are those of RandomGuess; nevertheless, the algorithm is also the least effective in terms of NDCG. No algorithm can offer both effectiveness and adapt to the offer of a country. When considering the rating-based representation, BPR is the most effective and has the lowest disparate visibility and exposure. Hence, the combination between factorization approaches and a pair-wise training can connect effectiveness and equity of visibility and exposure.

In the Books dataset, besides MostPop, all the approaches advantage the majority. This opposite trend in terms of disparate impact of the point-wise Matrix Factorization algorithms (BiasedMF and SVD++) w.r.t. the Movies dataset, can be explained by considering that the items having more ratings will lead to factors that have more weight at prediction stage; here, the majority is much larger than in the Movies dataset, so this leads to the group being advantaged in terms of visibility and exposure. This dataset is much also more sparse, so effectiveness is strongly reduced, and the point-wise Matrix Factorization approaches are the most effective. There is no connection between effectiveness and equity of exposure and visibility. Indeed, RandomGuess and UserKNN are, respectively, the best algorithms when considering the item-/rating-based representation of the groups. This good visibility and exposure provided by UserKNN in the rating-based setting can be connected to phenomena observed by Cañamares and Castells [29] since, under sparsity, the algorithm adapts to item popularity.

Observation 2. *Geographic imbalance almost always affects the minority group, since we feed algorithms with much more instances than their counterpart. Matrix Factorization based approaches can help the minority receive more visibility and exposure, with latent factors that capture preferences also of the minority. However, if the imbalance is too severe, the minority is always affected by disparate impact.*

2.5 Mitigating Disparate Impact

The previous section allowed us to observe a new phenomenon that departs from the existing algorithmic fairness studies, since *the minority group is not always the disadvantaged one when considering geographic imbalance*. Still, our results show that we can always observe a group receiving a disproportional visibility and exposure with respect to its representation in the data.

Table 2.2 Impact of mitigation on recommended lists with item-based representation. Normalized Discounted Cumulative Gain (NDCG); Disparate Visibility ($\Delta\mathcal{V}_I$) for the minority; Disparate Exposure ($\Delta\mathcal{E}_I$) for the minority. We report below gain/loss of each setting w.r.t. the original one (left side of Table 2.1).

	Mitigation Visibility and Exposure					
	Movies			**Books**		
Algorithm	**NDCG**	$\Delta\mathcal{V}_I$	$\Delta\mathcal{E}_I$	**NDCG**	$\Delta\mathcal{V}_I$	$\Delta\mathcal{E}_I$
MostPop	0.1052	-0.0017	-0.0017	0.0087	-0.0039	-0.0039
(gain/loss)	-0.0057	0.1785	0.1999	-0.0002	0.1200	0.1200
RandomG	0.0106	-0.0017	-0.0017	8.91E+11	-0.0039	-0.0039
(gain/loss)	0.0001	-0.0036	-0.0043	3.24E+09	-0.0052	-0.0055
UserKNN	0.1205	-0.0017	-0.0017	0.0050	-0.0039	-0.0039
(gain/loss)	-0.0042	0.1528	0.1652	-0.0003	0.0399	0.0321
ItemKNN	0.1173	-0.0017	-0.0017	0.0075	-0.0039	-0.0039
(gain/loss)	-0.0027	0.1727	0.1909	0.0000	0.0760	0.0751
BPR	**0.1372**	**-0.0017**	**-0.0017**	0.0055	-0.0039	-0.0039
(gain/loss)	-0.0023	0.1037	0.1070	0.0001	0.0218	0.0220
BiasedMF	0.0623	-0.0017	-0.0017	**0.0119**	**-0.0039**	**-0.0039**
(gain/loss)	0.0035	-0.0918	-0.0971	0.0016	0.1200	0.1200
SVD++	0.0712	-0.0017	-0.0017	0.0113	-0.0039	-0.0039
(gain/loss)	0.0028	-0.0759	-0.0779	0.0011	0.1200	0.1200

In this section, we mitigate these phenomena by presenting a re-ranking algorithm that introduces items of the disadvantaged group in the recommendation list, to reach a visibility and an exposure proportional to its representation.

A re-ranking algorithm is the only option when optimizing ranking-based metrics, like visibility and exposure. An in-processing regularization, such as those presented in [102, 14], would not be possible, since at prediction stage the algorithm does not predict *if and where* an item will be ranked in a list. Re-rankings have been introduced to reduce disparities, both for non-personalized rankings [192, 167, 15, 31, 191, 139] and for Recommender Systems [130, 27], with approaches such as Maximal Marginal Relevance [30]. These algorithms optimize only one property (visibility or exposure), so no direct comparison is possible.

2.5.1 Algorithm

The foundation behind our mitigation algorithm is to *move up in the recommendation list the item that causes the minimum loss in prediction for all the users.* We start by targeting the desired visibility, to make sure the items of the disadvantaged group are recommended enough times. Then we move items up inside the recommendation list to reach the target exposure.

The mitigation is described in Algorithm 1. The inputs are the recommendations (top-n items), the current visibility and exposure of the disadvantaged group and its representation in the data (our target), and the IDs of the advantaged and disadvantaged groups. The output is the re-ranked list of items.

Table 2.3 Impact of mitigation on recommended lists with rating-based representation. Normalized Discounted Cumulative Gain (NDCG); Disparate Visibility ($\Delta\mathcal{V}_R$) for the minority; Disparate Exposure ($\Delta\mathcal{E}_R$) for the minority. We report below gain/loss of each setting w.r.t. the original one (left side of Table 2.1).

	Mitigation Visibility and Exposure					
	Movies			**Books**		
Algorithm	**NDCG**	$\Delta\mathcal{V}_R$	$\Delta\mathcal{E}_R$	**NDCG**	$\Delta\mathcal{V}_R$	$\Delta\mathcal{E}_R$
MostPop	0.1076	-0.0003	-0.0003	0.0089	-0.0040	-0.0040
(gain/loss)	-0.0032	0.1085	0.1299	-0.0006	0.0800	0.0800
RandomG	0.0112	-0.0003	-0.0003	8.54E+11	-0.0040	-0.0040
(gain/loss)	0.0006	-0.0736	-0.0743	-2.37E+10	-0.0452	-0.0455
UserKNN	0.1239	-0.0003	-0.0003	0.0050	-0.0040	-0.0040
(gain/loss)	-0.0008	0.0828	0.0952	-0.0003	-0.0001	-0.0079
ItemKNN	0.1185	-0.0003	-0.0003	0.0075	-0.0040	-0.0040
(gain/loss)	-0.0015	0.1027	0.1209	0.0001	0.0360	0.0351
BPR	**0.1390**	**-0.0003**	**-0.0003**	0.0053	-0.0040	-0.0040
(gain/loss)	-0.0005	0.0337	0.0370	-0.0001	-0.0182	-0.0180
BiasedMF	0.0648	-0.0003	-0.0003	**0.0122**	**-0.0040**	**-0.0040**
(gain/loss)	0.0060	-0.1618	-0.1671	0.0016	0.0800	0.0800
SVD++	0.0735	-0.0003	-0.0003	0.0113	-0.0040	-0.0040
(gain/loss)	0.0051	-0.1459	-0.1479	0.0011	0.0800	0.0800

The $optimizeVisibilityExposure$ method (lines 1-6), executes the mitigation, firstly to regulate the visibility of the disadvantaged group (by adding their items to the top-k) and secondly to regulate the exposure (by moving their items up in the top-k). The $mitigation$ method (lines 7-23) regulates the visibility and exposure of the recommendation list. First, we loop over the users (lines 9-11) and call the $calculateLoss$ method, to calculate the loss (in terms of items' predicted relevance) we would have in each user's list when swapping the items of the two groups. The while loop (lines 12-21) swaps the items until the target visibility/exposure is reached; line 13 returns the user that causes the minimum loss and line 14 swaps their items. If the goal is to reach a target visibility, lines 15-16 increase the visibility of the group by 1; if the swap is done to reach a target exposure, lines 17-19 subtract the exposure of the old item and add that of the new one. Finally, the $calculateLoss$ method recalculates the loss for the user object of the swap and returns the re-ranked list.

The $calculateLoss$ method (lines 24-37) identifies the user causing the minimal loss of predicted relevance. We select two items in the list of each user. The first is the last item of the advantaged group in the top-k (line 26). If we are regulating visibility, lines 27-28 select the first item of the disadvantaged group out of the top-k (denoted as last-n). Lines 29-33 mitigate for exposure; the while selects an item of the disadvantaged group that in the top-k is currently ranked lower than that of its counterpart. Once we obtain the pair of items for the user, we calculate the loss by considering the $prediction$ attribute (line 34). Finally, line 35 collects the loss of the user, which is returned in line 36.

2.5.2 Impact of Mitigation

In this section, we assess the impact of our mitigation. Since we split data temporally, we cannot run statistical tests to assess the difference in the results, so we highlight the gain/loss obtained for each measure.

Results are reported in Tables 2.2 and 2.3 separating them between item- and rating-based representation of the groups. Trivially, given a target representation and a dataset, all algorithms achieve the same disparate visibility/exposure. Let us consider the trade-off between disparate visibility/exposure and effectiveness. Considering the Movies dataset, in both representations of the groups, BPR is the algorithm with the best trade-off between effectiveness and equity of visibility and exposure. It was already the most accurate algorithm, and thanks to our mitigation based on the minimum-loss principle, the loss in NDCG was negligible. In the Books dataset, BiasedMF confirms to be the best approach, in both effectiveness and equity of visibility and exposure. It is interesting to observe that, in both scenarios, MostPop is the second most effective algorithm and now provides the same visibility and exposure as the other algorithms; this is due to popularity bias phenomena [1], and their analysis is left as future work.

> **Observation 3**. *When providing a re-ranking based on minimal predicted loss, the effectiveness remains stable, but disparate visibility and disparate exposure are mitigated.*

2.6 Conclusions and Future Work

In this paper, we considered data imbalance in the items' country of production of items (*geographic imbalance*). We considered a group setting based on a majority-versus-rest split of the items and defined measures to assess disparate visibility and disparate exposure for groups. The results of five collaborative filtering approaches show that the minority group is not always disadvantaged.

We proposed a mitigation algorithm that produces a re-ranking, by adding to the recommendation lists items that cause the minimum loss in predicted relevance. Results show that *thanks to our approach, any recommendation algorithm can bring equity of visibility and exposure to providers, without impacting the end-users in terms of effectiveness*.

Future work will study geographic imbalance in education, to explore country-based disparities for teachers [9, 44, 42, 43]. Moreover, we will evaluate divergence-based disparity metrics [39]) and consider multi-class group settings. Other issues emerging from imbalanced groups, such as bribing [161, 148], will be considered.

```
Input: $recList$: ranked list (records contain $user$, $item$, $prediction$, $exposure$, $group$, $position$), $vis$:
       visibility of disadvantaged group, $exp$: exposure of disadvantaged group, $rep$: representation of
       disadvantaged group, $advG$: ID of advantaged group, $disadvG$: ID of disadvantaged group
Output: $reRankedList$: ranked list adjusted by visibility and exposure
define optimizeVisibilityExposure ($recList$, $vis$, $exp$, $rep$)
begin
    $reRankedList \leftarrow$ mitigation($recList$, $vis$, $rep$, $advG$, $disadvG$, ”visibility”)
    $reRankedList \leftarrow$ mitigation($reRankedList$, $exp$, $rep$, $advG$, $disadvG$, ”exposure”)
    return $reRankedList$
end
define mitigation ($list$, $VE$, $rep$, $advG$, $disadvG$, $rankingType$)
begin
    for $user \in list.users$ do
        $losses.add$(calculateLoss($list$, $user$, $rankingType$, $advG$, $disadvG$)
    end
    while $VE < rep$ do
        $minLoss \leftarrow losses.sortByLoss(0)$
        $list \leftarrow$ swap($list$, $minlLoss.itemAdvG$, $minLoss.itemDisadvG$)
        if $reRankingType ==$ “$visibility$” then
            $VE \leftarrow VE + 1$
        else
            $VE \leftarrow (VE - minLoss.itemDisadvG.exposure) + minLoss.itemAdvG.exposure$
        end
        $losses.add$(calculateLoss($list$, $user$, $rankingType$, $advG$, $disadvG$))
    end
    return $list$
end
define calculateLoss ($list$, $user$, $rankingType$, $advG$, $disadvG$)
begin
    $itemAdvGroup \leftarrow$ getlastItem($list$, $user$,top-k,$advGroup$)
    if $reRankingType ==$ “$visibility$” then
        $itemDisadvGroup \leftarrow$ getfirstItem($list$, $user$,last-n,$disadvGroup$)
    else
        while $itemAdvGroup.position > itemDisadvGroup.position$ do
            $itemDisadvGroup \leftarrow$ getnextItem($list$, $user$,top-k,$disadvGroup$)
        end
    end
    $loss \leftarrow itemAdvGroup.prediction$ - $itemDisadvGroup.prediction$
    $lossUser \leftarrow [user, itemAdvGroup, itemDisadvGroup, loss]$
    return $lossUser$
end
```

Algorithm 1: Visibility and exposure mitigation algorithm

CHAPTER 3

The Winner Takes it All: Geographic Imbalance and Provider (Un)fairness in Educational Recommender Systems

This chapter contains the paper entitled ”The Winner Takes it All: Geographic Imbalance and Provider (Un)fairness in Educational Recommender Systems”, which presents the Binary approach published in the Conference on Research and Development in Information Retrieval (SIGIR 2021). This article addresses the issue of geographic imbalance and inequity faced by educators in educational Recommender Systems, particularly on massive open online course (MOOC) platforms.

- Elizabeth Gómez, Carlos Shui Zhang, Ludovico Boratto, Maria Salamó, and Mirko Marras. 2021. **The Winner Takes It All: Geographic Imbalance and Provider (Un)fairness in Educational Recommender Systems**. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’21). Association for Computing Machinery, New York, NY, USA, 1808–1812. https://doi.org/10.1145/3404835.3463235 — **Rank: A* en CORE**.

The Winner Takes it All: Geographic Imbalance and Provider (Un)fairness in Educational Recommender Systems

The impact that educational Recommender Systems are having for learners is channeling most research efforts on the effectiveness of the recommended entities. While teachers have a central role in these platform, the impact of Recommender Systems for teachers in terms of the exposure they give to the courses is an under-explored area. In this paper, we consider data coming from a real-world platform and analyze the distribution of the recommendations w.r.t. geographical provenience of the teachers. We observe that data is highly imbalanced towards the United States, both in terms of offered courses and of interactions. These imbalances are exacerbated by Recommender Systems, which overexpose the country w.r.t. its representation in the data, thus generating unfairness for teachers outside the country. To introduce equity, we propose an approach that regulates the share of recommendations given to the items produced in a country (*visibility*) and the positions in which items are ranked in the recommendation list (*exposure*).

Keywords: Provider Fairness, Educational Recommender Systems, Data Imbalance.

3.1 Introduction

Learning paradigms are shifting towards online environments [103], thanks to Massive Online Open Courses (MOOC) platforms. The recent pandemic has dramatically accelerated the use of these platforms, which are reporting a 25-30% increase [166]. In this scenario, Recommender Systems are the means that allows MOOC platforms to direct appropriate resources to learners [19]. Course recommendation is by far the most common in these platforms, with a clear focus on the learners and the opportunities that are offered to them [127].

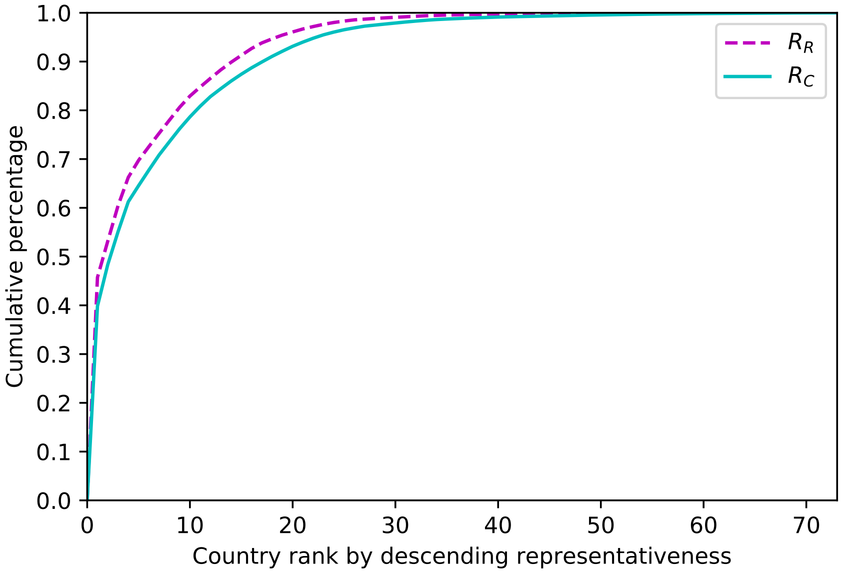


Fig. 3.1 Country imbalance. Cumulative percentage of ratings (in purple) and online courses (in blue) for each country in COCO [43].

The impact of Recommender Systems have on teachers is an under-explored perspective. However, teachers are a key stakeholders in a MOOC platform, since they are the ones that provide the courses, and they are directly affected by the way recommendation lists are shaped. Indeed, according to how many times the courses of a teacher are recommended (*visibility*) [59] and where they appear in the ranking, that teacher is given a certain *exposure* by the system [167]. Disparities in the visibility and exposure given to teachers might lead to undesired consequences, such as unfairness [86]. In this paper, we focus on *group unfairness*, shaping groups based on the *geographic provenience* of the teachers offering the courses. Our goal is to study if imbalances in the country of provenience of the teachers might affect the opportunities of teachers coming from certain parts of the world to offer their services, by being under-exposed. Specifically, we consider two demographic groups, the first covering the country with the highest representation of teachers in the platform (in our data, the United States), and the second containing the rest of the world. There are multiple reasons why this is an interesting setting. Considering the reference dataset for this study, COCO [43] (presented in

Section 3.2.4), Fig. 3.1 shows that United States cover more than 40% of the courses and nearly 50% of the ratings ($\mathcal{R}_R$ and $\mathcal{R}_C$ respectively characterize the percentage of ratings and courses that a country attracted, as presented in Section 3.2.2). The remaining 73 countries attract a very small percentage of ratings and courses, thus leading to an important *geographic imbalance* in the input data. However, in a binary setting such as the one we consider, the most represented country does not constitute an overall majority in the data. This offers an interesting benchmark to study the interplay between geographic imbalance and minority groups and their impact on unfairness.

If Recommender Systems overexpose teachers coming from the country with the highest representation, teachers from the rest of the world are unfairly affected by how recommendations are generated. In this work, we consider five state-of-the-art collaborative filtering models, covering both memory- and model-based approaches and point- and pair-wise approaches. We show that Recommender Systems exacerbate disparities emerging from geographic imbalance, under-exposing the teachers coming from the rest of the world. To overcome these phenomena, we propose a re-ranking approach that aims to re-distribute the recommendations between the United States and the rest of the world, following a notion of *equity* [176].

Specifically, our contributions are as follows: (i) we assess how Recommender Systems affect groups of teachers based on their provenience, (ii) we propose an approach to introduce equity in the recommendations' distribution, and (iii) we show that we can introduce equity without affecting recommendation effectiveness.

3.2 Preliminaries

3.2.1 Recommendation Scenario

Let $U = \{u_1, u_2, ..., u_n\}$ be a set of learners, $C = \{c_1, c_2, ..., c_j\}$ be a set of courses, and V be a totally ordered set of values used to express a preference. The set of ratings is a ternary relation $R \subseteq U \times C \times V$; each rating is denoted by r_{uc}.

We consider a temporal split of the data, where a fixed percentage of the ratings of the learners (ordered by timestamp) goes to the training and the rest goes to the test set [12].

The recommendation goal is to learn a function f that estimates the relevance ($\hat{r}_{uc}$) of the learner-course pairs that do not appear in the training data. We denote as $\hat{R}$ the set of recommendations, and as $\hat{R}_G$ those involving courses of a group G.

Let $A = \{a_1, a_2, ..., a_k\}$ be the set of geographic areas in which courses are organized. Specifically, we consider a geographic area as the country associated to a course. We denote as A_c the set of geographic areas of a course c. Note that, since teachers of a course could be from different geographical areas, several geographic areas may appear in a course. We shape two groups, the most represented area, $M = \{i \in I : 1 \in A_i\}$, and the rest, $m = \{i \in I : 1 \notin A_i\}$. Note that 1 identifies the most represented country.

3.2.2 Metrics

Representation. The representation of a group is the amount of times that group appears in the data. We consider two forms of representation, based on (i) the amount of courses offered by a group and (ii) the amount of ratings collected for that group. We define with $\mathcal{R}$ the *representation* of a group G ($\mathcal{R}_C$ denotes a course-based representation, while $\mathcal{R}_R$ a rating-based representation):

$$\mathcal{R}_C(G) = |G|/|C| \tag{3.1}$$

$$\mathcal{R}_R(G) = |\{r_{uc} : c \in G\}|/|R| \tag{3.2}$$

Eq. (3.1) accounts for the proportion of courses of a group, while Eq. (3.2) for the proportion of ratings associated to a group. The representation of a group is measured by considering only the training set. Given a group G, the representation of the other, $\overline{G}$, is computed as $\mathcal{R}_*(\overline{G}) = 1 - \mathcal{R}_*(G)$ (where '$*$' refers to C or R).

Disparate Impact. We assess unfairness with two notions of *disparate impact* generated by a recommender system.

Definition 3.2.3 (Disparate visibility)**.** *The* disparate visibility *of a group is the difference between the share of recommendations for items of that group and the representation of that group [59]:*

$$\Delta\mathcal{V}(G) = \frac{1}{|U|}\sum_{u \in U} \frac{|\{\hat{r}_{uc} : c \in \hat{R}_G\}|}{|\hat{R}|} - \mathcal{R}_*(G) \tag{3.3}$$

Its range is in $[-\mathcal{R}_*(G), 1 - \mathcal{R}_*(G)]$; it is 0 when there is no disparate visibility, while negative/positive values indicate that the group had a share of recommendations lower/higher than its representation.

Definition 3.2.4 (Disparate exposure)**.** *The* disparate exposure *of a group is the difference between the exposure obtained by the group in the recommendations [167] and the representation of that group:*

$$\Delta\mathcal{E}(G) = \frac{1}{|U|}\sum_{u \in U} \frac{\sum_{pos=1}^{k} \frac{1}{log_2(pos+1)}, \forall c \in \hat{R}_G}{\sum_{pos=1}^{k} \frac{1}{log_2(pos+1)}} - \mathcal{R}_*(G) \tag{3.4}$$

where pos is the position of an item in the top-k recommendations.

This metric also ranges in $[-\mathcal{R}_*(G), 1-\mathcal{R}_*(G)]$; it is 0 when there is no disparate exposure, while negative/positive values indicate that the exposure given to the group in the recommendations is lower/higher than its representation.

Notice that the disparate visibility/exposure of one group can be computed as the opposite of the value obtained for the other group.

3.2.3 Recommendation Algorithms

We consider five state-of-the-art Collaborative Filtering models. As memory-based approaches, we consider the UserKNN [92] and ItemKNN [159] algorithms. For the class of matrix factorization based approaches, we consider the BPR [149], BiasedMF [108], and SVD++ [107] algorithms. To contextualize our results, we consider two non-personalized algorithms (Most Popular and Random Guess).

3.2.4 Dataset

To the best of our knowledge, COCO [43] is the only educational dataset that contains the geographic provenience of the users. It was collected from an online course platform, and each course is associated to one or more teachers, belonging to 74 countries.

We pre-processed the dataset to remove all learners with less than 3 ratings. Our final dataset contains 12,472 courses and 298,644 learners, which provided 1,296,598 ratings. We encoded the country of a course with an integer, with the United States having ID 1, and the rest having subsequent IDs.

Other educational datasets, proposed by [61, 195, 144], include $(learner, course, rating)$ triplets only, as needed in traditional recommendation scenarios, thus not fitting the problem tackled in this study (the teachers' sensitive attributes are not available).

3.3 Disparate Impact Assessment

3.3.1 Experimental Setting

The test is composed by the most recent 20% of the ratings of each learner. We run the recommendation algorithms using the LibRec library (v.2). For each user, we store the first 100 results (top-n) to then mitigate disparities through a re-ranking. The recommendation list for each learner is composed by 20 courses (top-k).

Each algorithm was run with the following hyper-parameters: (i) **UserKNN.** similarity: Pearson; neighbors: 50; similarity shrinkage: 10; (ii) **ItemKNN.** similarity: Cosine; neighbors: 200; similarity shrinkage: 10; (iii) **BPR.** iterator learnrate: 0.01; iterator learnrate maximum: 0.01; iterator maximum: 100; user regularization: 0.01; item regularization: 0.01; factor number: 10; learnrate bolddriver: false; learnrate decay=1.0; (iv) **BiasedMF.** iterator learnrate: 0.01; iterator learnrate maximum: 0.01; iterator maximum: 10; user regularization: 0.01; item regularization: 0.01; bias regularization: 0.01; number of factors: 10; learnrate bolddriver: false; learnrate decay: 1.0; (v) **SVD++.** iterator learnrate: 0.01; iterator learnrate maximum: 0.01; iterator maximum: 13; user regularization: 0.01; item regularization: 0.01; impItem regularization: 0.001; number of factors: 10; learnrate bolddriver: false; learnrate decay: 1.0.

To evaluate recommendation quality, we measure the NDCG [99].

3.3.2 Characterizing User Behavior

In COCO, $\mathcal{R}_C(M) = 0.41$ and $\mathcal{R}_R(M) = 0.47$. Considering that the dataset contains 74 countries, we observe a strong geographic imbalance in terms of offered courses. This imbalance is worsened when we consider the ratings. We analyzed the learners behind these ratings, and observed that 24.5% of the raters are from the United States. Previous studies on this dataset [19] show that the vast majority of the ratings is 5. Also under this geographical setting, user satisfaction is equally distributed along the two groups.

> **Observation 1**. *There is a strong geographic imbalance in the representation of each group, in terms of offered items. The most represented group usually attracts more ratings, thus increasing the existing imbalance. There are cultural aspects behind this imbalance, with one fourth of the ratings coming from learners of the most represented country.*

Table 3.1 Effectiveness, disparate visibility, and disparate exposure of group m, considering both a course- and a rating-based representation of the groups.

Algorithm	**NDCG**	$\Delta\mathcal{V}_C$	$\Delta\mathcal{E}_C$	$\Delta\mathcal{V}_R$	$\Delta\mathcal{E}_R$
MostPop	0.0193	-0.3091	-0.2117	-0.2447	-0.1473
RandomG	0.0006	**0.0000**	**-0.0001**	0.0644	0.0643
UserKNN	0.0372	-0.0402	-0.1457	0.0242	-0.0813
ItemKNN	**0.2068**	-0.0862	-0.0783	-0.0218	-0.0139
BPR	0.1401	-0.0715	-0.0658	**-0.0071**	**-0.0014**
BiasedMF	0.0007	-0.1065	-0.0949	-0.0421	-0.0305
SVD++	0.0044	-0.0534	-0.0543	0.0110	0.0101

3.3.3 Assessing Effectiveness and Disparities

In Table 3.1, we report the results obtained by each model. Results show that ItemKNN is the most effective algorithm; considering that rating distribution in this dataset is skewed towards high values, these results connect to widely-known phenomena that make the algorithm successful [136], such as the size of the data we are working with and the fact that the neighborhoods will not change much, given that the ratings are very similar. We conjecture that these are the main drivers towards sound and accurate predictions. When considering a course-based representation, Random Guess is the algorithm providing the most equitable visibility and exposure. Hence, when picking the items to recommend at random, the recommendation lists are shaped following the distribution in the course offer; nevertheless, this is the algorithm returning the lowest effectiveness. Finally, BPR is the approach returning the most equitable recommendations when considering a rating-based representation. We connect these results to those of [38] [38], who showed that factorization approaches are able to build factors that capture all user preferences, leading to the recommendation of long-tail items. BPR also returns the second best NDCG.

> **Observation 2**. *Geographic imbalance leads to disparate visibility and exposure at the advantage of the most represented group. Recommendation effectiveness is decoupled from equity of visibility and exposure, with BPR returning the best trade-off between the two properties in the course-based representation.*

3.4 Mitigating Disparate Impact

We mitigate disparities with a re-ranking algorithm that introduces items of the disadvantaged group in the recommendation list.A re-ranking is the only option when optimizing ranking-based metrics, like visibility and exposure. An in-processing regularization, such as [102, 14], would not be possible, since at prediction stage a model does not predict *if and where* an item will be ranked. Re-rankings have been employed to reduce disparities, both for non-personalized rankings [192, 167, 15, 31, 191, 139] and for Recommender Systems [130, 27], with approaches such as Maximal Marginal Relevance [30]. These optimize either visibility or exposure, so no comparison is possible.

3.4.1 Algorithm

The idea behind our mitigation algorithm is to *move up in the recommendation list the course that causes the minimum loss in prediction for all the learners*. Algorithm 2 describes the process; the input is a recommendation list for all the learners (the top-n items) and the output is the re-ranked list of courses. The complete mitigation process is divided into three methods.

The first, $optimizeVisibilityExposure$ (lines 1-6), starts the mitigation. It makes two interventions: one based on visibility and the second one based on exposure. The second method, called $mitigation$ (lines 7-29), regulates the visibility and exposure inside the recommendation list. The $checkPosition$ method (lines 30-34) is responsible for checking the position of an item in the list, taking into account if we perform a visibility- or exposure-based mitigation. The role of each line is commented in blue in the algorithm.

3.4.2 Impact of Mitigation

Tables 3.2 and 3.3 report the results after mitigating considering the course- and rating-based representations of the groups. Given the temporal split of the data, we cannot perform statistical tests to validate the results so, under each metric, we report the gain/loss obtained after running our mitigation. Our results present a general pattern, which leads us to our third observation.

Table 3.2 Results for group m of the mitigation based on $\mathcal{R}_C$, both after optimizing for Visibility and after optimizing for Exposure (here, we report only the NDCG and the disparate exposure; visibility, by design, remains the same).

	Visibility			**Exposure**	
Algorithm	**NDCG**	$\Delta\mathcal{V}_C$	$\Delta\mathcal{E}_C$	**NDCG**	$\Delta\mathcal{E}_C$
MostPop	0.0181	0.0000	-0.0924	0.0166	0.0000
(gain/loss)	-0.0012	0.3091	0.1193	-0.0027	0.2117
RandomG	0.0006	0.0000	-0.0001	0.0006	0.0000
(gain/loss)	0.0000	0.0000	0.0000	0.0000	0.0000
UserKNN	0.0369	0.0000	-0.0233	0.0360	0.0000
(gain/loss)	-0.0003	0.0402	0.1225	-0.0012	0.1457
ItemKNN	**0.2061**	0.0000	-0.0301	**0.2038**	0.0000
(gain/loss)	-0.0008	0.0862	0.0481	-0.0030	0.0782
BPR	0.1395	0.0000	-0.0288	0.1375	0.0000
(gain/loss)	-0.0006	0.0714	0.0370	-0.0026	0.0658
BiasedMF	0.0007	0.0000	-0.0266	0.0006	0.0000
(gain/loss)	-0.0001	0.1064	0.0682	-0.0001	0.0948
SVD++	0.0043	0.0000	-0.0063	0.0043	0.0000
(gain/loss)	-0.0001	0.0534	0.0480	-0.0001	0.0543

Table 3.3 Results for group m of the mitigation based on $\mathcal{R}_R$, both after optimizing for Visibility and after optimizing for Exposure (here, we report only the NDCG and the disparate exposure; visibility, by design, remains the same).

	Visibility			**Exposure**	
Algorithm	**NDCG**	$\Delta\mathcal{V}_R$	$\Delta\mathcal{E}_R$	**NDCG**	$\Delta\mathcal{E}_R$
MostPop	0.0186	0.0000	-0.0986	0.0178	0.0000
(gain/loss)	-0.0007	0.2448	0.0488	-0.0015	0.1474
RandomG	0.0006	0.0000	0.0217	0.0006	0.0000
(gain/loss)	0.0000	-0.0644	-0.0426	0.0000	-0.0643
UserKNN	0.0369	0.0000	-0.0237	0.0364	0.0000
(gain/loss)	-0.0003	-0.0241	0.0577	-0.0008	0.0814
ItemKNN	**0.2068**	0.0000	-0.0113	**0.2061**	0.0000
(gain/loss)	0.0000	0.0219	0.0026	-0.0007	0.0139
BPR	0.1401	0.0000	-0.0083	0.1396	0.0000
(gain/loss)	0.0000	0.0071	-0.0069	-0.0005	0.0015
BiasedMF	0.0007	0.0000	-0.0060	0.0007	0.0000
(gain/loss)	0.0000	0.0421	0.0245	0.0000	0.0305
SVD++	0.0044	0.0000	-0.0575	0.0045	0.0000
(gain/loss)	0.0000	-0.0110	-0.0675	0.0001	-0.0101

Observation 3. *When providing a re-ranking based on minimal predicted loss, effectiveness remains stable, but disparate visibility and disparate exposure are mitigated. Interventions to adjust both visibility and exposure are needed to provide equity; if we mitigate only having a visibility goal, disparate exposure (fourth column, in red, in Tables 3.2 and 3.3) still occurs.*

3.5 Conclusions and Future Work

In this paper, we considered course Recommender Systems, with a focus on how teachers can be affected by the way courses are geographically distributed. Considering a real-world dataset coming from a MOOC platform, we assessed that our most represented country (the United States) is over-exposed by state-of-the-art recommendation models, thus affecting the teachers from the rest of the world. To overcome this issue, we proposed a re-ranking approach that aims to provide equity, by reaching the target visibility and exposure while causing the minimum loss in relevance. Results show that we can provide equity of visibility and exposure without affecting recommendation effectiveness.

This work sheds light on the unfairness generated by educational Recommender Systems and offers a first solution to mitigate these phenomena. Based on this work, in future work we will go beyond this type of mitigation of the disparities, to re-distribute the recommendations in equitable ways between the individual countries, taking into account for multiple aspects (e.g., the language of the courses offered in non-English countries and that of the learners). This opens interesting research scenarios to mitigate unfairness in educational Recommender Systems at a finer granularity.

```
Input: $recList$: ranked list (records contain $user$, $item$, $prediction$, $exposure$, $position$)
Output: $reRankedList$: ranked list adjusted by visibility and exposure
define optimizeVisibilityExposure ($recList$)
begin
    $reRankedList \leftarrow$ mitigation($recList$, “visibility”) ; // mitigation to target the desired visibility
    $reRankedList \leftarrow$ mitigation($reRankedList$, “exposure”) ; // mitigation to regulate the exposure
    return $reRankedList$ ; // return the re-ranked list
end
define mitigation ($recList$, $reRankingType$) // add the courses of the disadvantaged group to the top-$k$
begin
    for $user \in list.users$ do // for each user
        for $item \in$ top-n do // we loop over all items that belong to this user
            if checkPosition(item, itemsOut, reRankingType) is True then // check the position
                itemsOut.add(item) ; // add the item as possible candidate to move out to the list
            else if checkPosition(item, itemsOut, reRankingType) is False then
                itemsIn.add(item) ; // add the item as possible candidate to move in to the list
            end
        end
        while itemsIn not empty and itemsOut not empty do // computes all possible swaps and the loss of each one
            $itemsIn \leftarrow itemsIn.pop(first)$ ; $itemsOut \leftarrow itemsOut.pop(last)$ ;
             $loss \leftarrow itemsOut.last - itemsIn.first$;
            possibleSwaps.add(id,user,itemsOut.last,itemsIn.first,loss);
        end
    end
    if reRankingType == “visibility” then sortByLoss(possibleSwaps); // sort by loss in case of visibility ;
    else if reRankingType == “exposure” then sortByExposureLoss(possibleSwaps); // sort by exposure loss in case of exposure ;
    while $proportions < targetProportions$ and possibleSwaps not empty do // do swaps until the target proportions are reached
        $list \leftarrow$ swap($list, itemOut, itemIn$) ; // makes the swap of the candidate with minor loss
        $proportions \leftarrow updateProportions(itemOut, itemIn, reRankingType)$; // updates group proportions
    end
    return $list$; // returns the re-ranked list
end
define checkPosition($item, itemsOut, reRankingType$) // check the position of an item in the list
begin
    if reRankingType == “visibility” then return $item.position <$ top-k ;
    else if reRankingType == “exposure” then return $item.position < itemsOut.last.position$ ;
end
```

Algorithm 2: Visibility and exposure mitigation algorithm.

CHAPTER 4

Provider Fairness Across Continents in Collaborative Recommender Systems

This chapter contains the paper entitled ”Provider Fairness Across Continents in Collaborative Recommender Systems”, which presents the Multi-class approach published in the Journal Information Processing and Management (IPM 2022). This article explores the issue of fairness among providers from different continents in Recommender Systems, highlighting that providers from certain geographic regions, particularly underrepresented continents, receive less visibility and exposure in these systems. The published research manuscript included in this chapter is the following:

- Elizabeth Gómez, Ludovico Boratto, Maria Salamó. **Provider fairness across continents in collaborative Recommender Systems**, Information Processing and Management, Volume 59, Issue 1, 2022, 102719, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2021.102719 — **Quartile: Q1 - Impact Factor (JCR): 8.6**.

Provider Fairness Across Continents in Collaborative Recommender Systems

When a recommender system suggests items to the end-users, it gives a certain exposure to the providers behind the recommended items. Indeed, the system offers a possibility to the items of those providers of being reached and consumed by the end-users. Hence, according to how recommendation lists are shaped, the experience of under-recommended providers in online platforms can be affected. To study this phenomenon, we focus on movie and book recommendation and enrich two datasets with the continent of production of an item. We use this data to characterize imbalances in the distribution of the user-item observations and regarding where items are produced (*geographic imbalance*). To assess if Recommender Systems generate a disparate impact and (dis)advantage a group, we divide items into groups, based on their continent of production, and characterize how represented is each group in the data. Then, we run state-of-the-art Recommender Systems and measure the visibility and exposure given to each group. We observe disparities that favor the most represented groups. We overcome these phenomena by introducing equity with a re-ranking approach that regulates the share of recommendations given to the items produced in a continent (*visibility*) and the positions in which items are ranked in the recommendation list (*exposure*), with a negligible loss in effectiveness, thus controlling fairness of providers coming from different continents. A comparison with the state of the art shows that our approach can provide more equity for providers, both in terms of visibility and of exposure.

Keywords: Recommender Systems, Bias, Provider Fairness, Geographic Groups, Data Imbalance, Disparate Impact.

4.1 Introduction

Recommender Systems support users by suggesting items that might be of interest to them [153]. This is usually done by learning behavioral patterns from historical data, usually in the form of user-item interactions. However, imbalances in the input data can lead to biases in the results these algorithms produce [17]. The main example of this type of phenomenon is popularity bias, where popular items get over-recommended, to the detriment of long-tail ones [21]. If bias is associated with sensitive attributes of the users (such as gender or race), biased results might lead to unethical consequences, such as discrimination (*unfairness*) [17, 86]. Discrimination might affect both the end-users (often referred to as consumers), when those belonging to legally protected groups or certain individuals receive systematically worse recommendations (*consumer fairness*), and content producers, in case the items of those belonging to legally protected groups or individuals are under-recommended by an algorithm (*provider fairness*) [17, 11]. However, there are scenarios in which a recommender system works with imbalanced data not only because of a biased data collection, but because of the way an industry is composed. A clear example of this is the modern film industry, where the United States Cinema (Hollywood) takes the largest share of the market, both in terms of produced movies and of revenues[1]. Moreover, as observed by Bauer and Schedl, users belonging to different geographic areas have different item consumption patterns [10].

Given these considerations, it becomes natural to ask ourselves *if data imbalances associated with an industry can lead to unfairness for providers, according to the way recommendations are produced*. Specifically, we consider providers belonging to different geographic areas and assess if Recommender Systems exacerbate the natural imbalances existing in the input data, thus affecting the producers of smaller industries from a geographic point of view. In our recent work [78], we considered a binary setting, in which item producers were divided into two groups, a *majority* containing the items coming from the main country of production of the items, and a *minority* containing items produced in the rest of the world. We assessed how state-of-the-art collaborative filtering algorithms distributed the recommendations, and observed that the majority items are over-represented in the recommendation lists, both in terms of the number of recommendations (*visibility*) and in their position in the rankings (*exposure*). We presented an approach that redistributes the recommendations, so that the majority group has a representation in the recommendations that corresponds to that in the input data.

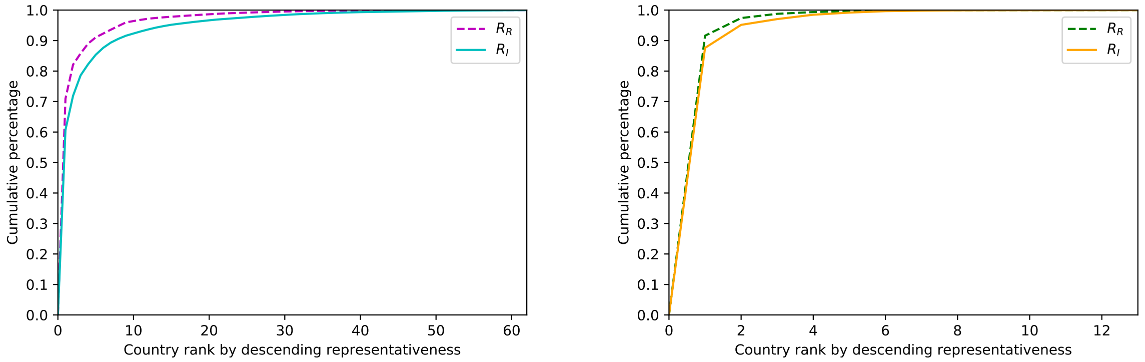
However, when dealing with provider fairness, it is important to *understand how recommendations are distributed across different provider groups*. Indeed, even if we ensure that the providers in the majority group are not over-recommended (as we did in [78]), we still do not have guarantees that the different provider groups belonging to the minority are recommended in equitable ways. In the context of geographic groups, this means that *the problem of how recommendations of items produced by providers in small regions are distributed, and of how*

[1] https://www.boxofficepro.com/mpa-2019-global-box-office-and-home-entertainment-surpasses-100-billion/

to mitigate possible disparities, remains open.

Unfair recommendations for providers, based on their geographic provenience, is an issue that goes beyond a biased functioning of an automated decision-support system and has consequences at multiple levels, by denying the opportunity to providers to offer their items (*ethical perspective*), thus limiting their possibility to work (*business perspective*); on top of this, unfair outputs are also forbidden by current regulations, such as GDPR (*legal perspective*). Hence, ensuring that providers coming from different parts of the world are not affected by the fact that they belong to a region that has a low share in the market, is a problem of central importance.

To address this problem, in this paper, we move from a binary to a multi-group setting, to study *unfairness for providers belonging to different continents*. We consider two of the main recommendation domains, namely movies and books, and assess how state-of-the-art collaborative filtering algorithms distribute the recommendations. We observe that both the original models and the mitigation we introduced for binary groups create disparities in both the visibility and exposure given to content providers in different continents and that the less represented is a group in the data, the worse is this disparity created by a recommendation model. To overcome these phenomena, we propose an approach that introduces fairness for providers belonging to different geographic areas, by re-distributing the recommendations across continents following a notion of *equity* [176]. Concretely, our mitigation strategy gives a provider group a visibility and an exposure equal to the representation of the group in the input data.



(a) Country representation in MovieLens-1M.

(b) Country representation in Book-Crossing.

Fig. 4.1 Country representation in the input data. Representation of each country in the MovieLens-1M (a) and Book-Crossing (b) datasets. Representation is computed either considering the amount of items produced by a country ($\mathcal{R}_I$) or the amount of ratings it attracted ($\mathcal{R}_R$).

Our choice to introduce equity considering a continent as our granularity level is motivated by Fig. 4.1, where we show the representation of each country of production in the input data. The first thing that emerges is that the dataset we considered in the movie domain (MovieLens-1M) contains items produced in over 60 countries. Hence, introducing equity at the country level would require adjusting the recommendation lists to ensure that each of these countries received a representation equal to its representation in the input data. While this is challenging, due to the large number of countries and the limited size of a recommendation list, a second issue emerges when we observe how represented is each country in the data, both considering the number of items it produced ($\mathcal{R}_I$) and the number of ratings it received ($\mathcal{R}_R$)[2]. Again, from the MovieLens-1M dataset, we can see that the imbalance in the representation is severe, with the main country (the United States) who produced over 60% of the movies and attracted around 70% of the ratings. Given the very small representation of almost all the other countries, a regulation at the country level would lead to a severe drop in the recommendation effectiveness for the users. Readers can also see that, while our second dataset, Book-Crossing, has much fewer countries, the imbalance in the data is even more severe. Hence, introducing provider fairness at the continent level allows us to work with a stable setting and to contrast among them the results obtained with the two datasets.

Our contributions can be summarized as follows:

- We assess unfairness for groups of providers belonging to different geographic continents, considering state-of-the-art recommendation models;
- We propose a re-ranking algorithm to introduce provider fairness for multiple groups, following a notion of equity that distributes the recommendations according to the representation of the groups in the input data;
- We evaluate our algorithm in two recommendation domains and study its effectiveness at producing fair but effective recommendations.

Concretely, we extend the study presented in [78] in the following ways: (i) we extend our related work, to improve the coverage of the existing literature; (ii) we analyze how our previous proposal deals with more than two groups; (iii) we introduce a new problem setting, which has a multi-group fairness goal; (iv) we introduce a new algorithm to introduce equity for more than two provider groups; (v) we add a comparison with the state of the art, to show why a mitigation tacking both visibility and exposure is needed to ensure provider fairness.

The rest of the paper is structured as follows: in Section 4.2, we present related work and in Section 4.3 we provide the foundation to our work. We continue by presenting, in Section 4.4, a summary of our work in [78], to have a reference on provider fairness for two provider groups in this context. Given this setting, in Section 4.5, we assess the capability of the state-of-the-art models and of our binary mitigation to provide fairness to groups shaped at continent level. In Section 4.6, we propose a mitigation algorithm to overcome unfairness scenarios in presence of multiple groups and we evaluate it in Section 4.7. Finally, we provide concluding remarks in Section 4.8.

[2] Our two notions of group representation, $\mathcal{R}_I$ and $\mathcal{R}_R$, are formally presented in Section 4.3.1.

4.2 Related Work

This section covers related studies. First of all, Section 4.2.1 starts from the concepts of visibility and exposure in ranking. Next, in Section 4.2.2, we continue with the impact of recommendation for providers. Finally, Section 4.2.3 concludes by contextualizing our work with the existing studies.

4.2.1 Visibility and Exposure in Rankings

Given a ranking, visibility measures the amount of times an item is presented in the rankings [59, 192] whereas exposure assesses *where* an item is ranked [15, 191]. Visibility and exposure were first introduced in the context of non-personalized rankings, where the objects being ranked are individual users, such as job candidates. These metrics can operate at the *individual* or *group* levels.

At the *individual* level these metrics are devoted to guaranteeing that similar individuals are treated similarly [15, 45]. For instance, Biega et al. [15] defined measures to capture unfairness at the level of individual subjects. Conversely, at the *group* level these metrics make sure that users belonging to different groups are given adequate visibility or exposure [45, 192, 191]. One example is the ranked group fairness definition presented in [192], which extends group fairness using the standard notion of protected groups. Zehlike and Castillo [191] describe an approach that measures discrimination and unequal opportunity in rankings at training time in terms of discrepancies in the average group exposure.

Under the group setting, the visibility/exposure of a group is proportional to its representation in the data [139, 158, 185, 147]. Since our study considers group settings, we embrace this class of metrics, assessing both the visibility and exposure in the recommendation lists.

4.2.2 Impact of Recommendations for Providers

The concepts of visibility and exposure have a direct impact on the providers behind the recommended items. *Provider fairness (P-fairness)* is the impact of the generated recommendations on the item providers. It guarantees that the providers of the recommended objects that belong to different groups are similar at the individual level, will get recommended according to their representation in the data. Provider fairness was mostly tackled through post-processing approaches.

Defining when a user or a group of users gets discriminated by an Artificial Intelligence (AI) system highly depends on the context that is being studied [93, 85, 42, 9].

In the context of books recommendation, Ekstrand et al. [51] assessed that collaborative filtering algorithms recommend author's books of a given gender with a distribution that differs from that of the original user profiles. Liu and Burke [120] consider P-fairness in the Kiva.org platform, which grants loans to low-income entrepreneurs. It is achieved through a re-ranking function (based on xQuad), which balances recommendation accuracy and fairness, by dynamically adding a bonus to the items of the uncovered providers. In the same domain, Sonboli and Burke [168] define the concept of local fairness, to identify protected groups through consideration of local conditions. This is done to avoid discriminating between types of loans and to equalize access to capital across all types of businesses. Abdollahpouri et al. [2] analyze the unfairness of popularity bias in movies recommendation, while Kowald et al. [109] analyze the same problem in the music domain.

Mehrotra et al. [130] assess unfairness based on the popularity of the providers. More specifically, they focus on a two-sided marketplace, with the consumers being users who listen to music, and the artists being the providers. If only highly popular artists are recommended to users, this creates a disadvantage for the less popular ones. For this reason, artists are divided into ten bins based on their popularity, and a fairness metric that rewards recommendation lists that are diverse in terms of popularity bins is defined. Several policies are defined to study the trade-offs between user relevance and fairness, with the ones that balance the two aspects being those who achieve the best trade-off.

Several policies are defined to study the trade-offs between user relevance and fairness. Kamishima et al. [102] introduce the concept of recommendation independence. Given a sensitive feature (which can be associated with the consumers, the providers, or the items), they present a framework to generate fair recommendations, in the sense that the outcome is statistically independent of a specified sensitive feature. Specifically, an objective function with three components (a loss function, an independence term, and a regularization term) is introduced, so that the prediction function returns an expected value of the loss function as small as possible and an independent term as large as possible.

4.2.3 Contextualizing our Work

To the best of our knowledge, this is the first time that unfairness phenomena for content providers belonging to different continents are tackled. Considering the UNESCO[3], our two study domains (i.e, cinema and literature) are powerful vehicles for culture, education, leisure, and propaganda. This report also highlights the importance of smaller cinematographic industries at the global level. In our movie dataset, India represents 0.004% of the total amount of items.

Moreover, both domains have an impact on the economy of a country, with (sometimes public) investments for the production of movies/books that are expected to generate a return. Hence, considering how Recommender Systems can push the consumption of items of a country is a related but different problem w.r.t. provider fairness.

In conclusion, studying the disparities emerging from the geographic imbalances in the composition of an industry is a problem that goes beyond the impact for content providers.

[3] https://publications.parliament.uk/pa/cm200203/cmselect/cmcumeds/667/667.pdf

Denying visibility and exposure to the items of a continent has a negative impact (i) on the cultural impact that a country can have and (ii) at an economic level.

4.3 Preliminaries

In this section we present the preliminaries, to provide foundations to our work. First, Section 4.3.1 details the recommendation scenario. Next, the metrics are described in Section 4.3.2. In Section 4.3.3, we present the recommendation algorithms. Finally, we describe the datasets used in this study in Section 4.3.4.

4.3.1 Recommendation Scenario

We consider a set of users, $U = \{u_1, u_2, ..., u_n\}$, a set of items, $I = \{i_1, i_2, ..., i_j\}$, and let V be a totally ordered set of values that can be used to express a preference together with a special symbol $\perp$. The set of ratings result from a map $r : U \times I \rightarrow V$, where V is the ratings' domain. If $r(u, i) = \perp$ then we say that user, u, did not rate item, i. To simplify notation, we denote $r(u, i)$ by r_{ui}. We define the set of ratings as $R = \{(u, i, r_{ui}) \, : \, u \in U, \, i \in I, \, r_{ui} \neq \perp\}$. These ratings can directly feed an algorithm in the form of triplets (point-wise approaches) or shape learner-course observations (pair-wise approaches).

To assess the real impact of the recommendations, we consider a temporal split of the data, where a fixed percentage of the ratings of the learners (ordered by timestamp) goes to the training and the rest goes to the test set [12].

The recommendation goal is to learn a function f that estimates the relevance ($\hat{r}_{ui}$) of the user-item pairs that do not appear in the training data (i.e., $r_{ui} = \perp$). We denote as $\hat{R}$ the set of recommendations, and as $\hat{R}_G$ those involving items of a group G, i.e., $\hat{R}_G = \{\hat{r}_{ui} \, : \, u \in U, \, i \in G \subseteq I\}$.

Let $A = \{a_1, a_2, ..., a_g\}$ denote the set of g geographic areas in which items are organized. Specifically, we consider a geographic area as the continent of provenience of each item provider. We denote as A_i the set of geographic areas associated with an item i. Note that, since the providers of an item could be from different geographical areas, several geographic areas may appear in an item, and thus, $|A_i| \geq 1$. In case two providers belong to the same geographic area, it appears only once. We use the geographic areas to shape k demographic groups, where the t^{th} demographic group is defined as $G_t = \{i \in I \, : \, a_t \in A_i\}$, for $t = 1, \ldots, g$. Finally, Table 4.1 summarizes the terminology used in this article.

4.3.2 Metrics

This section describes the metrics used in our analysis and experiments, i.e., the representation of a group, disparate visibility, and disparate exposure.

Table 4.1 Summary of the terminology used in the article. First column details the concept, while the second presents the notation for this concept.

Concept	Term
Set of users	U
Set of items	I
Set of preferences	V
Set of ratings	R
Rating of user u over item i	r_{ui}
Predicted relevance of item i for user u	$\hat{r}_{ui}$
Set of recommendations	$\hat{R}$
Set of recommendations involving items of a group G	$\hat{R}_G$
Set of geographic areas	A
Set of geographic areas associated with a item i	A_i
Demographic group	G_t
Item-based representation of a group	$\mathcal{R}_I(G)$
Rating-based representation of a group	$\mathcal{R}_R(G)$
Disparate visibility of a group	$\Delta\mathcal{V}(G)$
Disparate exposure of a group	$\Delta\mathcal{E}(G)$

Representation. The representation of a group is the amount of times that group appears in the data. We consider two forms of representation, based on (i) the amount of items offered by a group and (ii) the amount of ratings collected for that group. We define with $\mathcal{R}$ the *representation* of a group G ($\mathcal{R}_I$ denotes an item-based representation, while $\mathcal{R}_R$ a rating-based representation):

$$\mathcal{R}_I(G) = |G| \big/ |I| \tag{4.1}$$

$$\mathcal{R}_R(G) = |\{r_{ui} : u \in U, i \in G \subseteq I\}| \big/ |R| \tag{4.2}$$

Eq. (4.1) accounts for the proportion of items of a group, while Eq. (4.2) for the proportion of ratings associated with a group. Both metrics are between 0 and 1. We compute the representation of a group only considering the training set. Trivially, given a perspective (either item- or rating-based), the sum of the representations of all groups is equal to 1, $\sum_{i=1}^{k} \mathcal{R}_*(G_i) = 1$ (where '*' refers to I or R).

Disparate Impact. We assess unfairness with two notions of *disparate impact* generated by a recommender system. Specifically, we assess disparate impact with two metrics.

Definition 4.3.5 (Disparate visibility). *Given a group G, the* disparate visibility *returned by a recommender system for that group is measured as the difference between the share of recommendations for items of that group and the representation of that group in the input data:*

$$\Delta\mathcal{V}(G) = \left(\frac{1}{|U|}\sum_{u\in U}\frac{|\{\hat{r}_{ui} : \hat{r}_{ui} \in \hat{R}_G,\, i \in G \subseteq I\}|}{|\hat{R}|}\right) - \mathcal{R}_*(G) \tag{4.3}$$

where ‘*’ refers to I (i.e., item-based representation) or R (i.e., rating-based representation). The range of values for this score is $[-\mathcal{R}_*(G), 1 - \mathcal{R}_*(G)]$; specifically, it is 0 when the recommender system has no disparate visibility, while negative/positive values indicate that the group received a share of recommendations that is lower/higher than its representation. This metric is based on that considered by Fabbri et al. in [59].

Definition 4.3.6 (Disparate exposure)**.** *Given a group G, the* disparate exposure *returned by a recommender system for that group is measured as the difference between the exposure given to that group in the recommendation lists [167] and its representation:*

$$\Delta\mathcal{E}(G) = \left(\frac{1}{|U|}\sum_{u\in U}\frac{\sum_{pos=1}^{k}\frac{1}{log_2\left(\hat{r}^u_G(pos)+1\right)}}{\sum_{pos=1}^{k}\frac{1}{log_2\left(\hat{r}^u_I(pos)+1\right)}}\right) - \mathcal{R}_*(G), \tag{4.4}$$

where $\hat{r}^u_G(pos)$ denotes the rating $\hat{r}_{ui}$ that takes position pos in the list $\hat{R}^u_G = \{\hat{r}_{vi} \,:\, v = u,\, i \in G \subseteq I\}$, $u \in U$, sorted by decreasing order.

This metric also ranges in $[-\mathcal{R}_*(G), 1 - \mathcal{R}_*(G)]$ range; concretely, a value equal to 0 indicates that the recommender system has no disparate exposure, while negative/positive values indicate that the exposure given to the group is lower/higher than its representation in the data.

Remark. *Since the goal of our paper is to* allow the items of a group to be recommended enough times (visibility) and with enough exposure*, a unique “disparate impact” metric would not allow us to balance both perspectives, by combining everything in a single number. For this reason, we keep both disparate visibility as disparate exposure as goals to enable provider fairness in the context of geographic imbalance.*

4.3.3 Recommendation Algorithms

In this study, we focus on well-known state-of-the-art Collaborative Filtering approaches. In particular, we focus on both classes of point-wise and pair-wise approaches, and considered memory-based and model-based algorithms.

As *memory-based approaches*, we consider two approaches: **UserKNN** and **ItemKNN** algorithms. UserKNN [92] selects the K neighbors closest to the target user, and recommends the elements like by other users more similar to the target one. Similarly, ItemKNN [159] recommends to the target user those items that are more similar to other items that they liked before.

For the class of *matrix factorization based approaches*, we consider the **BPR**, **BiasedMF**, and **SVD++** algorithms. Matrix factorization algorithms divide the data into matrices, representing them in latent factors to determine the degree of affinity that users and items have with those factors. In particular, Bayesian Personalized Ranking (in short, BPR) [149] is an algorithm optimized to generate recommendation lists, based on a Bayesian probability function. The preference function is based on the ratings of pairs of items. BiasedMF [108] performs basic factorization of the matrix that includes global mean, user bias, and item bias whereas SVD++ [107] takes into account the implicit interactions, as well as the user's latent factors and the item's latent factors.

Our baselines are two *non-personalized algorithms* (**MostPopular** and **RandomGuess**), which allow us to contextualize our results. MostPopular recommends items based on how their popularity in the dataset, by counting the number of times an item was rated. In this way, the algorithm considers only the item perspective, without associating the ratings to the individual users and their preferences. On the other hand, RandomGuess establishes the maximum and minimum rating in the data and returns a random rating for each user-item pair to predict.

4.3.4 Datasets

We analyze data from two different contexts, movies and books, exploring the role of the geographic provenience of providers in the recommendation process. In what follows, we describe the characteristics of each dataset:

- **MovieLens-1M (Movies).** The dataset provides 1M ratings (in the range [1-5]), given by 6,040 users to 3,600 movies. Each user rated at least 20 ratings. For each user, the dataset provides demographic information (namely, their gender, age, occupation, and zip code), which have not been considered for this study to focus on provider fairness; these attributes will be considered in future work, to study the interplay between provider fairness and the characteristics of the users. For each movie, the dataset provides its IMDb ID, which allowed us to associate it to its continent of production, thanks to the OMDB APIs[4] (note that *each movie may have more than one continent of production*). The dataset also offers the title and genre of each movie, which are not relevant in the context of this study and, thus, have not been considered.
- **Book-Crossing (Books).** The dataset contains 356k ratings (in the range [1-10]), given by 10,409 users, to 14,137 books. Also this dataset provides demographic information about its users, by offering age and location attributes, if the user has provided them. Also in this case, these attributes are not relevant for our study, so any additional information of the users offered by this study will be considered in future work. For each book, the dataset provides its ISBN code; this code allowed us to retrieve information about the production continent, by exploiting the APIs offered by the Global Register of Publishers[5]. Additional book information such as its title, author, publisher, and cover image, is not relevant for our study and, thus, has not been considered.

[4] http://www.omdbapi.com/

[5] https://grp.isbn-international.org/search/piid_cineca_solr

We shape demographic groups considering the following continents: Africa, Asia, Europe, North America, Oceania, and South America for Movies, and Europe, North America, Oceania, and South America for Books. We remark that no items from Africa or Asia were available in the Books dataset and there are no items from the seventh continent (Antarctica) in both datasets.

4.4 Provider Fairness with a Binary Perspective

This section frames our previous study, which deals with provider fairness for two groups in presence of geographic imbalance. It summarizes the main observations obtained from our experiments to enhance the need for a cross-continent provider perspective.

In our previous work [78], we observed that imbalances in the data distribution can affect the visibility and exposure given to providers. Our study was focused on analyzing the items into two different groups based on the country of production, in a majority-versus-rest setting, and assessed if Recommender Systems generate a disparate impact and (dis)advantage a group. Recall that, in this work, our setting is focused on cross-continent provider fairness. Another observation extracted from our study is that the produced recommendations by the Recommender Systems can amplify these imbalances and create biases. To study this phenomenon, we enriched two datasets and characterize data imbalance w.r.t. the country of production of an item (geographic imbalance). We conducted the experiments in two domains, movies (MovieLens-1M) and books (Book-Crossing), where both datasets are imbalanced towards the United States.

4.4.1 Group Representation

We assessed disparate impact by comparing the visibility and exposure given to a group of providers with the representation of the group in the data. As we are doing in this study, we studied two forms of representation, based on (i) the number of items a group offers, or (ii) the number of ratings given to the items of a group.

Let C_i be the set of production countries of an item i. We use it to shape two groups, a majority $G_M = \{i \in I : 1 \in C_i\}$, and a minority $G_m = \{i \in I : 1 \notin C_i\}$. Note that 1 identifies the country associated with the majority group.

In the binary perspective, the representation of the minority group in the Movies dataset is $\mathcal{R}_I(m) = 0.3$ and $\mathcal{R}_R(m) = 0.23$, considering item and rating, respectively. In the Books dataset, instead, the representation of the minority group is $\mathcal{R}_I(m) = 0.12$ and $\mathcal{R}_R(m) = 0.08$. As it can be observed, both datasets show a strong geographic imbalance, with the majority group covering 70% of the items in the first dataset and 88% in the second. This imbalance is worsened when we consider the ratings, since in the movie context the ratings associated with the majority are 77%, while in the book data the rating representation for the majority is 92%. However, the minority items are not considered as of lower quality for the users, since the average rating for both groups is nearly the same in both datasets. In the Movies dataset, the average rating for the majority group is 3.56, while that of the minority group is 3.61. In the Books dataset, we observed an average rating of 4.38 for the majority, and of 4.43 for the minority. This shows that the preference of the users for the two groups does not differ.

4.4.2 Metrics and Algorithms

We characterized both the visibility and exposure given to the providers of a group by a recommendation algorithm. To evaluate recommendation effectiveness, we measured the ranking quality of the lists by measuring the *Normalized Discounted Cumulative Gain* (NDCG). We ran the state-of-the-art Recommender Systems described in Section 4.3.3, using the LibRec library. The test set was composed of the most recent 20% of the ratings of each user.

4.4.3 Assessment

In our initial analysis of both datasets (i.e, Movies and Books), the phenomenon that emerges is that both groups can be affected by disparate impact and that, when one group receives more visibility, it also receives more exposure; hence, when a group is favored in the number of recommendations, it is also ranked higher.

Concretely, the results showed the presence of a disparate impact that mostly favors the majority, since we feed algorithms with much more instances than their counterpart. However, factorization approaches are still capable of capturing the preferences for the minority items with latent factors, thus creating a positive impact for the group. But, if the imbalance is too severe, the minority is always affected by the disparate impact.

4.4.4 Approach

To mitigate disparities, we proposed a binary re-ranking approach that optimizes both the visibility and exposure given to providers in a binary (i.e., majority-vs-rest) setting, based on their representation in the data. Specifically, our approach introduces, in the recommendations, items that increase the visibility and exposure for the disadvantaged group, causing the minimum possible loss in user relevance. For each user, we generated 150 recommendations (denoted as the top-n) so that we can mitigate the disparate impact through a re-ranking algorithm. The final recommendation list for each user is composed of 20 items (denoted as top-k) and measured the visibility and exposure given to each group.

In what follows, we provide a summary of the steps followed by our re-ranking approach:

1. We identify the user causing the minimal loss in terms of items' predicted relevance;
2. We select two items in the list of the user, namely the last item of the advantaged group in the top-k and the first item of a disadvantaged group out of the top-k;

3. We swap the items and move to step 1 until the target visibility is reached.

After the target visibility is reached, we consider the top-k to regulate the exposure of the disadvantaged group. We swap inside the list the pair of items belonging to different groups that cause the minimum loss of predicted relevance, until the desired exposure for the disadvantaged group is reached.

4.4.5 Impact of Mitigation

Briefly, the impact of our proposed re-ranking algorithm for mitigating disparities in the binary setting is three-fold. First, our approach leads to the goal visibility and exposure. Given a target representation and a dataset, all algorithms achieve the same disparate visibility/exposure. Second, thanks to our mitigation based on the minimum-loss principle, the loss in NDCG was negligible. Finally, the most effective approach before mitigation is confirmed as such also after mitigation.

4.5 Disparate Impact Assessment

The first goal of this study is to evaluate the presence of unfairness in the state-of-the-art collaborative recommendation models, so as to understand if and where a problem exists. Concretely, our task is to analyze the recommendations these models generate and assess the presence of disparities in the way recommendations are distributed across different provider groups.

To accomplish this goal, in this section, we run the algorithms presented in Section 4.3.3 and measure their effectiveness and the disparate impact they generate for providers belonging to different continents.

At the end of this section, we will be able to understand which models create disparities and under which conditions.

4.5.1 Experimental Setting

For both datasets presented in Section 4.3.4, the test set was composed of the most recent 20$\%$ of the ratings of each user. To run the recommendation algorithms presented in Section 4.3.3, we considered the LibRec library (version 2). For each user, we generate 150 recommendations (denoted in the paper as the top-n) to then mitigate disparities through a re-ranking algorithm. The final recommendation list for each user is composed of 20 items (denoted as the top-k).

Each algorithm was run with the following hyper-parameters:

- **UserKNN.** We used Pearson similarity and 50 neighbors. The similarity shrinkage was set up to 10;

- **ItemKNN.** We used the Cosine similarity for the Movies dataset and Pearson similarity for the Books one. The number of neighbors was 200 for Movies and 50 for Books dataset. The similarity shrinkage was set up to 10;
- **BPR.** We configured the iterator learnrate to 0.1, the iterator learnrate maximum to 0.01, the iterator maximum to 150, the user regularization to 0.01; the item regularization to 0.01; the factor number to 10; the learnrate bolddriver to false, and the learnrate decay to 1.0;
- **BiasedMF.** We adjusted the iterator learnrate to 0.01, the iterator learnrate maximum to 0.01, the iterator maximum to 20 for the Movies dataset and 1 for the Books one, the user regularization to 0.01, the item regularization to 0.01; the bias regularization to 0.01, the number of factors to 10, the learnrate bolddriver to false, and the learnrate decay to 1.0;
- **SVD++.** We set up the iterator learnrate to 0.01, the iterator learnrate maximum to 0.01, the iterator maximum to 10 for the Movies dataset and 1 for the Books one, the user regularization to 0.01, the item regularization to 0.01, the impItem regularization to 0.001, the number of factors to 10, the learnrate bolddriver to false, and the learnrate decay to 1.0.

Recommendation effectiveness is assessed by measuring the ranking quality of the list, using the *Normalized Discounted Cumulative Gain* (NDCG) [99].

$$DCG@k = \sum_{u \in U} \hat{r}_{ui}^{pos} + \sum_{pos=2}^{k} \frac{\hat{r}_{ui}^{pos}}{log_2(pos)} \quad NDCG@k = \frac{DCG@k}{IDCG@k} \tag{4.5}$$

where $\hat{r}_{ui}^{pos}$ is relevance of item i recommended to user u at position pos. The ideal DCG ($IDCG$) is calculated by sorting items based on decreasing true relevance (true relevance is 1 if the user interacted with the item in the test set, 0 otherwise). The higher the better.

4.5.2 Characterizing Representation

The first step towards the assessment of disparate impact is to characterize the representation of the different groups in the data, which we present in Table 4.2. Note that the Books dataset does not contain books from Africa and Asia.

We can observe that North America represents the most represented continent in both datasets, covering 69% of the produced items ($\mathcal{R}_I$) in the Movies data and almost 90% of the items in Books. This existing imbalance is increased if we consider the rating-based representation ($\mathcal{R}_R$), where North America has a share of 76.6% and 93% of the ratings, respectively in the Movies and Books data. This leads us to our first observation.

Table 4.2 Group representation. item-based ($\mathcal{R}_I$) and rating-based ($\mathcal{R}_R$) representations of each group (AF: Africa, AS: Asia, EU: Europe, NA: North America, OC: Oceania, SA: South America). Groups appear in alphabetical order by the name of the continent.

	MOVIES		BOOKS	
	$\mathcal{R}_I$	$\mathcal{R}_R$	$\mathcal{R}_I$	$\mathcal{R}_R$
AF	0.0038	0.0028	-	-
AS	0.0392	0.0234	-	-
EU	0.2469	0.1946	0.1043	0.0698
NA	0.6937	0.7659	0.8951	0.9299
OC	0.0139	0.0128	0.0005	0.0002
SA	0.0025	0.0003	0.0001	0.0001

Observation 1. *Both datasets have a strong geographic imbalance towards North America, which is the most represented group from both item- and rating-based perspectives. The imbalance is strengthened when considered the rating-based representation, meaning that the largest groups attract a share of ratings that is even higher than the amount of items it offers. This clearly has a price for the smaller groups, which are able to attract a percentage of ratings that is lower than the amount of items they offer. Hence,* user-item interactions favor the largest group and exacerbate imbalances that already existed in the item offer, even before we run a recommendation algorithm.

4.5.3 Assessing Effectiveness and Disparate Impact

In this section, we assess the effectiveness (in terms of NDCG) and the disparate impact (both in terms of visibility and exposure) returned by the state-of-the-art algorithms. Moreover, we assess if the binary mitigation (for two groups) proposed in [78] is capable of enabling fairness for multiple groups shaped at the continent level.

Results are visually reported in Fig. 4.2 for the Movies dataset and in Fig. 4.3 for the Books one. To present our results in a reproducible way, Tables A.1 and A.2 (placed in the Appendix) report the values that shape our figures, for Movies and Books respectively. If we look at how the original models (thick bars) behave according to the different types of representation, we can observe that they adjust better to the rating-based representation of the groups. Indeed, in both figures we can observe that in (c) and (d) the disparities returned when considering an item-based representation are more prominent than their rating-based counterparts, in (a) and (b). In other words, recommendation models adapt better to the interactions between users and items than to the amount of items a group has to offer. The only exception to this is RandomGuess which, by picking items at random, better adapts to the distribution of the items. Indeed, as subfigures (c) and (d) show, it is the approach that returns the most equitable results, in both Fig.s 4.2 and 4.3. Nevertheless, as it is shown by the NDCG values at the bottom, it is also the least effective approach. One can notice that the thin bars, reporting the results after the binary mitigation, are generally closer to 0 than the original models, indicating that, even though we are not providing fairness at the continent level, still the approach in [78] distributes the recommendation in a more equitable way than the original models.

Going more in-depth to the behavior of the system in each domain, Fig. 4.2, shows that for Movies the algorithm that better adjusts to the different continents, when considering the rating-based representation is BPR; indeed, the algorithm is the most effective one, returning the highest NDCG, and the one that adjusts better to the desired equity in terms of disparate visibility and exposure. One interesting phenomenon we can observe is that, when a model over- or under-recommends one of the two most represented groups (i.e., North America and Europe), the other is directly affected. We can see this clear pattern for MostPop, RandomGuess, UserKNN, and ItemKNN, who create disparate visibility and exposure at the advantage of North America, while affecting the most the second most represented country (Europe). On the contrary, when Europe is over-recommended, North America is the most affected country, as shown by BiasedMF and SVD++. This last phenomenon we observed for the point-wise recommendation models (BiasedMF and SVD+) connects to the observations of Cremonesi et al. [38], who showed the capability of factorization approaches to recommend long-tail items. Disparate impact clearly affects less the smallest groups, since they have a representation almost equal to 0, which is reflected in the visibility and exposure they are given; nevertheless, when a country is under- or over-recommended, with disparate visibility and exposure values lower/higher than 0, (e.g., Oceania and Africa), all models follow the same pattern.

Moving to the behavior of the binary mitigation proposed in [78], clearly, since our original mitigation was based on providing the United States with an equitable number of recommendations, North America is the continent that benefits the most from our original mitigation, especially with BiasedMF and SVD++.

Moving to our Books data, in Fig. 4.3 we can observe that BiasedMF is the most effective approach. One interesting aspect we can notice here is that not even the point-wise matrix factorization based models are able to contrast the imbalance, favoring North America in terms of visibility and exposure. This leads us to our second observation.

> **Observation 2**. *Recommendation models better adjust to the rating distribution than to the item offer associated with a group. Factorization-based approaches are able to account for the needs of smaller groups, unless the imbalance in the input data is too severe. Even though we are working in a multi-group setting, Recommender Systems mostly operate as if two big groups existed; when one group is favored, the other is affected, both in terms of visibility and exposure. A mitigation for binary groups helps reducing disparities, but is not enough to introduce fairness for groups shaped at the continent level. Indeed, integrating more recommendations of the items from the minority group does not ensure that these recommendations are equally distributed among the different continents, so disparities still emerge.*

4.6 Mitigating Disparate Impact

In the previous section, we assessed the presence of disparities, having a negative impact mainly for the less represented provider groups. To overcome this limitation and introduce provider fairness for the different geographic groups, in this section, we detail the motivation of our approach and describe the re-ranking algorithm proposed to mitigate disparities at the continent level.

4.6.1 Motivation Behind our Approach

From the previous section, considering the representation of each group in the data, we noticed that some groups receive disproportional visibility and exposure. As a result of this observation, we propose to mitigate disparities with a re-ranking algorithm. Specifically, the goal of the proposed algorithm is to reach a visibility and exposure for each group proportional to their representation by moving items of the disadvantaged groups in the recommendation list.

A re-ranking algorithm is the unique option when optimizing metrics such as visibility and exposure. We cannot perform an in-processing regularization (e.g., [102, 14]) because at the prediction stage it is not possible to know if and where an item is ranked in a recommendation list. For this reason, it is not possible to do a comparison with this class of approaches. The reason why this comparison is not possible is not due to the algorithms we chose in our study, as this consideration also applies to list-wise approaches. Note that re-ranking algorithms have been introduced in the context of recommendation [130, 27, 30] as well as in non-personalized rankings [192, 167, 15, 31, 191, 139], but all of them are optimizing just one metric (i.e, visibility or exposure). Hence, in this section, we present an approach that provides fairness guarantees to all the provider groups in the data, considering both visibility and exposure metrics.

4.6.2 Algorithm

Our mitigation algorithm is based on the idea of *promoting in the recommendation list the item, that considering all the users, minimizes the loss in prediction*. Algorithm 3 describes pipeline followed by our mitigation method and Algorithm 4 presents our regulation of visibility and exposure in the recommendation lists. Finally, Algorithm 5 presents the support methods that are called by our mitigation method. Algorithm 3 takes as inputs (i) the recommendation list, $recList$, for all the users (consisting of the top-n items) and (ii) how recommendations should be distributed across continents after the mitigation ($targetProportions$). The output is the new list of re-ranked items, $reRankedList$.

Algorithm 3 consists of one main method, called $optimizeContinentsVisibilityExposure$ (lines 1-6). It makes two interventions, one based on visibility and the second one based on exposure. After each method is called, it returns the recommendation list, optimized for visibility (line 3) and exposure (line 4).

```
Input: $recList$: ranked list (records contain $user$, $item$, $prediction$, $exposure$, $continent$, $position$)
       $targetProportions$: list with the target proportions of each continent
Output: $reRankedList$: ranked list adjusted by visibility and exposure
define optimizeContinentsVisibilityExposure ($recList$, $targetProportions$)
begin
    // mitigation to target the desired visibility
    $reRankedList \leftarrow$ mitigationContinent($recList$, “visibility”, $targetProportions$);
    // mitigation to regulate the exposure
    $reRankedList \leftarrow$ mitigationContinent($reRankedList$, “exposure”, $targetProportions$);
    return $reRankedList$ ; // re-ranked list adjusted by visibility and
     exposure
end
```

Algorithm 3: Muticlass mitigation algorithm based on Visibility and Exposure

Algorithm 4 contains the method that performs our mitigation process, called $mitigation Continent$ (lines 1-34). Concretely, the method regulates, in a recommendation list, the visibility or exposure, so that it reaches the representation of each continent.

After setting some supporting data structures (line 3) and assessing the current disparity we observe for each continent (lines 4 and 5), in lines 6-20, we create two lists of candidate items, respectively to be removed from and added in the recommendation list, named $itemsOut$ and $itemsIn$. Concretely, the first list contains items currently recommended to the user that belong to an advantaged group, while the second contains items of a disadvantaged group currently not recommended to the user. In lines 14-19, we create a list, named $possibleSwaps$, containing pairs of candidate items that cause the minimum possible loss in terms of predicted relevance for the users. This list is sorted by loss in line 21. Finally, in lines 23-32, we swap the items and update the proportions, until we reach the desired visibility or exposure in the recommendation list. The re-ranked list is returned in line 33.

Finally, Algorithm 5, contains the methods we call in Algorithm 4. Concretely, the $check Position$ method (lines 1-5) is responsible for checking the position of an item in the list, taking into account if we perform a visibility- or exposure-based mitigation. The $checkDisadvantaged Group$ method (lines 6-10) verifies whether the item belongs to a disadvantaged continent or not. Note that the method contains a for loop, since multiple continents may occur in an item. In that case, we compute the total sum of disparities to define a global disparity of the item. The method returns true when the disparity is positive, false otherwise. The $initialProportions$ method (lines 11-24), returns the visibility and exposure of each continent before running our mitigation. The last method, $updatePositions$ (lines 25-33) is responsible for updating the visibility and exposure given to a group after an item is added to the recommendation list.

4.7 Impact of Mitigation

The final goal of our study is to assess if the approach we presented in Section 4.6 is capable of providing fairness, by distributing the recommendations in equitable ways between the different provider groups. Moreover, we want to analyze if our approach can accomplish this goal in a better way than the existing approaches at the state of the art. Concretely, our task is to analyze the recommendations generated after running our mitigation strategy and a well-known state-of-the-art algorithm, to assess if disparities are still present and where and how recommendation effectiveness is impacted by our mitigation.

To accomplish this goal, in this section we analyze the impact of our mitigation approach, by assessing recommendation effectiveness and the presence of disparate impact for providers belonging to different continents. Section 4.7.1 shows the results of our mitigation algorithm and the advantages of employing an approach that can account for the presence of multiple groups, rather than the binary-group perspective proposed in [78]. Next, in Section 4.7.2, we compare our proposal against a well-known re-ranking approach, proposed in [120].

4.7.1 Impact of Mitigating for Multiple Groups

In this section, we analyze the impact of our mitigation algorithm for multiple groups, analyzing both the recommendation effectiveness and the visibility and exposure given to the different groups.

We report our results in Fig. 4.4 for the Movies dataset, and in Fig. 4.5 for the Books one. To present our results in a reproducible way, Tables A.3 and A.4 (see Appendix) report the values that shape our figures, for Movies and Books respectively.

One aspect that can be appreciated is that, given a reference representation and a dataset, all algorithms disparities are almost equal to 0, indicating we can provide a fair distribution of the recommendations, based on the distribution of the continents in the input data. This can be noticed by observing the thin bars in each subfigure.

Let us consider the trade-off between disparate visibility/exposure and effectiveness. Considering the Movies data (Fig. 4.4), in both the rating- (subfigures a and b) and item-based (subfigures c and d) representations of the groups, BPR is the algorithm with the best trade-off between effectiveness and equity of visibility and exposure. It was already the most accurate algorithm, and thanks to our mitigation based on the minimum-loss principle, the loss in NDCG was negligible. Moving to the Books dataset (Fig. 4.5), BiasedMF confirms to be the best approach, in both effectiveness and equity of visibility and exposure. It is interesting to observe that, in both scenarios, MostPop is the second most effective algorithm and now provides the same visibility and exposure as the other algorithms; we conjecture that this might be due to popularity bias phenomena [21], and their analysis is left as future work.

> **Observation 3**. *When providing a re-ranking based on minimal predicted loss, the effectiveness remains stable, but disparate visibility and disparate exposure are mitigated. The most effective approach remains the best one after the mitigation.*

In Section 4.5.3, we have shown that a mitigation considering only two groups is not enough to introduce equity in the presence of multiple groups. To assess the difference the benefits of considering multiple groups in a mitigation strategy, in Tables 4.3 and 4.4, we compare the results we can obtain with our multi-group mitigation and the one considering only a binary perspective, in the Movies and Books datasets, respectively. The results clearly show that a mitigation at a more fine-grained granularity can provide fairness to providers in different groups.

4.7.2 Comparison with the State of the Art

We compare the results of our mitigation with that proposed in [120]. This approach aims at introducing provider fairness via a re-ranking approach, as our approach. Differently from us, in the mitigation proposed in [120] the predicted relevance is increased if a provider has not appeared yet in the top-k of a user. Since we are dealing with a provider fairness setting, we increase the predicted rating if a geographic area has not appeared yet in the ranking of a user. We remind readers to [120] for the technical details of the re-ranking approach we compare with. Hyperparameter λ of the original algorithm proposed in [120] was set to 2.

Tables 4.5 and 4.6 report the obtained results, where *multi* refers to our re-ranking multi-group mitigation algorithm and *baseline* is the compared algorithm.

We observe that our approach in most cases is capable of introducing equity by mitigating both disparate visibility and exposure in all algorithms. In general, our algorithm achieves better disparities than the baseline (indeed, in our results disparities are almost always close to 0). The baseline algorithm is able to minimize the disparities for those groups that are more represented but not for the less represented ones. Our proposal reduces slightly disparities with respect to the baseline in Most Popular, UserKnn, and ItemKnn only in South America (SA), which is the group with the smallest representation. In the remaining continents and algorithms, our proposal is highly effective in mitigating the disparities. We consider that the baseline is not mitigating both visibility and exposure to a greater extent because it favors the introduction, in the top-k, of items produced in more than one geographic group. This means that, while a disadvantaged group might gain visibility and/or exposure, the accompanying group also receives the same treatment, even though it might be advantaged.

Observation 4. *Introducing provider fairness requires interventions at the recommendation-list level. Mitigating by boosting predicted relevance for the disadvantaged groups does not provide guarantees of equity of visibility and exposure are fully mitigated. Disparities are only partially mitigated.*

4.8 Conclusions and Future Work

Recommender Systems usually emphasize biases that emerge because of the way data has been collected. In this work, we focused on a scenario in which imbalances are associated with the way an industry is composed, with certain geographic areas that produce more items of certain types. This is the case for movies and books, which have been the main use-cases in our work.

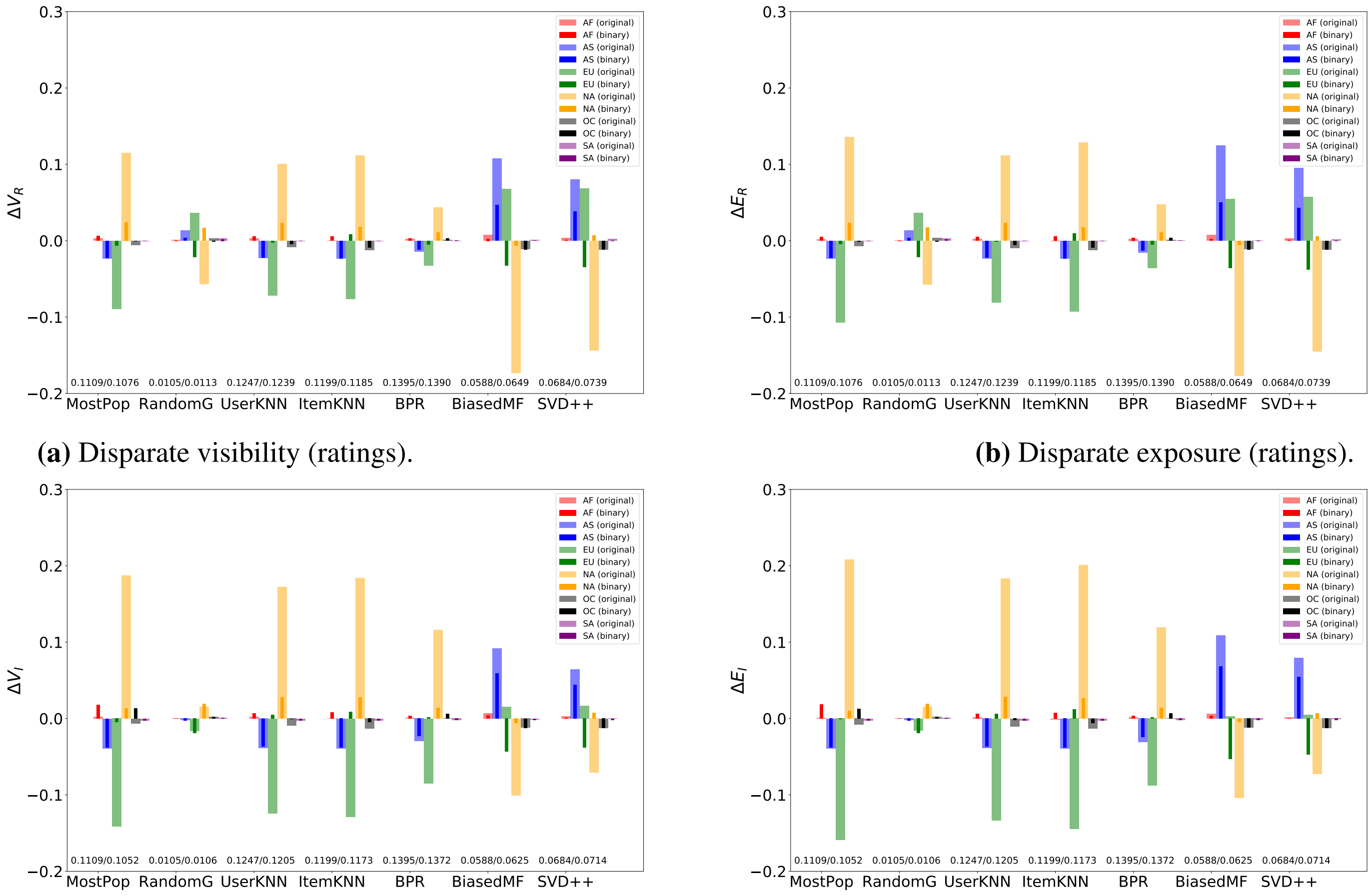
Concretely, we assessed how Recommender Systems dealt with data imbalances, studying their capability to recommend items of providers coming from different continents and possible unfairness phenomena emerging from the way recommendations are distributed. We considered state-of-the-art collaborative filtering models and assessed that all of them create disparities in the way recommendations are produced, both in terms of visibility and of exposure given to providers.

To overcome these phenomena, we analyzed a binary re-ranking approach [78], which improves geographic imbalance in a binary (i.e., majority-vs-rest) setting by maintaining recommendation effectiveness. However, we have observed that in a group setting the approach does not reach equity for all groups. Accordingly to this observation, we proposed a multi-group re-ranking approach that re-distributes the recommendation across provider groups (i.e., geographic continents) based on a notion of equity, that assigns to each group a share of recommendation proportional to its representation in the input data. Experimental results show that our approach can introduce provider fairness without affecting recommendation effectiveness.

Considering that in this study we observed that the mitigation of data imbalances needs intervention at a fine-grained level, in future work we will assess the interplay between the representation of individual providers and the geographic area they belong to. Concretely, we will consider settings with more fine-grained groups (e.g., at country level), to assess if, with more groups, each with a lower representation in the data, our approach can still enable fairness for provider groups, and possibly refine our approach. Moreover, we will consider additional domains such as education, to explore disparities for teachers [9, 44, 42, 43]. Finally, we will also consider other issues emerging from imbalanced groups, such as bribing [161, 148].

4.9 Appendix

In Appendix A, we show Tables that will help the reader to reproduce the results obtained in our experiments.



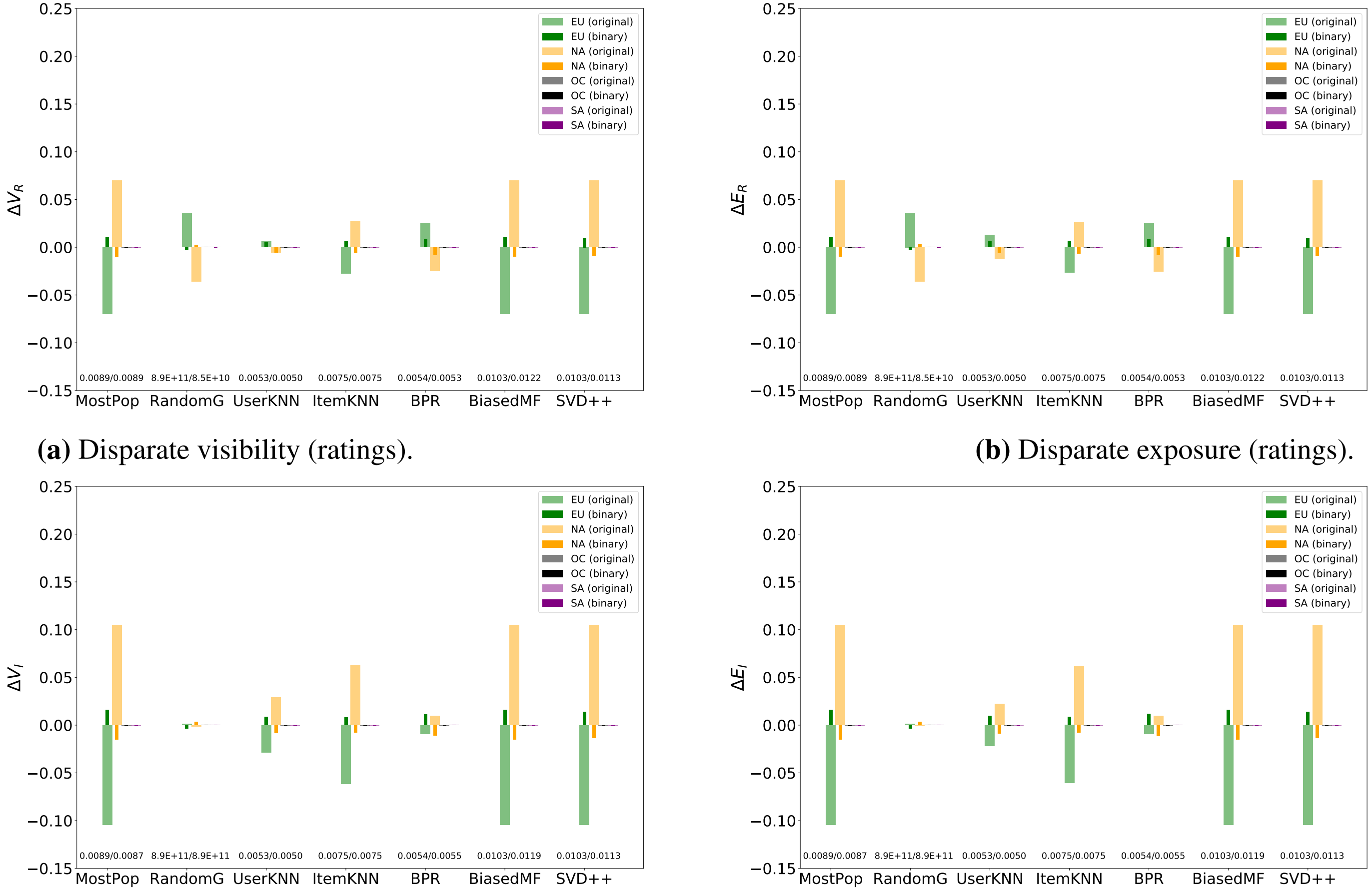
(a) Disparate visibility (ratings).

(b) Disparate exposure (ratings).

(c) Disparate visibility (items).

(d) Disparate exposure (items).

Fig. 4.2 Disparate impact in the Movies dataset. Disparate impact returned by the state-of-the-art models (thick bars) and by the binary mitigation proposed in [78] (thin bars). Each figure contains one section for each algorithm and one color for each continent. The text at the bottom of each figure contains the NDCG returned by the original model and after the binary mitigation, separated by a "/". In (a) and (b), we report the disparate visibility and disparate exposure obtained when considering a rating-based representation, while in (c) and (d), the disparate visibility and disparate exposure obtained when considering an item-based representation.



(a) Disparate visibility (ratings).

(b) Disparate exposure (ratings).

(c) Disparate visibility (items).

(d) Disparate exposure (items).

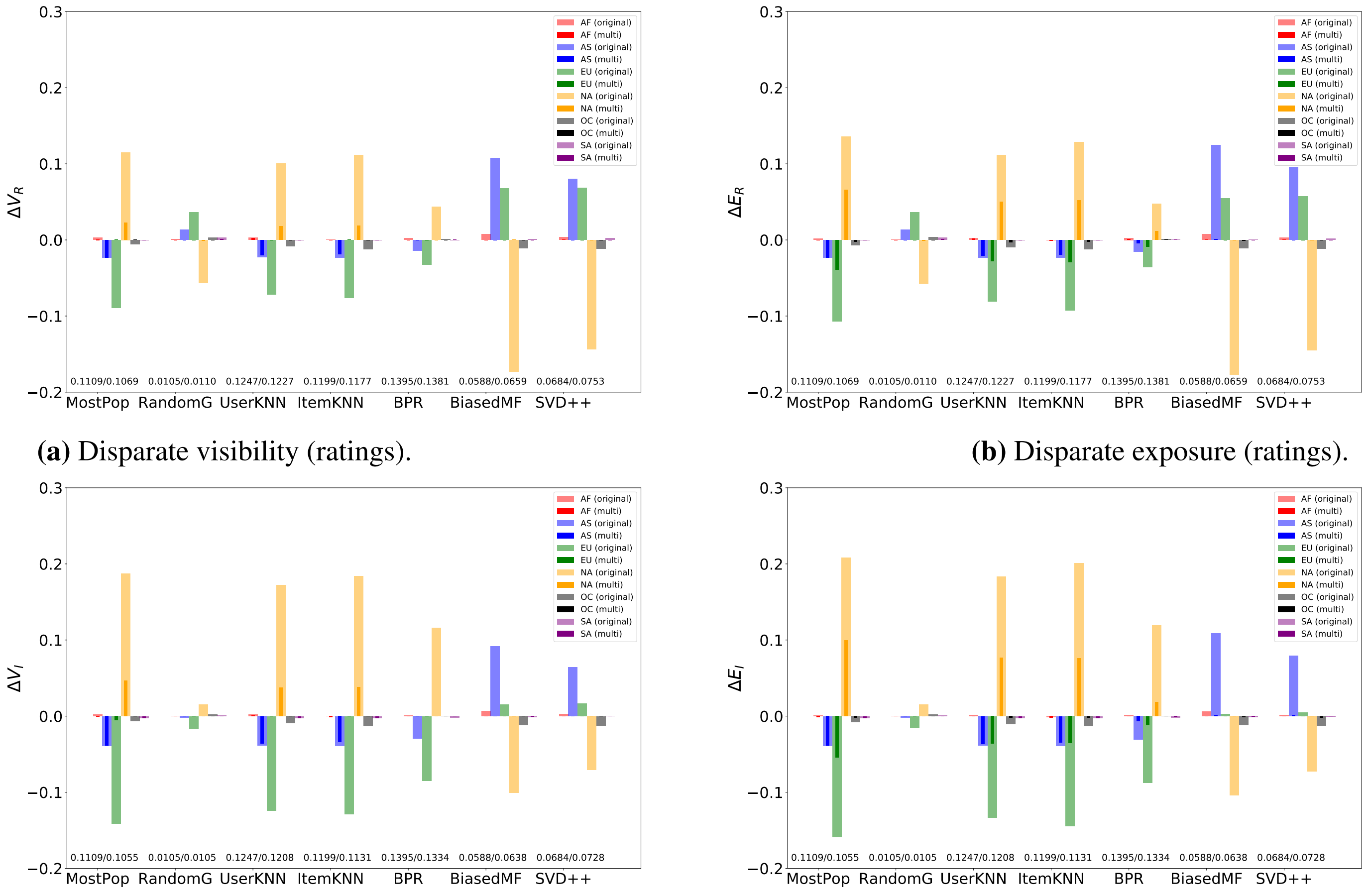
Fig. 4.3 Disparate impact in the Books dataset. Disparate impact returned by the state-of-the-art models (thick bars) and by the binary mitigation proposed in [78] (thin bars). Each figure contains one section for each algorithm and a color for each continent. The text at the bottom of each figure contains the NDCG returned by the original model and after the binary mitigation, separated by a "/". In (a) and (b), we report the disparate visibility and disparate exposure obtained when considering a rating-based representation, while in (c) and (d), the disparate visibility and disparate exposure obtained when considering an item-based representation.

```
define mitigationContinent ($list$, $reRankingType$, $targetProportions$)
begin
    // initializes four empty lists to store candidate items to add, candidate items to remove, all possible swaps of items, and the disparities per continent, respectively
    $itemsIn, itemsOut, possibleSwaps, continentList \leftarrow list(), list(), list(), list()$ ;
    $proportions \leftarrow initialProportions(list, reRankingType)$; // compute continents' proportions in the ranked list
    $continentList \leftarrow proportions - targetProportions$ ; // updates disparity of each continent
    foreach $user \in list$ do // for each user
        foreach $list.item \in$ top-n do // we loop over all items that belong to this user
            if checkPosition(list.item, itemsOut, reRankingType)==True and checkDisadvantagedGroup(list.continent,continentList)==False then
                itemsOut.add(list.item) ; // adds the item as possible candidate to move out if it belongs to an advantaged group and belongs to the top-k
            else if checkPosition(list.item, itemsOut, reRankingType)==False and checkDisadvantagedGroup(list.continent,continentList)==True then
                itemsIn.add(list.item) ; // adds the item as possible candidate to move in if it belongs to a disadvantaged group and it is not in the top-k
            end
        end
        while !itemsIn.empty() and !itemsOut.empty() do
            $itemIn \leftarrow itemsIn.pop(first)$; // item ranked higher in the top-n, outside the top-k
            $itemOut \leftarrow itemsOut.pop(last)$; // item ranked lower in the top-k
            $loss \leftarrow itemOut.prediction - itemIn.prediction$ ; // computes the loss
            possibleSwaps.add(id, user, itemOut, itemIn, loss); // adds the possible swap
        end
    end
    sortByLoss(possibleSwaps); // sort candidate swaps by loss, from minor to major
    $i \leftarrow 0$;
    // do swaps until the target proportions are reached or no more swaps
    while $proportions < targetProportions$ and i < len(possibleSwaps) do
        $elem \leftarrow possibleSwaps.get(i)$ ; // gets candidate swap with the minor loss
        if checkPosition(elem.id, elem.itemOut, reRankingType)==True and checkDisadvantagedGroup(elem.itemIn.continent,continentList)==False then
            $list \leftarrow$ swap$(list, elem.itemOut, elem.itemIn)$; // makes the swap of items
            // computes exposure difference
            $exp \leftarrow itemOut.exposure - itemIn.exposure$ ;
            // reduces continents' proportions for the itemOut
            $proportions \leftarrow updateProportions(elem.itemOut, reRankingType, exp, -1)$;
            // adds continents' proportions for the itemIn
            $proportions \leftarrow updateProportions(elem.itemIn, reRankingType, exp, 1)$;
            // updates continent's disparities
            $continentList \leftarrow proportions - targetProportions$ ;
        $i \leftarrow i + 1$ ; // advances to the next possible swap with minor loss
    end
    return $list$ ; // re-ranked list
end
```

Algorithm 4: Support method to the multiclass mitigation algorithm

```
define checkPosition($item, itemsOut, reRankingType$) // check the position of an item
begin
    if $reRankingType$ == "visibility" then return $item.position < top\text{-}k$ ;
    else if $reRankingType$ == "exposure" then return $item.position < itemsOut.last.position$ ;
end
define checkDisadvantagedGroup (continent, continentList) // check disadvantaged continent
begin
    for $cont \in continent$ do sumDeltas += continentList.get(cont) ; // adds the disparity of the continent
    return (sumDeltas $>$ 0);
end
define initialProportions($list, reRankingType$) // check initial continents' proportions
begin
    $proportions \leftarrow 0$; // set up each continent' proportion to 0
    foreach $user \in list$ do // for each user
        foreach $list.item \in top\text{-}k$ do // we loop over the top-k items that belong to this user
            if $reRankingType$ == "visibility" then
                for $cont \in list.continent$ do proportions[cont] += 1 ;
            else if $reRankingType$ == "exposure" then
                for $cont \in list.continent$ do proportions[cont] += list.exposure ;
            end
        end
    end
    return proportions
end
define updateProportions($item, reRankingType, exp, value$) // update proportions after a swap
begin
    if $reRankingType$ == "visibility" then
        for $cont \in item.continent$ do proportions[cont] += (1 $\times$ value) ;
    else if $reRankingType$ == "exposure" then
        for $cont \in item.continent$ do proportions[cont] += ( exp $\times$ value) ;
    end
    return proportions
end
```

Algorithm 5: Support methods for the mitigationContinent method



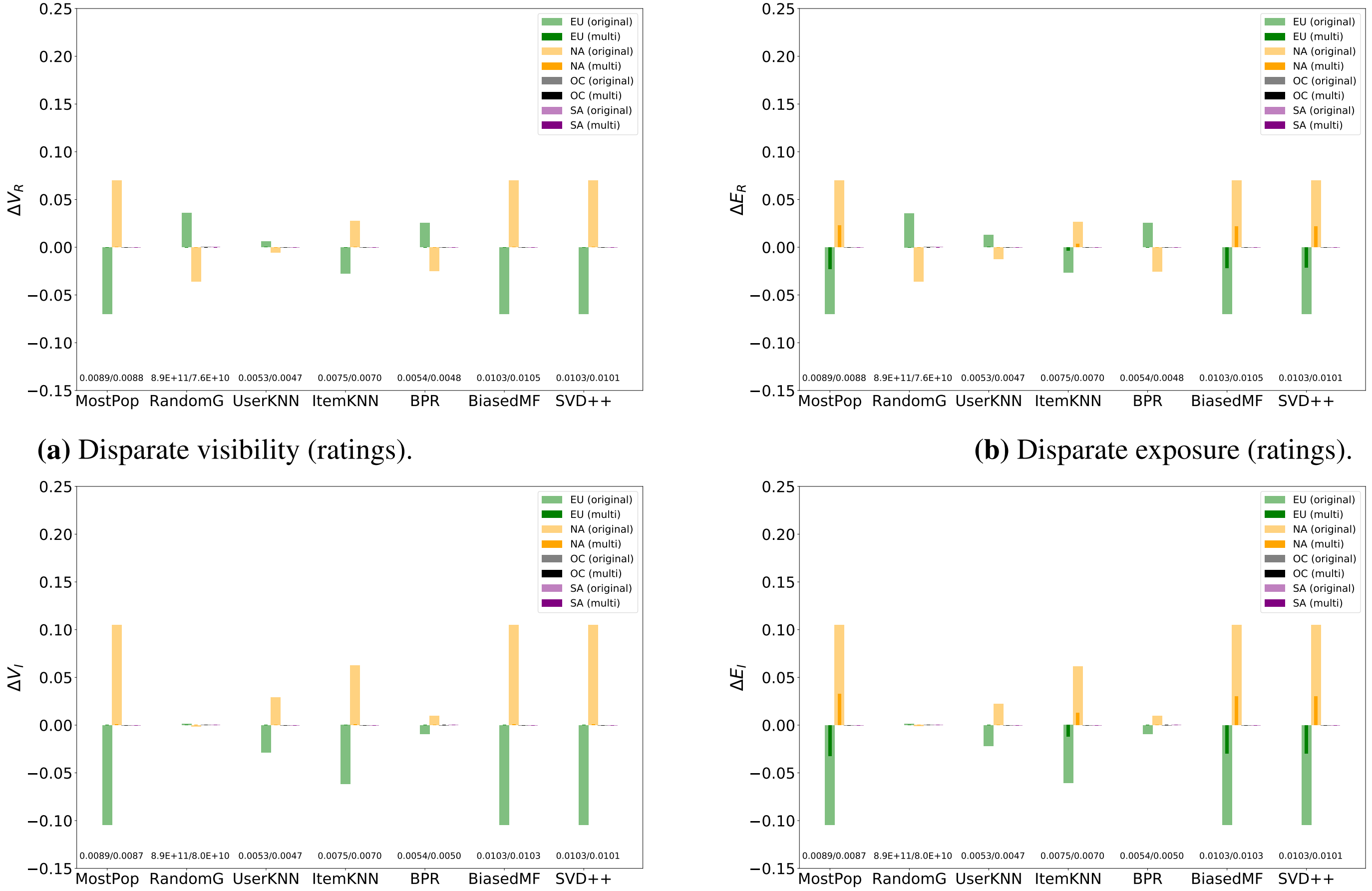
(a) Disparate visibility (ratings).

(b) Disparate exposure (ratings).

(c) Disparate visibility (items).

(d) Disparate exposure (items).

Fig. 4.4 Disparate impact in the Movies dataset after mitigation. Disparate impact returned by the state-of-the-art models (thick bars) and by our multi-group mitigation algorithm (thin bars). Each figure contains one section for each algorithm and a color for each continent. The text at the bottom of each figure contains the NDCG returned by the original model and after the binary mitigation, separated by a "/". In (a) and (b), we report the disparate visibility and disparate exposure obtained when considering a rating-based representation, while in (c) and (d), the disparate visibility and disparate exposure obtained when considering an item-based representation.



(a) Disparate visibility (ratings).

(b) Disparate exposure (ratings).

(c) Disparate visibility (items).

(d) Disparate exposure (items).

Fig. 4.5 Disparate impact in the Books dataset after mitigation. Disparate impact returned by the state-of-the-art models (thick bars) and by our multi-group mitigation algorithm (thin bars). Each figure contains one section for each algorithm and a color for each continent. The text at the bottom of each figure contains the NDCG returned by the original model and after the binary mitigation, separated by a "/". In (a) and (b), we report the disparate visibility and disparate exposure obtained when considering a rating-based representation, while in (c) and (d), the disparate visibility and disparate exposure obtained when considering an item-based representation.

Table 4.3 Disparate impact with different mitigation strategies in the Movies dataset. Disparate impact metrics returned by the different models for each continent (AF: Africa, AS: Asia, EU: Europe, NA: North America, OC: Oceania, SA: South America) considering the Movies data. For each algorithm, we report the results obtained by the binary and by our multi-group mitigation, in terms of disparate visibility and exposure when considering the rating-based representation as a reference ($\Delta\mathcal{V}_R$ and $\Delta\mathcal{E}_R$ lines) and with the item-based representation ($\Delta\mathcal{V}_I$ and $\Delta\mathcal{E}_I$ lines). Under each metric, we report the gain or loss we obtained when moving from the binary to our multi-group mitigation.

		MOVIES											
		AF		**AS**		**EU**		**NA**		**OC**		**SA**	
		binary	**multi**	**binary**	**multi**	**binary**	**multi**	**binary**	**multi**	**binary**	**multi**	**binary**	**multi**
MostPop	$\Delta\mathcal{V}_R$	0.0060	0.0007	-0.0228	-0.0230	-0.0062	0.0001	0.0237	0.0226	-0.0003	0.0000	-0.0003	-0.0003
	(gain/loss)	0.0028	-0.0025	0.0005	0.0003	0.0831	0.0894	-0.0914	-0.0925	0.0050	0.0053	0.0000	0.0000
	$\Delta\mathcal{E}_R$	0.0051	-0.0005	-0.0229	-0.0231	-0.0042	-0.0392	0.0233	0.0655	-0.0009	-0.0024	-0.0003	-0.0003
	(gain/loss)	0.0033	-0.0023	0.0005	0.0002	0.1026	0.0676	-0.1124	-0.0702	0.0060	0.0046	0.0000	0.0000
	$\Delta\mathcal{V}_I$	0.0182	-0.0005	-0.0378	-0.0389	-0.0047	-0.0053	0.0133	0.0468	0.0136	0.0003	-0.0025	-0.0025
	(gain/loss)	0.0161	-0.0026	0.0012	0.0002	0.1369	0.1363	-0.1741	-0.1406	0.0200	0.0067	0.0000	0.0000
	$\Delta\mathcal{E}_I$	0.0188	-0.0015	-0.0380	-0.0389	-0.0007	-0.0543	0.0100	0.0995	0.0124	-0.0022	-0.0025	-0.0025
	(gain/loss)	0.0180	-0.0023	0.0011	0.0002	0.1584	0.1047	-0.1979	-0.1085	0.0205	0.0058	0.0000	0.0000
RandomG	$\Delta\mathcal{V}_R$	-0.0002	0.0000	0.0036	0.0000	-0.0209	0.0000	0.0168	-0.0011	-0.0012	0.0000	0.0020	0.0011
	(gain/loss)	-0.0009	-0.0007	-0.0100	-0.0136	-0.0569	-0.0360	0.0734	0.0555	-0.0045	-0.0033	-0.0011	-0.0019
	$\Delta\mathcal{E}_R$	-0.0003	0.0000	0.0036	0.0000	-0.0211	-0.0001	0.0170	-0.0009	-0.0011	0.0000	0.0020	0.0010
	(gain/loss)	-0.0009	-0.0006	-0.0099	-0.0135	-0.0577	-0.0367	0.0740	0.0561	-0.0044	-0.0033	-0.0011	-0.0020
	$\Delta\mathcal{V}_I$	-0.0003	0.0000	-0.0027	0.0000	-0.0191	0.0000	0.0193	0.0000	0.0020	0.0000	0.0009	0.0000
	(gain/loss)	0.0000	0.0003	-0.0005	0.0022	-0.0029	0.0163	0.0036	-0.0157	-0.0002	-0.0022	0.0000	-0.0009
	$\Delta\mathcal{E}_I$	-0.0004	0.0000	-0.0029	0.0000	-0.0191	0.0000	0.0196	0.0000	0.0020	0.0000	0.0008	0.0000
	(gain/loss)	0.0000	0.0004	-0.0006	0.0023	-0.0034	0.0157	0.0042	-0.0153	-0.0002	-0.0022	-0.0001	-0.0008
UserKNN	$\Delta\mathcal{V}_R$	0.0056	0.0023	-0.0220	-0.0199	-0.0022	0.0000	0.0231	0.0179	-0.0042	0.0000	-0.0003	-0.0003
	(gain/loss)	0.0025	-0.0007	0.0008	0.0029	0.0697	0.0719	-0.0771	-0.0824	0.0041	0.0083	0.0000	0.0000
	$\Delta\mathcal{E}_R$	0.0052	0.0019	-0.0222	-0.0208	-0.0009	-0.0277	0.0235	0.0499	-0.0053	-0.0029	-0.0003	-0.0003
	(gain/loss)	0.0028	-0.0006	0.0008	0.0022	0.0802	0.0534	-0.0878	-0.0614	0.0040	0.0064	0.0000	0.0000
	$\Delta\mathcal{V}_I$	0.0069	0.0009	-0.0365	-0.0359	0.0049	-0.0002	0.0278	0.0377	-0.0006	0.0000	-0.0025	-0.0025
	(gain/loss)	0.0048	-0.0012	0.0021	0.0027	0.1290	0.1239	-0.1447	-0.1349	0.0087	0.0094	0.0000	0.0000
	$\Delta\mathcal{E}_I$	0.0065	0.0004	-0.0368	-0.0364	0.0064	-0.0360	0.0286	0.0768	-0.0021	-0.0023	-0.0025	-0.0025
	(gain/loss)	0.0050	-0.0011	0.0020	0.0024	0.1397	0.0974	-0.1550	-0.1068	0.0083	0.0081	0.0000	0.0000
ItemKNN	$\Delta\mathcal{V}_R$	0.0059	0.0000	-0.0230	-0.0184	0.0083	0.0002	0.0180	0.0185	-0.0088	0.0000	-0.0003	-0.0003
	(gain/loss)	0.0053	-0.0006	0.0004	0.0050	0.0847	0.0767	-0.0937	-0.0932	0.0033	0.0121	0.0000	0.0000
	$\Delta\mathcal{E}_R$	0.0055	-0.0008	-0.0231	-0.0192	0.0097	-0.0294	0.0173	0.0520	-0.0091	-0.0023	-0.0003	-0.0003
	(gain/loss)	0.0058	-0.0005	0.0003	0.0042	0.1021	0.0630	-0.1115	-0.0768	0.0032	0.0101	0.0000	0.0000
	$\Delta\mathcal{V}_I$	0.0085	-0.0015	-0.0382	-0.0341	0.0090	0.0000	0.0281	0.0381	-0.0049	0.0000	-0.0025	-0.0025
	(gain/loss)	0.0089	-0.0011	0.0010	0.0050	0.1378	0.1288	-0.1559	-0.1459	0.0084	0.0132	0.0000	0.0000
	$\Delta\mathcal{E}_I$	0.0078	-0.0022	-0.0383	-0.0345	0.0123	-0.0354	0.0264	0.0764	-0.0057	-0.0018	-0.0025	-0.0025
	(gain/loss)	0.0091	-0.0009	0.0009	0.0047	0.1570	0.1092	-0.1747	-0.1247	0.0077	0.0117	0.0000	0.0000
BPR	$\Delta\mathcal{V}_R$	0.0028	0.0000	-0.0117	-0.0001	-0.0047	0.0000	0.0106	0.0000	0.0029	0.0000	0.0000	0.0001
	(gain/loss)	0.0006	-0.0022	0.0024	0.0140	0.0280	0.0326	-0.0330	-0.0436	0.0020	-0.0008	0.0000	0.0001
	$\Delta\mathcal{E}_R$	0.0033	0.0009	-0.0129	-0.0041	-0.0050	-0.0086	0.0109	0.0113	0.0035	0.0003	0.0001	0.0002
	(gain/loss)	0.0007	-0.0017	0.0023	0.0110	0.0308	0.0271	-0.0363	-0.0359	0.0024	-0.0007	0.0000	0.0001
	$\Delta\mathcal{V}_I$	0.0033	0.0001	-0.0229	0.0000	0.0016	0.0000	0.0137	0.0000	0.0063	0.0002	-0.0021	-0.0003
	(gain/loss)	0.0021	-0.0011	0.0069	0.0298	0.0865	0.0849	-0.1022	-0.1159	0.0066	0.0005	0.0001	0.0018
	$\Delta\mathcal{E}_I$	0.0038	0.0005	-0.0243	-0.0065	0.0015	-0.0117	0.0142	0.0183	0.0068	0.0001	-0.0021	-0.0007
	(gain/loss)	0.0022	-0.0012	0.0067	0.0244	0.0895	0.0763	-0.1053	-0.1012	0.0069	0.0001	0.0001	0.0015
BiasedMF	$\Delta\mathcal{V}_R$	0.0026	0.0000	0.0468	0.0000	-0.0322	0.0000	-0.0060	0.0000	-0.0113	0.0000	0.0000	0.0000
	(gain/loss)	-0.0051	-0.0077	-0.0608	-0.1076	-0.0996	-0.0674	0.1669	0.1729	-0.0005	0.0108	-0.0009	-0.0010
	$\Delta\mathcal{E}_R$	0.0026	0.0001	0.0501	0.0009	-0.0358	0.0000	-0.0056	0.0000	-0.0112	-0.0009	0.0000	-0.0001
	(gain/loss)	-0.0048	-0.0073	-0.0744	-0.1236	-0.0907	-0.0549	0.1711	0.1767	-0.0005	0.0098	-0.0007	-0.0007
	$\Delta\mathcal{V}_I$	0.0040	0.0003	0.0594	0.0005	-0.0432	0.0000	-0.0060	0.0000	-0.0122	0.0000	-0.0020	-0.0007
	(gain/loss)	-0.0027	-0.0065	-0.0324	-0.0913	-0.0583	-0.0151	0.0945	0.1005	-0.0003	0.0119	-0.0007	0.0005
	$\Delta\mathcal{E}_I$	0.0038	0.0003	0.0685	0.0015	-0.0532	0.0000	-0.0049	0.0000	-0.0122	-0.0011	-0.0021	-0.0007
	(gain/loss)	-0.0026	-0.0061	-0.0402	-0.1072	-0.0558	-0.0026	0.0994	0.1044	-0.0003	0.0107	-0.0005	0.0009
SVD++	$\Delta\mathcal{V}_R$	0.0012	0.0004	0.0381	0.0000	-0.0343	0.0000	0.0070	0.0000	-0.0116	-0.0004	-0.0003	0.0000
	(gain/loss)	-0.0027	-0.0034	-0.0419	-0.0800	-0.1029	-0.0686	0.1503	0.1433	-0.0004	0.0109	-0.0024	-0.0021
	$\Delta\mathcal{E}_R$	0.0009	0.0005	0.0427	0.0000	-0.0374	0.0000	0.0058	0.0000	-0.0117	-0.0004	-0.0003	-0.0001
	(gain/loss)	-0.0020	-0.0024	-0.0526	-0.0954	-0.0943	-0.0569	0.1510	0.1452	-0.0003	0.0111	-0.0017	-0.0015
	$\Delta\mathcal{V}_I$	0.0011	0.0000	0.0440	0.0000	-0.0384	0.0000	0.0076	0.0000	-0.0125	0.0000	-0.0018	0.0000
	(gain/loss)	-0.0018	-0.0029	-0.0202	-0.0642	-0.0547	-0.0164	0.0786	0.0710	-0.0001	0.0124	-0.0017	0.0001
	$\Delta\mathcal{E}_I$	0.0005	0.0000	0.0547	0.0019	-0.0474	0.0001	0.0069	0.0000	-0.0127	-0.0019	-0.0021	-0.0001
	(gain/loss)	-0.0014	-0.0019	-0.0249	-0.0777	-0.0521	-0.0046	0.0798	0.0729	-0.0001	0.0106	-0.0013	0.0007

Table 4.4 Disparate impact with different mitigation strategies in the Books dataset. Disparate impact metrics returned by the different models for each continent (EU: Europe, NA: North America, OC: Oceania, SA: South America) considering the Books data. For each algorithm, we report the results obtained by the binary and by our multi-group mitigation, in terms of disparate visibility and exposure when considering the rating-based representation as a reference ($\Delta\mathcal{V}_R$ and $\Delta\mathcal{E}_R$ lines) and with the item-based representation ($\Delta\mathcal{V}_I$ and $\Delta\mathcal{E}_I$ lines). Under each metric, we report the gain or loss we obtained when moving from the binary to our multi-group mitigation.

BOOKS									
		EU		**NA**		**OC**		**SA**	
		binary	**multi**	**binary**	**multi**	**binary**	**multi**	**binary**	**multi**
MostPop	$\Delta\mathcal{V}_R$	0.0102	0.0000	-0.0099	0.0003	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)	0.0800	0.0697	-0.0800	-0.0697	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_R$	0.0102	-0.0227	-0.0099	0.0230	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)	0.0800	0.0471	-0.0800	-0.0471	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{V}_I$	0.0157	0.0000	-0.0151	0.0006	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)	0.1199	0.1042	-0.1199	-0.1042	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_I$	0.0157	-0.0322	-0.0151	0.0328	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)	0.1200	0.0720	-0.1200	-0.0720	0.0000	0.0000	0.0000	0.0000
RandomG	$\Delta\mathcal{V}_R$	-0.0026	0.0000	0.0025	0.0000	0.0001	0.0000	0.0000	0.0000
	(gain/loss)	-0.0383	-0.0357	0.0385	0.0360	-0.0002	-0.0003	0.0000	0.0000
	$\Delta\mathcal{E}_R$	-0.0028	0.0000	0.0027	0.0000	0.0001	0.0000	0.0000	0.0000
	(gain/loss)	-0.0384	-0.0356	0.0386	0.0359	-0.0002	-0.0003	0.0000	-0.0001
	$\Delta\mathcal{V}_I$	-0.0033	0.0000	0.0033	0.0000	0.0000	0.0000	0.0000	0.0000
	(gain/loss)	-0.0045	-0.0012	0.0045	0.0012	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_I$	-0.0035	0.0000	0.0035	0.0000	-0.0001	0.0000	0.0000	0.0000
	(gain/loss)	-0.0046	-0.0011	0.0046	0.0011	0.0000	0.0000	0.0000	0.0000
UserKNN	$\Delta\mathcal{V}_R$	0.0057	0.0001	-0.0055	0.0000	-0.0001	-0.0001	-0.0001	-0.0001
	(gain/loss)	-0.0002	-0.0058	0.0002	0.0057	0.0000	0.0001	0.0000	0.0000
	$\Delta\mathcal{E}_R$	0.0062	0.0001	-0.0060	0.0000	-0.0001	-0.0001	-0.0001	-0.0001
	(gain/loss)	-0.0064	-0.0125	0.0064	0.0124	0.0000	0.0001	0.0000	0.0000
	$\Delta\mathcal{V}_I$	0.0088	0.0000	-0.0082	0.0004	-0.0004	-0.0002	-0.0001	-0.0001
	(gain/loss)	0.0373	0.0286	-0.0374	-0.0288	0.0000	0.0002	0.0000	0.0000
	$\Delta\mathcal{E}_I$	0.0095	0.0000	-0.0089	0.0004	-0.0004	-0.0003	-0.0001	-0.0001
	(gain/loss)	0.0313	0.0219	-0.0314	-0.0220	0.0000	0.0002	0.0000	0.0000
ItemKNN	$\Delta\mathcal{V}_R$	0.0062	0.0000	-0.0060	0.0002	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)	0.0335	0.0273	-0.0336	-0.0273	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_R$	0.0065	-0.0032	-0.0063	0.0034	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)	0.0328	0.0231	-0.0328	-0.0231	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{V}_I$	0.0081	0.0000	-0.0075	0.0006	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)	0.0699	0.0618	-0.0699	-0.0618	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_I$	0.0085	-0.0121	-0.0079	0.0127	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)	0.0692	0.0486	-0.0693	-0.0486	0.0000	0.0000	0.0000	0.0000
BPR	$\Delta\mathcal{V}_R$	0.0080	0.0000	-0.0079	0.0000	-0.0001	0.0000	0.0000	0.0000
	(gain/loss)	-0.0172	-0.0252	0.0172	0.0251	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_R$	0.0081	0.0000	-0.0080	0.0000	-0.0001	0.0000	0.0000	0.0000
	(gain/loss)	-0.0171	-0.0252	0.0171	0.0251	0.0000	0.0001	0.0000	0.0000
	$\Delta\mathcal{V}_I$	0.0114	0.0000	-0.0110	0.0000	-0.0003	0.0000	-0.0001	0.0000
	(gain/loss)	0.0207	0.0093	-0.0208	-0.0098	0.0000	0.0003	0.0000	0.0001
	$\Delta\mathcal{E}_I$	0.0116	0.0000	-0.0112	0.0001	-0.0003	-0.0001	-0.0001	0.0000
	(gain/loss)	0.0209	0.0093	-0.0209	-0.0096	0.0000	0.0003	0.0000	0.0001
BiasedMF	$\Delta\mathcal{V}_R$	0.0102	0.0000	-0.0099	0.0003	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)	0.0800	0.0698	-0.0800	-0.0698	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_R$	0.0102	-0.0215	-0.0099	0.0218	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)	0.0800	0.0483	-0.0800	-0.0483	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{V}_I$	0.0157	0.0000	-0.0151	0.0006	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)	0.1200	0.1043	-0.1200	-0.1043	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_I$	0.0157	-0.0296	-0.0151	0.0302	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)	0.1200	0.0747	-0.1200	-0.0747	0.0000	0.0000	0.0000	0.0000
SVD++	$\Delta\mathcal{V}_R$	0.0094	0.0000	-0.0091	0.0003	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)	0.0792	0.0698	-0.0792	-0.0698	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_R$	0.0094	-0.0213	-0.0091	0.0216	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)	0.0792	0.0485	-0.0792	-0.0485	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{V}_I$	0.0141	0.0000	-0.0134	0.0006	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)	0.1183	0.1043	-0.1183	-0.1043	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_I$	0.0140	-0.0296	-0.0134	0.0302	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)	0.1183	0.0747	-0.1183	-0.0747	0.0000	0.0000	0.0000	0.0000

Table 4.5 Disparate impact with different mitigation strategies. Disparate impact metrics returned by the different models for each continent (AF: Africa, AS: Asia, EU: Europe, NA: North America, OC: Oceania, SA: South America). For each algorithm, we report the results obtained by the baseline and by our multi-group mitigation, in terms of disparate visibility and exposure when considering the rating-based representation as a reference ($\Delta\mathcal{V}_R$ and $\Delta\mathcal{E}_R$ lines) and with the item-based representation ($\Delta\mathcal{V}_I$ and $\Delta\mathcal{E}_I$ lines).

MOVIES													
		AF		**AS**		**EU**		**NA**		**OC**		**SA**	
		multi	**baseline**	**multi**	**baseline**	**multi**	**baseline**	**multi**	**baseline**	**multi**	**baseline**	**multi**	**baseline**
MostPop	$\Delta\mathcal{V}_R$	0.0007	0.0096	-0.0230	-0.0218	0.0001	-0.0853	0.0226	0.1017	0.0000	-0.0039	-0.0003	-0.0003
	$\Delta\mathcal{E}_R$	-0.0005	0.0058	-0.0231	-0.0225	-0.0392	-0.1043	0.0655	0.1274	-0.0024	-0.0060	-0.0003	-0.0003
	$\Delta\mathcal{V}_I$	-0.0005	0.0086	-0.0389	-0.0375	-0.0053	-0.1376	0.0468	0.1740	0.0003	-0.0050	-0.0025	-0.0025
	$\Delta\mathcal{E}_I$	-0.0015	0.0048	-0.0389	-0.0383	-0.0543	-0.1566	0.0995	0.1997	-0.0022	-0.0071	-0.0025	-0.0025
RandomG	$\Delta\mathcal{V}_R$	0.0000	0.0136	0.0000	0.0359	0.0000	0.0402	-0.0011	-0.1336	0.0000	0.0280	0.0011	0.0159
	$\Delta\mathcal{E}_R$	0.0000	0.0072	0.0000	0.0255	-0.0001	0.0387	-0.0009	-0.0971	0.0000	0.0163	0.0010	0.0095
	$\Delta\mathcal{V}_I$	0.0000	0.0127	0.0000	0.0202	0.0000	-0.0121	0.0000	-0.0613	0.0000	0.0269	0.0000	0.0137
	$\Delta\mathcal{E}_I$	0.0000	0.0062	0.0000	0.0097	0.0000	-0.0136	0.0000	-0.0248	0.0000	0.0152	0.0000	0.0073
UserKNN	$\Delta\mathcal{V}_R$	0.0023	0.0075	-0.0199	-0.0211	0.0000	-0.0681	0.0179	0.0866	0.0000	-0.0045	-0.0003	-0.0003
	$\Delta\mathcal{E}_R$	0.0019	0.0050	-0.0208	-0.0220	-0.0277	-0.0788	0.0499	0.1031	-0.0029	-0.0070	-0.0003	-0.0003
	$\Delta\mathcal{V}_I$	0.0009	0.0065	-0.0359	-0.0369	-0.0002	-0.1204	0.0377	0.1589	0.0000	-0.0056	-0.0025	-0.0025
	$\Delta\mathcal{E}_I$	0.0004	0.0041	-0.0364	-0.0378	-0.0360	-0.1310	0.0768	0.1754	-0.0023	-0.0081	-0.0025	-0.0025
ItemKNN	$\Delta\mathcal{V}_R$	0.0000	0.0060	-0.0184	-0.0226	0.0002	-0.0722	0.0185	0.0996	0.0000	-0.0105	-0.0003	-0.0003
	$\Delta\mathcal{E}_R$	-0.0008	0.0028	-0.0192	-0.0230	-0.0294	-0.0898	0.0520	0.1217	-0.0023	-0.0114	-0.0003	-0.0003
	$\Delta\mathcal{V}_I$	-0.0015	0.0050	-0.0341	-0.0384	0.0000	-0.1245	0.0381	0.1719	0.0000	-0.0115	-0.0025	-0.0025
	$\Delta\mathcal{E}_I$	-0.0022	0.0018	-0.0345	-0.0388	-0.0354	-0.1421	0.0764	0.1940	-0.0018	-0.0124	-0.0025	-0.0025
BPR	$\Delta\mathcal{V}_R$	0.0000	0.0046	-0.0001	-0.0091	0.0000	-0.0293	0.0000	0.0292	0.0000	0.0043	0.0001	0.0002
	$\Delta\mathcal{E}_R$	0.0009	0.0041	-0.0041	-0.0122	-0.0086	-0.0337	0.0113	0.0386	0.0003	0.0031	0.0002	0.0002
	$\Delta\mathcal{V}_I$	0.0001	0.0037	0.0000	-0.0249	0.0000	-0.0815	0.0000	0.1015	0.0002	0.0032	-0.0003	-0.0019
	$\Delta\mathcal{E}_I$	0.0005	0.0031	-0.0065	-0.0280	-0.0117	-0.0860	0.0183	0.1109	0.0001	0.0020	-0.0007	-0.0020
BiasedMF	$\Delta\mathcal{V}_R$	0.0000	0.0177	0.0000	0.1061	0.0000	0.0707	0.0000	-0.1994	0.0000	-0.0065	0.0000	0.0114
	$\Delta\mathcal{E}_R$	0.0001	0.0130	0.0009	0.1238	0.0000	0.0566	0.0000	-0.1909	-0.0009	-0.0085	-0.0001	0.0060
	$\Delta\mathcal{V}_I$	0.0003	0.0168	0.0005	0.0903	0.0000	0.0185	0.0000	-0.1271	0.0000	-0.0076	-0.0007	0.0092
	$\Delta\mathcal{E}_I$	0.0003	0.0120	0.0015	0.1080	0.0000	0.0043	0.0000	-0.1186	-0.0011	-0.0096	-0.0007	0.0038
SVD++	$\Delta\mathcal{V}_R$	0.0004	0.0163	0.0000	0.0769	0.0000	0.0765	0.0000	-0.1838	-0.0004	-0.0066	0.0000	0.0207
	$\Delta\mathcal{E}_R$	0.0005	0.0098	0.0000	0.0938	0.0000	0.0610	0.0000	-0.1669	-0.0004	-0.0090	-0.0001	0.0112
	$\Delta\mathcal{V}_I$	0.0000	0.0153	0.0000	0.0611	0.0000	0.0243	0.0000	-0.1116	0.0000	-0.0076	0.0000	0.0185
	$\Delta\mathcal{E}_I$	0.0000	0.0089	0.0019	0.0780	0.0001	0.0088	0.0000	-0.0946	-0.0019	-0.0101	-0.0001	0.0091

Table 4.6 Disparate impact with different mitigation strategies. Disparate impact metrics returned by the different models for each continent (AF: Africa, AS: Asia, EU: Europe, NA: North America, OC: Oceania, SA: South America). For each algorithm, we report the results obtained by the baseline and by our multi-group mitigation, in terms of disparate visibility and exposure when considering the rating-based representation as a reference ($\Delta\mathcal{V}_R$ and $\Delta\mathcal{E}_R$ lines) and with the item-based representation ($\Delta\mathcal{V}_I$ and $\Delta\mathcal{E}_I$ lines).

BOOKS									
		EU		**NA**		**OC**		**SA**	
		multi	**baseline**	**multi**	**baseline**	**multi**	**baseline**	**multi**	**baseline**
MostPop	$\Delta\mathcal{V}_R$	0.0000	-0.0592	0.0003	0.0595	-0.0002	-0.0002	-0.0001	-0.0001
	$\Delta\mathcal{E}_R$	-0.0227	-0.0643	0.0230	0.0646	-0.0002	-0.0002	-0.0001	-0.0001
	$\Delta\mathcal{V}_I$	0.0000	-0.0937	0.0006	0.0943	-0.0005	-0.0005	-0.0001	-0.0001
	$\Delta\mathcal{E}_I$	-0.0322	-0.0988	0.0328	0.0995	-0.0005	-0.0005	-0.0001	-0.0001
RandomG	$\Delta\mathcal{V}_R$	0.0000	0.0378	0.0000	-0.0395	0.0000	0.0013	0.0000	0.0004
	$\Delta\mathcal{E}_R$	0.0000	0.0369	0.0000	-0.0379	0.0000	0.0007	0.0000	0.0002
	$\Delta\mathcal{V}_I$	0.0000	0.0034	0.0000	-0.0047	0.0000	0.0010	0.0000	0.0003
	$\Delta\mathcal{E}_I$	0.0000	0.0024	0.0000	-0.0030	0.0000	0.0004	0.0000	0.0002
UserKNN	$\Delta\mathcal{V}_R$	0.0001	0.0094	0.0000	-0.0093	-0.0001	-0.0001	-0.0001	-0.0001
	$\Delta\mathcal{E}_R$	0.0001	0.0157	0.0000	-0.0156	-0.0001	-0.0001	-0.0001	-0.0001
	$\Delta\mathcal{V}_I$	0.0000	-0.0251	0.0004	0.0256	-0.0002	-0.0004	-0.0001	-0.0001
	$\Delta\mathcal{E}_I$	0.0000	-0.0187	0.0004	0.0193	-0.0003	-0.0004	-0.0001	-0.0001
ItemKNN	$\Delta\mathcal{V}_R$	0.0000	-0.0248	0.0002	0.0251	-0.0002	-0.0002	-0.0001	-0.0001
	$\Delta\mathcal{E}_R$	-0.0032	-0.0248	0.0034	0.0251	-0.0002	-0.0002	-0.0001	-0.0001
	$\Delta\mathcal{V}_I$	0.0000	-0.0593	0.0006	0.0599	-0.0005	-0.0005	-0.0001	-0.0001
	$\Delta\mathcal{E}_I$	-0.0121	-0.0593	0.0127	0.0599	-0.0005	-0.0005	-0.0001	-0.0001
BPR	$\Delta\mathcal{V}_R$	0.0000	0.0262	0.0000	-0.0262	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_R$	0.0000	0.0259	0.0000	-0.0258	0.0000	-0.0001	0.0000	0.0000
	$\Delta\mathcal{V}_I$	0.0000	-0.0083	0.0000	0.0087	0.0000	-0.0003	0.0000	-0.0001
	$\Delta\mathcal{E}_I$	0.0000	-0.0086	0.0001	0.0091	-0.0001	-0.0004	0.0000	-0.0001
BiasedMF	$\Delta\mathcal{V}_R$	0.0000	-0.0698	0.0003	0.0701	-0.0002	-0.0002	-0.0001	-0.0001
	$\Delta\mathcal{E}_R$	-0.0215	-0.0698	0.0218	0.0701	-0.0002	-0.0002	-0.0001	-0.0001
	$\Delta\mathcal{V}_I$	0.0000	-0.1043	0.0006	0.1049	-0.0005	-0.0005	-0.0001	-0.0001
	$\Delta\mathcal{E}_I$	-0.0296	-0.1043	0.0302	0.1049	-0.0005	-0.0005	-0.0001	-0.0001
SVD++	$\Delta\mathcal{V}_R$	0.0000	-0.0698	0.0003	0.0700	-0.0002	-0.0002	-0.0001	-0.0001
	$\Delta\mathcal{E}_R$	-0.0213	-0.0698	0.0216	0.0700	-0.0002	-0.0002	-0.0001	-0.0001
	$\Delta\mathcal{V}_I$	0.0000	-0.1042	0.0006	0.1049	-0.0005	-0.0005	-0.0001	-0.0001
	$\Delta\mathcal{E}_I$	-0.0296	-0.1042	0.0302	0.1049	-0.0005	-0.0005	-0.0001	-0.0001

CHAPTER 5

Enabling Cross-continent Provider Fairness in Educational Recommender Systems

This chapter contains the paper entitled ”Enabling Cross-continent Provider Fairness in Educational Recommender Systems”, which presents the Multi-class approach published in the Journal Future Generation Computer Systems (FGCS 2022). In this paper, we tackle the issue of fairness among teachers from various continents within educational Recommender Systems, particularly on massive open online course (MOOC) platforms.

- Elizabeth Gómez, Carlos Shui Zhang, Ludovico Boratto, Maria Salamó, Guilherme Ramos. **Enabling cross-continent provider fairness in educational Recommender Systems**, Future Generation Computer Systems, Volume 127, 2022, Pages 435-447, ISSN 0167-739X, https://doi.org/10.1016/j.future.2021.08.025 — **Quartile: Q1 - Impact Factor (JCR): 7.5**.

Enabling Cross-continent Provider Fairness in Educational Recommender Systems

With the widespread diffusion of Massive Online Open Courses (MOOCs), educational Recommender Systems have become central tools to support students in their learning process. While most of the literature has focused on students and the learning opportunities that are offered to them, the teachers behind the recommended courses get a certain *exposure* when they appear in the final ranking. Underexposed teachers might have reduced opportunities to offer their services, so accounting for this perspective is of central importance to generate equity in the recommendation process. In this paper, we consider groups of teachers based on their geographic provenience and assess provider (un)fairness based on the continent they belong to. We consider measures of visibility and exposure, to account (i) in how many recommendations and (ii) wherein the ranking of the teachers belonging to different groups appear. We observe disparities that favor the most represented groups, and we overcome these phenomena with a re-ranking approach that provides each group with the expected visibility and exposure, thus controlling fairness of providers coming from different continents (*cross-continent provider fairness*). Experiments performed on data coming from a real-world MOOC platform show that our approach can provide fairness without affecting recommendation effectiveness.

Keywords: Educational Recommender Systems, Provider Fairness, Geographic Groups.

5.1 Introduction

Historically, Recommender Systems have been used to promote the consumption of items [153]. Their recent employment in domains such as tourism [157, 173], health [18, 174], and education [19, 21], has shown that this class of algorithms can support users in their decision-making processes, beyond pure sales and streams.

Educational Recommender Systems have particularly flourished, due to the widespread use of Massive Online Open Courses (MOOCs) [62]. In MOOC platforms, Recommender Systems learn users' learning needs and preferences, and direct them towards possible resources of interest [19]. With the recent pandemics, the subscription to MOOC platforms has increased by 25-30%[1], which makes the research on Recommender Systems in these platforms more and more relevant. Among the many types of entities that can be recommended in MOOC platforms, we focus on the main one, i.e., *course recommendation*.

Producing effective recommendations is not the sole goal in a domain such as education. Indeed, the emergence of biases, such as course popularity, can push the recommendation of only popular courses [19] or affect users' learning opportunities [127]. If we go beyond the learners' perspective and of how recommendations can affect them, to consider a multi-stakeholder perspective [3, 11], we can observe that teachers are also directly affected by how recommendations are produced. Indeed, when their courses are recommended by an algorithm, they receive a certain *exposure* in the final ranking. Under- or over-exposing, certain providers might generate or exacerbate disparities and affect the opportunities that are given to teachers to offer their services. When these disparities are associated with sensitive attributes, a recommender system unfairly *discriminates* teachers (*provider unfairness*) [3, 20].

In this paper, we focus on possible unfairness emerging from the provenience of the teachers offering the courses. Specifically, we tackle a *continent-based perspective*, considering demographic groups formed by the continent of provenience of the teachers[2]. Previous studies have shown that geographic perspectives can impact the way users consume items [10]. Delving into the context of our study, considering a geographic perspective to provider fairness is a problem of central relevance in the context of course recommendation to (i) avoid affecting teachers belonging to geographic areas that have low representation in the data, by under-recommending their courses, and (ii) increase cultural diversity in the recommendation process, by putting learners in touch with courses coming from different parts of the world. Hence, equity for providers from a geographical perspective can provide benefits to both teachers and learners.

Our study begins by assessing unfairness, considering the share of recommendations associated with a demographic group, and contextualizing it to the *representation* of the group in the data. We will consider two forms of representation, based on (i) the number of courses the

[1] https://www.classcentral.com/report/mooc-stats-pandemic/

[2] In the context of this work, we will refer to a group of teachers belonging to a certain continent simply as a "demographic group".

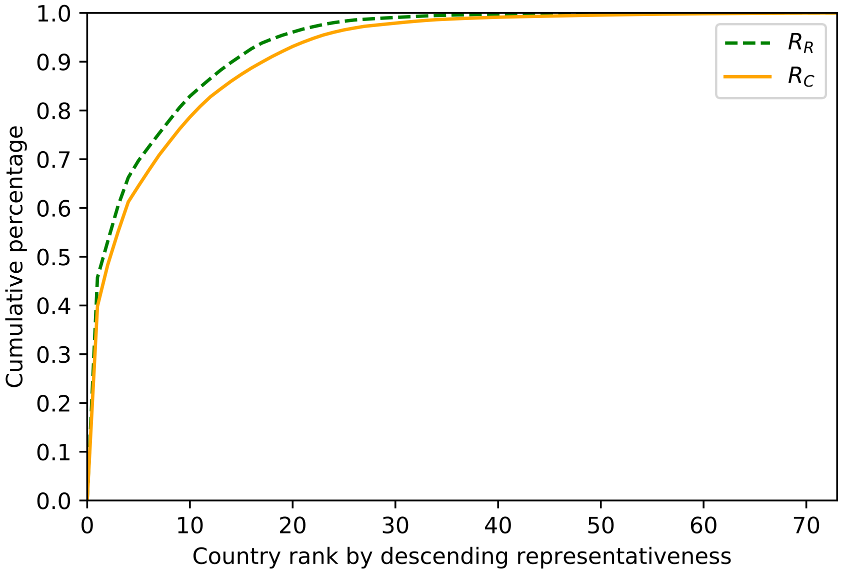


Fig. 5.1 Country imbalance. Cumulative percentage of learners' feedback (in blue) and online courses (in green) for each country in COCO [43].

teachers in a group offer and (ii) the number of interactions between learners and the courses offered by that demographic group. Specifically, we assess unfairness by considering both the *visibility* received by the teachers in a group (i.e., the percentage of recommendations having them as teachers) and their *exposure*, which accounts for the position in which courses are recommended [167]. Hence, with these two metrics, we measure, respectively, (i) the share of recommendations of a group and (ii) the relevance that is given to that group. Both metrics are relevant to assess disparate impact in this context. Visibility alone might lead a group of teachers not being reached by learners in case they appear only at the bottom of the list, and exposure alone might not guarantee that the courses of a group are being offered to enough learners (indeed, if we optimized only for exposure, then a single course at the top of the recommendation list for one learner would lead that group to get high exposure, but might mean that the opportunities for that group to get recommended to other learners are strongly reduced). We do this assessment on state-of-the-art collaborative recommendation approaches, covering both model- and memory-based approaches and point- and pair-wise algorithms.

Our choice to shape demographic groups based on their continent of provenience was made because a country-based perspective led to a too fine-grained granularity. Considering the data we work with (presented in detail in Section 5.3), the teachers come from 74 different countries. Fig. 5.1 presents the imbalance in the rating and course distributions, considering the countries in descending order, based on our two notions of representation. We can observe that the top-20 countries respectively attract and cover around 90% of the ratings and courses. This severe imbalance shows that mitigating unfairness at the country level would be unfeasible, due to the very high number of countries we deal with and the low representation of the great majority of countries. We discuss in Section 5.6 how to deal with fairness at the country level.

We mitigate disparities emerging from our previous assessment with a novel multi-class re-ranking strategy, which optimizes both the visibility and exposure given to teachers, based on their representation in the data. Thanks to our approach, we can regulate how recommendations are distributed along with the different demographic groups (*cross-continent provider fairness*), following a distributive norm based on *equity* [176].

Our contributions can be summarized as follows:

- We consider, for the first time in the literature of educational recommendation, provider fairness for demographic groups based on their geographic provenience;
- We assess unfairness on real-world data coming for a MOOC platform;
- We mitigate unfairness with a novel approach and evaluate its effectiveness.

The rest of the paper is structured as follows: in Section 5.2 we cover related work, and in Section 5.3 we provide the foundation to our study. We assess unfairness in Section 5.4 and mitigate disparities in Section 5.5. Finally, we conclude our paper in Section 5.6.

5.2 Related Work

This section presents literature related to our work. We divided it into different sections, according to the topics we analyze. First of all, we start with education Recommender Systems. Next, we overview related work on visibility and exposure in rankings. We continue by analyzing provider fairness in Recommender Systems literature and then focus on the specific topic of our work, fairness in education Artificial Intelligence. Finally, we conclude this section contextualizing our work with respect to the existing literature.

5.2.1 Educational Recommender Systems

Recommender Systems in educational platforms can involve the suggestion of different entities, such as courses [19, 186, 193], threads [184, 36], peers with whom to connect [110, 143, 83], and learning elements [58, 83]. In this section, we focus on course recommendation, which is the main focus of this paper. When designing course Recommender Systems, several sources of data are considered, such as previous user preferences [193, 194, 60] the combination between user preferences, demographic data, and pre-requisites [138], or the learning style of learners [84]. The classic recommendation models are employed to process the recommendations, namely collaborative filtering [19, 138, 193, 194], content-based filtering [186, 193], and hybrid approaches [68]. Specifically, in this work, we focus on collaborative filtering algorithms.

5.2.2 Visibility and Exposure in Rankings

Given a ranking, visibility, and exposure metrics respectively assess the number of times an item is present in the rankings [59, 192] and *where* an item is ranked [15, 191]. They were introduced in the context of non-personalized rankings, where the objects being ranked are individual users (e.g., job candidates). These metrics can operate at the *individual* level, thus guaranteeing that similar individuals are treated similarly [15, 45], or at *group* level, by making sure that users belonging to different groups are given adequate visibility or exposure [192, 191]. Under the group setting, the visibility/exposure of a group is proportional to its representation in the data [139, 158, 185].

5.2.3 Provider Fairness in Recommender Systems

The concepts of visibility and exposure have a direct impact on the providers behind the recommended items. When a system does not discriminate providers based on sensitive attributes, it is known to offer *provider fairness (P-fairness)*. P-fairness guarantees that the providers of the recommended objects that belong to different groups or are similar at the individual level, will get recommended according to their representation in the data. In this domain, Ekstrand et al. [51] assessed that collaborative filtering methods recommend books of authors of a given gender with a distribution that differs from that of the original user profiles. Liu and Burke [120] propose a re-ranking function, which balances recommendation accuracy and fairness, by dynamically adding a bonus to the items of the uncovered providers. Sonboli and Burke [168] define the concept of local fairness, to equalize access to capital across all types of businesses. Mehrotra et al. [130] assess unfairness based on the popularity of the providers. Several policies are defined to study the trade-offs between user-relevance and fairness. Kamishima et al. [102] introduce recommendation independence, which leads to recommendations that are statistically independent of sensitive features.

5.2.4 Fairness in Educational Artificial Intelligence

Defining when a user or a group of users gets discriminated by an Artificial Intelligence (AI) system highly depends on the context that is being studied [93, 85, 42, 9]. Yu et al. [188] assessed that a fair prediction, for the under-represented groups, of long- and short-term students' success is only possible if institutional data is integrated with the learning management system data. In the context of adaptive learning technologies, Doroudi and Brunskill [48] have shown that the existing algorithms can be inequitable when they rely on inaccurate models; the integration of the additive factor model, usually employed to perform knowledge tracing, can improve fairness in these systems. Hu and Rangwala [94] have focused on models that ensure individual fairness when predicting students at risk of underperforming. Individual fairness was also guaranteed to learners in course Recommender Systems, by ensuring equal learning opportunities [127].

5.2.5 Contextualizing our Work

As our analysis of the existing literature shows, our work provides novelty in the intersection of the four areas we have analyzed. Specifically, the concepts of visibility and exposure were never analyzed for demographic groups based on their provenience. None of the educational AI systems has dealt with our notion of fairness. Specifically, our work is the first to provide fairness guarantees to teachers based on their provenience, thus enabling Recommender Systems to tackle equity in the learning process from a novel perspective.

5.3 Preliminaries

Here, we present the preliminaries to provide foundations for our work. First of all, Section 5.3.1 details the recommendation scenario. Next, the metrics are described in Section 5.3.2. In Section 5.3.3, we present the recommendation algorithms. Finally, we describe the dataset used in this study in Section 5.3.4.

5.3.1 Recommendation Scenario

Let $U = \{u_1, u_2, ..., u_n\}$ be a set of learners, $C = \{c_1, c_2, ..., c_j\}$ be a set of courses, and V be a totally ordered set of values that can be used to express a preference together with a special symbol $\perp$. The set of ratings result from a map $r : U \times C \to V$, where V is the ratings' domain. If $r(u, c) = \perp$ then we say that u did not rate c. To easy notation, we denote $r(u, c)$ by r_{uc}. Now, we can define the set of ratings as $R = \{(u, c, r_{uc}) \ : \ u \in U, \, c \in C, \, r_{uc} \neq \perp\}$. These ratings can directly feed an algorithm in the form of triplets (point-wise approaches) or shape learner-course observations (pair-wise approaches).

To assess the real impact of the recommendations, we consider a temporal split of the data, where a fixed percentage of the ratings of the learners (ordered by timestamp) goes to the training and the rest goes to the test set [12].

The recommendation goal is to learn a function f that estimates the relevance ($\hat{r}_{uc}$) of the learner-course pairs that do not appear in the training data (i.e., $r_{uc} = \perp$). We denote as $\hat{R}$ the set of recommendations, and as $\hat{R}_G$ those involving courses of a group G, i.e., $\hat{R}_G = \{\hat{r}_{uc} \ : \ u \in U, \, c \in G \subseteq C\}$.

Let $A = \{a_1, a_2, ..., a_g\}$ denote the set of g geographic areas in which courses are organized. Specifically, we consider a geographic area as the continent of provenience of each teacher providing a course. We denote as A_c the set of geographic areas associated with a course c. Note that, since teachers of a course could be from different geographical areas, several geographic areas may appear in a course, and thus, $|A_c| \geq 1$. In case two teachers belong to the same geographic area, it appears only once. We use the geographic areas to shape g

demographic groups, where the ith demographic group is defined as $G_i = \{c \in C : a_i \in A_c\}$, for $i = 1, \ldots, g$.

5.3.2 Metrics

In this section, we describe the metrics used in our analysis and experiments, i.e., the representation of a group, disparate visibility, and disparate exposure.

Representation. The representation of a group is the number of times in which that group appears in the data. We consider two forms of representation, based on (i) the number of courses offered by a group and (ii) the number of ratings collected for that group. We define with $\mathcal{R}$ the *representation* of a group G ($\mathcal{R}_C$ denotes a course-based representation, while $\mathcal{R}_R$ a rating-based representation):

$$\mathcal{R}_C(G) = |G| \big/ |C| \tag{5.1}$$

$$\mathcal{R}_R(G) = |\{r_{uc} \,:\, u \in U,\, c \in G \subseteq C\}| \big/ |R| \tag{5.2}$$

Eq. (5.1) accounts for the proportion of courses of a group, while Eq. (5.2) for the proportion of ratings associated with a group. Both metrics are between 0 and 1.

The representation of a group is measured by considering only the training set.

Disparate Impact. We assess unfairness with notions of *disparate impact* generated by a recommender system. Specifically, we assess disparate impact with two metrics.

Definition 5.3.7 (Disparate visibility)**.** *The* disparate visibility *of a group is computed as the difference between the share of recommendations for items of that group and the representation of that group:*

$$\Delta\mathcal{V}(G) = \left(\frac{1}{|U|} \sum_{u \in U} \frac{|\{\hat{r}_{uc} : \hat{r}_{uc} \in \hat{R}_G,\, c \in G \subseteq C\}|}{|\hat{R}|} \right) - \mathcal{R}_*(G) \tag{5.3}$$

where '*' refers to C or R. Its range is in $[-\mathcal{R}_*(G), 1 - \mathcal{R}_*(G)]$; it is 0 when there is no disparate visibility, while negative/positive values indicate that the group received a share of recommendations lower/higher than its representation. This metric is based on that considered by Fabbri et al. [59].

Definition 5.3.8 (Disparate exposure)**.** *The* disparate exposure *of a group is the difference between the exposure obtained by the group in the recommendation lists [167] and the representation of that group:*

$$\Delta\mathcal{E}(G) = \left(\frac{1}{|U|} \sum_{u \in U} \frac{\sum_{pos=1}^{k} \frac{1}{log_2\left(\hat{r}_G^u(pos)+1\right)}}{\sum_{pos=1}^{k} \frac{1}{log_2\left(\hat{r}_C^u(pos)+1\right)}} \right) - \mathcal{R}_*(G), \tag{5.4}$$

where $\hat{r}^u_G(pos)$ denotes the rating $\hat{r}_{uc}$ that takes position pos in the list $\hat{R}^u_G = \{\hat{r}_{vc} \,:\, v = u,\, c \in G \subseteq C\}$, $u \in U$, sorted by a decreasing order.

This metric also ranges in $[-\mathcal{R}_*(G), 1-\mathcal{R}_*(G)]$; it is 0 when there is no disparate exposure, while negative/positive values indicate that the exposure given to the group in the recommendations is lower/higher than its representation.

> **Remark**. *We do not define a unique "disparate impact" metric, to control both visibility and exposure, so that* teachers are recommended enough times and with enough exposure. *A unique metric would not allow us to balance both, by compressing everything in a unique number. Later in this paper, we show why both metrics are relevant to enable provider fairness in this context.*

5.3.3 Recommendation Algorithms

In this work, we consider five state-of-the-art Collaborative Filtering approaches, which are known to be the most employed class of algorithms for course recommendation [19]. We cover both classes of point-wise and pair-wise approaches and memory-based and model-based algorithms. In addition, we consider two baseline algorithms.

Our baselines are non-personalized algorithms, which will allow us to contextualize the results obtained with different classes of approaches.

- **MostPopular** recommends items based on their popularity in the dataset, by counting the number of items an item was rated. In this way, the algorithm considers only the item perspective, without associating the ratings to the individual users and their preferences.

- **RandomGuess** establishes the maximum and minimum ratings in the data and returns a random rating for each user-item pair to predict.

For the class of memory-based approaches, we consider the following neighborhood-based algorithms:

- **UserKNN** [92] selects the K neighbors closest to the target user, and recommends the elements that other users more similar to him liked.

- **ItemKNN** [159] works in a similar way to the previous one, but in this case the target user is recommended the items that are more similar to other items that they liked before.

Matrix Factorization algorithms divide the data into matrices, representing them in latent factors to determine the degree of affinity that users and items have with those factors. For this class of approaches, we consider the following algorithms:

- **BPR.** [149] Bayesian Personalized Ranking is a state-of-the-art algorithm, optimized to generate recommendation lists, creating a probability function from the Bayesian probability function. The preference function is based on the ratings of pairs of items.

- **BiasedMF.** [108] Basic factorization of the matrix that includes the global mean, user bias, and item bias.
- **SVD++** [107] takes into account the implicit interactions, as well as the user's latent factors and the item's latent factors.

5.3.4 Dataset

We analyze data from the educational context, exploring the role of the geographic provenience of teachers in the recommendation process. We remark that the experimentation is made difficult because there are very few large-scale educational datasets coming from this specific field of online education. To the best of our knowledge, COCO [43] is the only educational dataset that contains the geographic provenience of the users. The dataset was collected from an online course platform, and includes 43,045 courses and 4,123,127 learners who gave 6,564,870 ratings. Each course is associated with one or more teachers, belonging to 74 different countries.

We pre-processed the dataset to remove all users with less than 3 ratings. Our final dataset contains 12,472 courses and 298,644 learners, which provided 1,296,598 ratings. Out of these courses, 379 are associated with two or more continents, while the rest to only to one.

We shape demographic groups considering the following continents: Africa, Asia, Europe, North America, Oceania, and South America. No course from the seventh continent (Antarctica) was available in the dataset.

Other educational datasets, proposed by [61, 195, 144], generally include $(learner, course, rating)$ triplets only, as needed in traditional recommendation scenarios, thus not fitting the problem tackled in this study (no information about the teachers' sensitive attributes is available).

5.4 Disparate Impact Assessment

In this section, we run the algorithms presented in Section 5.3.3 to assess their effectiveness and the disparate impact they generate. Before doing so, we present the experimental setting and analyze the training data, to get insights into the representation of the different groups.

5.4.1 Experimental Setting

For the dataset presented in Section 5.3.4, the test set was composed of the most recent 20% of the ratings of each learner. To run the recommendation algorithms presented in Section 5.3.3, we considered the LibRec library (version 2). For each user, we generate 100 recommendations (denoted in the paper as the top-n) so that we can mitigate the disparate impact through a re-ranking algorithm. The final recommendation list for each learner is composed of 20 courses (denoted as top-k).

We performed a grid search to optimize the hyper-parameters of each algorithm and we chose the ones that achieved the best NDCG. Intending to facilitate the reproducibility of our experiments, we detail the hyper-parameters used to run each algorithm:

- **UserKNN.** similarity: Pearson; neighbors: 50; similarity shrinkage: 10;
- **ItemKNN.** similarity: Cosine; neighbors: 200; similarity shrinkage: 10;
- **BPR.** iterator learnrate: 0.01; iterator learnrate maximum: 0.01; iterator maximum: 100; user regularization: 0.01; item regularization: 0.01; factor number: 10; learnrate bolddriver: false; learnrate decay=1.0;
- **BiasedMF.** iterator learnrate: 0.01; iterator learnrate maximum: 0.01; iterator maximum: 10; user regularization: 0.01; item regularization: 0.01; bias regularization: 0.01; number of factors: 10; learnrate bolddriver: false; learnrate decay: 1.0;
- **SVD++.** iterator learnrate: 0.01; iterator learnrate maximum: 0.01; iterator maximum: 13; user regularization: 0.01; item regularization: 0.01; impItem regularization: 0.001; number of factors: 10; learnrate bolddriver: false; learnrate decay: 1.0.

To evaluate recommendation effectiveness, we measure the ranking quality of the lists by measuring the *Normalized Discounted Cumulative Gain* (NDCG) [99].

$$DCG@k = \sum_{u \in U} \hat{r}^u_G(pos) + \sum_{pos=2}^{k} \frac{\hat{r}^u_G(pos)}{log_2(pos)}$$

$$NDCG@k = \frac{DCG@k}{IDCG@k},$$

The ideal DCG ($IDCG$) is computed by sorting courses based on decreasing true relevance (true relevance is 1 if the learner interacted with the course in the test set, 0 otherwise). The higher the better.

Table 5.1 Group representation. Course-based ($\mathcal{R}_C$) and rating-based ($\mathcal{R}_R$) representations of each group. Groups appear in alphabetical order by the name of the continent.

	$\mathcal{R}_C$	$\mathcal{R}_R$
Africa	0.0569	0.0492
Asia	0.1043	0.0526
Europe	0.1974	0.1812
North America	0.5268	0.5796
Oceania	0.0443	0.0694
South America	0.0702	0.0680

5.4.2 Characterizing Representation

The first step towards the assessment of disparate impact is to characterize the representation of the different groups in the data, which we present in Table 5.1.

The first phenomenon we can observe is that the ranking of the groups is the same, regardless of the form of representation we consider. Most of the courses are taught by North American teachers, covering almost 52.7% of the courses. Europe follows with 19.7% of the courses, and Asia takes a 10.4% share. The remainder of the groups (Africa, Oceania, and South America) have less than 10% representation. This imbalance associated with North America is exacerbated when considering the rating-based representation, where the group covers around 60% of the ratings. This leads the rest of the groups to have a lower representation w.r.t. the course-based one, regardless of Oceania, which accounts for 6.9% of the ratings. We conjecture that learners might interact with courses from Oceania because its main language is English. We performed an additional analysis of the language of the courses, which confirmed that the vast majority of the courses where teachers are from Oceania are taught in English. This analysis connects the vast number of interactions between learners and courses from North America with their interactions with courses from Oceania.

We cannot draw similar conclusions for the two spoken languages both in Europe and South America, Spanish and Portuguese. Indeed, we observed that Spanish learners following courses in Spanish, mainly do from courses that are also organized in Spain. The same holds for South American learners and the courses in Spanish they interact with, which are mainly organized in South America. For the courses in Portuguese, learners from Portugal and Brazil mainly interact with courses provided in their own country. Hence, the representations of Europe and South America are not directly affected by the fact that the continents share two languages.

Observation 1. *North America represents the majority group, with over 50% of the offered courses. These courses attract even more interactions by the learners, thus increasing the group's rating-based representation. All the other groups have a rating-based representation that is lower than the course-based one, minus Oceania. Hence, when courses are offered in English, a group attracts a share of ratings higher than the rate of courses it offers. The same does not hold for courses in Spanish and Portuguese, where learners mainly follow courses in these languages organized in their own country.*

5.4.3 Assessing Effectiveness and Disparate Impact

In this section, we report the results in terms of effectiveness (NDCG) obtained by each algorithm, and the disparate visibility and exposure associated with each demographic group, based on the two forms of representation. Table 5.2 summarizes the results.

Table 5.2 Results of state-of-the-art Recommender Systems before mitigation. Each column reports the results of an algorithm, with the first line containing the global Normalized Discounted Cumulative Gain (NDCG). The table continues with one block per demographic group, reporting (i) the Disparate Visibility when considering the rating-based representation as a reference ($\Delta\mathcal{V}_R$), (ii) Disparate Exposure when considering the rating-based representation as a reference ($\Delta\mathcal{E}_R$), (iii) Disparate Visibility when considering the course-based representation as a reference ($\Delta\mathcal{V}_C$), and (iv) Disparate Exposure when considering the course-based representation as a reference ($\Delta\mathcal{E}_C$). The underlined values indicate the best ones for each metric and demographic group, while those in bold indicate the overall best result for each metric.

		AF	**AS**	**EU**	**NA**	**OC**	**SA**
MostPop	$\Delta\mathcal{V}_R$	-0.0492	-0.0054	-0.0393	0.1364	0.0254	-0.0680
	$\Delta\mathcal{E}_R$	-0.0152	-0.0145	-0.0538	0.1057	-0.0028	-0.0194
	$\Delta\mathcal{V}_C$	-0.0569	-0.0571	-0.0556	0.1893	0.0505	-0.0702
	$\Delta\mathcal{E}_C$	-0.0226	-0.0260	-0.0575	0.1206	0.0082	-0.0228
RandomG	$\Delta\mathcal{V}_R$	0.0066	0.0504	0.0159	-0.0497	-0.0257	0.0024
	$\Delta\mathcal{E}_R$	0.0028	0.0125	0.0064	-0.0148	-0.0077	0.0009
	$\Delta\mathcal{V}_C$	-0.0011	-0.0013	-0.0003	0.0031	-0.0006	0.0001
	$\Delta\mathcal{E}_C$	-0.0004	-0.0005	-0.0002	0.0012	-0.0002	0.0000
UserKNN	$\Delta\mathcal{V}_R$	0.0259	0.0043	0.0378	0.0367	0.0505	0.0456
	$\Delta\mathcal{E}_R$	-0.0105	-0.0069	-0.0180	0.0397	0.0012	-0.0056
	$\Delta\mathcal{V}_C$	-0.0065	-0.0527	0.0081	0.0367	-0.0008	0.0153
	$\Delta\mathcal{E}_C$	-0.0128	-0.0102	-0.0196	0.0411	0.0016	-0.0001
ItemKNN	$\Delta\mathcal{V}_R$	-0.0120	-0.0122	0.0058	0.0045	0.0120	0.0019
	$\Delta\mathcal{E}_R$	-0.0052	-0.0043	-0.0016	0.0050	0.0070	-0.0009
	$\Delta\mathcal{V}_C$	-0.0197	-0.0639	-0.0105	0.0573	0.0371	-0.0016
	$\Delta\mathcal{E}_C$	-0.0066	-0.0197	-0.0066	0.0248	0.0098	-0.0003
BPR	$\Delta\mathcal{V}_R$	-0.0053	-0.0124	0.0025	0.0009	0.0096	0.0047
	$\Delta\mathcal{E}_R$	-0.0016	-0.0044	-0.0054	0.0061	0.0049	0.0004
	$\Delta\mathcal{V}_C$	-0.0130	-0.0642	-0.0138	0.0537	0.0347	0.0025
	$\Delta\mathcal{E}_C$	-0.0038	-0.0213	-0.0106	0.0275	0.0083	-0.0001
BiasedMF	$\Delta\mathcal{V}_R$	0.0331	0.0021	0.0332	0.0060	-0.0672	-0.0073
	$\Delta\mathcal{E}_R$	0.0269	-0.0072	0.0038	-0.0138	-0.0219	0.0123
	$\Delta\mathcal{V}_C$	0.0254	-0.0496	0.0169	0.0589	-0.0421	-0.0095
	$\Delta\mathcal{E}_C$	0.0271	0.0000	-0.0033	0.0049	-0.0151	0.0108
SVD++	$\Delta\mathcal{V}_R$	0.0012	-0.0010	0.0244	-0.0161	-0.0259	0.0175
	$\Delta\mathcal{E}_R$	0.0024	0.0028	0.0040	-0.0189	-0.0065	0.0162
	$\Delta\mathcal{V}_C$	-0.0130	-0.0642	-0.0138	0.0537	0.0347	0.0025
	$\Delta\mathcal{E}_C$	-0.0001	-0.0150	-0.0014	-0.0008	0.0017	0.0157

The first aspect that emerges is that the most effective algorithm in terms of NDCG is ItemKNN. Interestingly, this leads the algorithm to return, for several groups, visibility or exposure proportional to the number of ratings. This scenario is the case for the exposure in Europe and North America, obtaining the lowest $\Delta\mathcal{E}_R$, and for South America in terms of visibility ($\Delta\mathcal{V}_R$). The second most performing algorithm in terms of NDCG is BPR; we can connect this result to the analysis of the dataset made in [19], where it was observed that most of the ratings were equal to 5. Hence, most of these interactions can be treated as binary observations, leading to the capability of the algorithm to produce a good ranking in this context. For the remaining groups, this is the approach that better adjusts to the rating-based representation, in terms of visibility ($\Delta\mathcal{V}_R$) for Europe, North America, and Oceania, and of exposure ($\Delta\mathcal{E}_R$) for Africa and South America. North America and South America are also, respectively, the two groups receiving the best visibility and exposure, given to them by BPR.

Observation 2. *Ranking effectiveness is associated with good visibility and exposure when considering a rating-based representation of the groups. The ratings given by learners help to produce good recommendations and to adapt to the preferences (in terms of ratings) that each demographic group had received.*

Focusing on the course-based representation, two interesting phenomena can be observed. The first is that Random Guess is the one adapting best to the offer in terms of courses. This phenomena is the case for the visibility, $\Delta\mathcal{V}_C$, in all the groups, and for the exposure, $\Delta\mathcal{E}_C$, in Europe, Oceania, and South America. South America is also the place where the best (and almost perfect) visibility and exposure are given to a group, also thanks to Random Guess. Nevertheless, this is also the algorithm that achieves the worst NDCG. Hence, a random choice of the courses to recommend adapts well to the offer of each group but is not effective. The other algorithm offering a good course-based visibility exposure is SVD++. What we can observe here is the presence of exposure equity for both the majority group (North America) and one of the smallest ones (Africa). This means that the factors built by the algorithm capture well the original distribution of the data, thus adapting well to the course offer. Also, in this case, the NDCG of the algorithm is very low, leading to the following observation.

Observation 3. *If an algorithm can provide a group with equitable visibility and exposure, when considering its representation in terms of offered courses, then its effectiveness is very low.*

Finally, we can analyze the scenarios in which the most severe disparities can be observed. Trivially, Most Popular is the algorithm associated with the highest disparate impact values, which can be observed for North America. This result connects to previous studies on popularity bias in educational recommendation [19, 21], and extends them to the unfairness provided by an algorithm.

Observation 4. *Popularity-based recommendation exacerbates disparities, favoring the largest group and at the expense of the smallest ones.*

5.5 Mitigating Disparate Impact

The previous section allowed us to observe that groups are receiving disproportional visibility and exposure concerning their representation in the data. In this section, we propose a

Table 5.3 Results of state-of-the-art Recommender Systems after full mitigation (both visibility and exposure). Normalized Discounted Cumulative Gain (NDCG) of the original algorithm, after mitigating based on the rating-based representation ($\mathcal{V}_R \rightarrow \mathcal{E}_R$), after mitigating based on the course-based representation ($\mathcal{V}_C \rightarrow \mathcal{E}_C$), and after mitigating with the baseline.

NDCG	MostPop	RandomG	UserKNN	ItemKNN	BPR	BiasedMF	SVD++
Original	0.0193	0.0006	0.0372	**0.2068**	0.1401	0.0007	0.0044
$\mathcal{V}_R$	0.0195	0.0006	0.0368	0.2066	0.1398	0.0007	0.0045
$\mathcal{V}_C$	0.0187	0.0006	0.0367	0.2039	0.1373	0.0007	0.0043
$\mathcal{V}_R \rightarrow \mathcal{E}_R$	0.0183	0.0006	0.0340	0.2008	0.1334	0.0007	0.0045
$\mathcal{V}_C \rightarrow \mathcal{E}_C$	0.0173	0.0006	0.0342	0.1952	0.1334	0.0007	0.0043
Baseline	0.0193	0.0002	0.0376	**0.2075**	0.1400	0.0005	0.0036

re-ranking algorithm to mitigate disparities. The algorithm introduces courses of the disadvantaged groups in the recommendation list, to reach visibility and exposure proportional to their representation.

A re-ranking algorithm is the only option when optimizing ranking-based metrics, such as visibility and exposure. An in-processing regularization, such as those that have been presented in [102, 14], would not be possible, since at the prediction stage the algorithm does not predict *if and where* an item will be ranked in a recommendation list; hence, no direct comparison with these approaches is possible. This is not due to the specific choice of algorithms, since this consideration would also hold for list-wise approaches. Re-rankings have been introduced to reduce disparities, both in the context of non-personalized rankings [192, 167, 15, 31, 191, 139] and of Recommender Systems [130, 27, 120], with approaches such as Maximal Marginal Relevance [30]. However, all these algorithms optimize only one property (either visibility or exposure). As we will show later in our ablation study, optimizing for one metric is not enough. Nevertheless, we studied the impact of the approach by Liu and Burke [120] in our context, which aims at introducing provider fairness via a re-ranking approach. Concretely, the predicted relevance is increased if a provider has not appeared yet in the top-k of a user. Since we are dealing with a provider fairness setting, we increase the predicted rating if a geographic area has not appeared yet in the ranking of a user. We remind readers to [120] for the technical details of the re-ranking approach we compare with. Hyperparameter λ of the original algorithm proposed in [120] was set to 2.

5.5.1 Algorithm

Our mitigation algorithm is based on the idea to *move up in the recommendation list the course that causes the minimum loss in prediction for all the learners, until the target visibility or exposure is reached.* Our approach at introducing fairness via a re-ranking is the only one providing guarantees that equity of visibility and exposure is possible since we keep changing the recommendation list until equity from both perspectives is reached. The approaches at the state of the art, based on Maximal Marginal Relevance, make interventions on the predicted relevance for the items, thus not optimizing and not offering guarantees for the final visibility and exposure goals.

The mitigation algorithm is described in Algorithm 6, while Algorithm 7 describes the its support methods. The input is a recommendation list for all the learners (the top-n items) and the target proportions to reach of each continent. The output is the re-ranked list of courses.

The first method, called $optimizeVisibilityExposure$ (lines 1-6), calls our mitigation function twice, to have the first intervention in terms of visibility and the second one in terms of exposure. The first $mitigation$ call (line 3) is devoted to targeting the desired visibility, to make sure the courses of the disadvantaged groups are recommended enough times. This mitigation step adds the courses of the disadvantaged groups to the top-k. The second mitigation call (line 4) is devoted to regulating the exposure, by moving courses up in the top-k inside the recommendation list, to reach the target exposure.

In lines 7-40, the $mitigation$ method regulates the visibility and exposure inside the recommendation list. First of all, several lists are initialized (line 9). Next, in lines 10 and 11, the continent's proportions and their disparities are computed. Following, from line 12 to 26, the algorithm computes for each user all possible swaps of disadvantaged groups that can be done in their recommendations list. Note that it loops over all items (i.e., courses) that belong to each learner and it checks two situations, (i) the course's position in the list and (ii) if the course is in a disadvantaged group or not. So, in the end, $possibleSwaps$ contains a set of swaps, where each swap contains the user, the item to extract, the item to add in the recommendation list (top-k) of that user, and the loss we would observe if the swap was done. After that, we sort the possible swaps by loss (line 27). Next, a while loop deals with all the swaps (lines 29-38). We iterate through all possible swaps until the target proportions are reached or there are no more swaps available. Before the $swap$ method is called, we check that the candidate swap still makes sense. That is, the candidate course to move up still belongs to a disadvantaged group and the candidate to move down is still in an advantaged group. If the conditions are satisfied by the candidate swap, we proceed to make the swap and update both the group proportions and the disparities. Finally, the method returns the re-ranked list (line 39).

Algorithm 7 details the support methods called in Algorithm 6. The $checkPosition$ method (lines 1-5) is responsible for checking the position of an item in the list, taking into account if we perform a visibility or exposure mitigation. In lines 6-10, the method $checkDisadvantaged$ $Group$ verifies whether the item belongs to a disadvantaged continent or not. Note that the method contains a for loop since multiple continents may occur in a course. In that case, we compute the total sum of disparities to define a global disparity of the course. The method returns true when the disparity is positive, false otherwise. The method $initialProportions$ (lines 11-24) computes the proportion of each continent. In case of mitigating visibility it accounts the number of courses per continent and, when it mitigates exposure, it computes the sum of exposure per continent. Specifically, the $updateProportions$ method (see lines 25-33) updates the proportions per group, based on the ranking type. In case of mitigating visibility, it updates the number of courses per continent and, when it mitigates exposure, modifies the sum of exposure of each continent. Finally, the method $updateDisparity$ (lines 34-38) computes the differences between the current proportions per continent and the target proportions.

Table 5.4 Results of state-of-the-art Recommender Systems after mitigating only for visibility. Each column reports the results of an algorithm, with the first two line containing the global Normalized Discounted Cumulative Gain (NDCG) obtained after the two mitigations. The table continues with one block per demographic group, reporting (i) the Disparate Visibility when considering the rating-based representation as a reference ($\Delta\mathcal{V}_R$), (ii) Disparate Exposure when considering the rating-based representation as a reference ($\Delta\mathcal{E}_R$), (iii) Disparate Visibility when considering the course-based representation as a reference ($\Delta\mathcal{V}_C$), and (iv) Disparate Exposure when considering the course-based representation as a reference ($\Delta\mathcal{E}_C$).

		AF	**AS**	**EU**	**NA**	**OC**	**SA**
MostPop	$\Delta\mathcal{V}_R$	-0.0004	0.0000	0.0000	0.0000	0.0004	0.0000
	$\Delta\mathcal{E}_R$	0.0152	0.0145	0.0538	-0.1057	0.0028	0.0194
	$\Delta\mathcal{V}_C$	-0.0081	0.0000	0.0000	0.0000	0.0081	0.0000
	$\Delta\mathcal{E}_C$	0.0226	0.0260	0.0575	-0.1206	-0.0082	0.0228
RandomG	$\Delta\mathcal{V}_R$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_R$	-0.0028	-0.0125	-0.0064	0.0148	0.0077	-0.0009
	$\Delta\mathcal{V}_C$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_C$	0.0004	0.0005	0.0002	-0.0012	0.0002	0.0000
UserKNN	$\Delta\mathcal{V}_R$	0.0000	0.0000	0.0000	-0.0006	0.0007	0.0000
	$\Delta\mathcal{E}_R$	0.0105	0.0069	0.0180	-0.0397	-0.0012	0.0056
	$\Delta\mathcal{V}_C$	0.0000	0.0001	0.0000	-0.0001	0.0000	0.0000
	$\Delta\mathcal{E}_C$	0.0128	0.0102	0.0196	-0.0411	-0.0016	0.0001
ItemKNN	$\Delta\mathcal{V}_R$	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_R$	0.0052	0.0043	0.0016	-0.0050	-0.0070	0.0009
	$\Delta\mathcal{V}_C$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_C$	0.0066	0.0197	0.0066	-0.0248	-0.0098	0.0016
BPR	$\Delta\mathcal{V}_R$	0.0003	0.0001	-0.0001	-0.0002	-0.0001	0.0000
	$\Delta\mathcal{E}_R$	0.0016	0.0044	0.0054	-0.0061	-0.0049	-0.0004
	$\Delta\mathcal{V}_C$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_C$	0.0038	0.0213	0.0106	-0.0275	-0.0083	0.0001
BiasedMF	$\Delta\mathcal{V}_R$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_R$	-0.0269	0.0072	-0.0038	0.0138	0.0219	-0.0123
	$\Delta\mathcal{V}_C$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_C$	-0.0271	0.0000	0.0033	-0.0049	0.0151	-0.0108
SVD++	$\Delta\mathcal{V}_R$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_R$	-0.0024	-0.0028	-0.0040	0.0189	0.0065	-0.0162
	$\Delta\mathcal{V}_C$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	$\Delta\mathcal{E}_C$	0.0004	0.0005	0.0002	-0.0012	0.0002	0.0000

5.5.2 Impact of Mitigation

In this section, we analyze the impact of our mitigation algorithm, analyzing both the recommendation effectiveness and the visibility and exposure given to the different groups.

Remark. *Since our study is based on a temporal split of the data, we could not run any statistical test to assess the difference in the results between the original algorithm and our re-ranking.*

In Table 5.3, we report the results obtained by our algorithm after mitigating to regulate both visibility and exposure, having as target the rating- and course-based representations of the group.[3] Readers should note that we are reporting only the NDCG values, because we successfully mitigated both disparate visibility and exposure for all groups; all the values were exactly 0, with some minor deviations at the third or fourth decimal in very few cases. What we can observe is that the effectiveness of the algorithm shows negligible losses in both cases.

Observation 5. *Cross-continent provider fairness for demographic groups of teachers can be achieved without having a negative impact in terms of recommendation effectiveness. Thanks to our approach, we can distribute the recommendation in equitable ways between the different groups, without affecting the learners.*

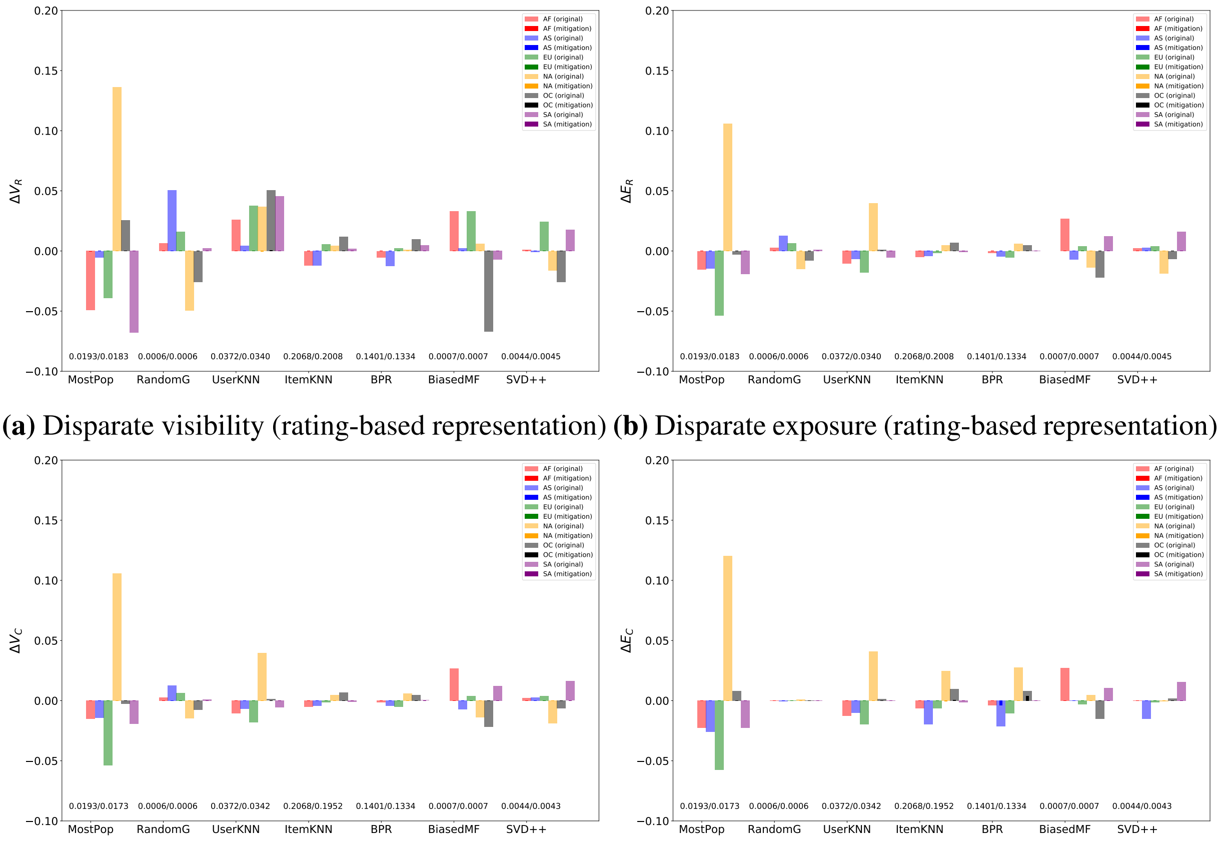
In Fig. 5.2, we visually show the benefits of moving from the original models to our mitigation in terms of disparate visibility and exposure, considering both a rating- and a course-based representation of the groups. The results confirm that we can provide consistent benefits and introduce equity, regardless of the algorithm, the metric, and the form of representation we consider.

To validate our mitigation strategy, which optimizes for both the target visibility and exposure, we run an ablation study, where we mitigate only for visibility. Results are reported in Table 5.4. The disparate visibility is mitigated by design. What we can observe is that in all of the groups and all the representations, disparate exposure is never fully mitigated. Referring to the phenomena we previously highlighted, Most Popular still over-exposes North America by 10%, at the expense of other groups, such as Europe (-5%). More broadly, we can observe that the disparate exposure values remain more or less the same as those of Table 5.2.

Observation 6. *Regulating the visibility given to a group does not provide the group with enough exposure. Disparities in terms of exposure are attenuated, but not fully mitigated. Specific interventions to regulate the given exposure are needed.*

To sum up, the ablation study shows that it is not enough to mitigate unfairness for demographic groups only considering the visibility received by the teachers in a group. Thus, our proposal of mitigating both visibility and exposure is an imperative need. The novelty of our approach comes from the idea of considering both metrics, visibility, and exposure, to address provider unfairness. It is important to remark that our results show that the proposed algorithm (see Algorithm 6) can reach the target proportions with a minimal loss in NDCG.

[3] The last line, indicating the NDCG values returned after running the mitigation with the baseline approach, will be analyzed in the context of Section 5.5.3.



(a) Disparate visibility (rating-based representation) **(b)** Disparate exposure (rating-based representation)

(c) Disparate visibility (course-based representation) **(d)** Disparate exposure (course-based representation)

Fig. 5.2 Disparate impact. Disparate impact returned by the state-of-the-art models (thick bars) and by the mitigation proposed in [76] (thin bars). Each figure contains one section for each algorithm and a color for each continent. The text at the bottom of each figure contains the NDCG returned by the original model and after the mitigation, separated by a "/". In (a) and (b), we report the disparate visibility and disparate exposure obtained when considering a rating-based representation, while in (c) and (d), the disparate visibility and disparate exposure obtained when considering a course-based representation representation.

5.5.3 Contextualization with the State of the Art

In this section, we compare the results of our mitigation with that proposed in [120]. Table 5.5 reports the obtained results.

While our approach is capable of introducing equity by mitigating both disparate visibility and exposure, as we have previously observed, this is not the case for the baseline approach in our context. Indeed, disparities are reduced by little concerning those returned by the original models, and, in some cases, they are even slightly worse. This effect is because the baseline approach favors the introduction in the top-k of courses produced in more than one continent (in other words, belonging to more than one geographic group). This observation means that, while a disadvantaged group might gain visibility and/or exposure, the accompanying group also receives the same treatment, even though it might be advantaged.

Table 5.5 Disparate impact with different mitigation strategies. Disparate impact metrics returned by the different models for each continent (AF: Africa, AS: Asia, EU: Europe, NA: North America, OC: Oceania, SA: South America). For each algorithm we report the results obtained by the baseline and by our multiclass mitigation, in terms of disparate visibility and exposure when considering the rating-based representation as a reference ($\Delta\mathcal{V}_R$ and $\Delta\mathcal{E}_R$ lines) and with the course-based representation ($\Delta\mathcal{V}_C$ and $\Delta\mathcal{E}_C$ lines).

		AF		**AS**		**EU**		**NA**		**OC**		**SA**	
		mitigation	**baseline**	**mitigation**	**baseline**	**mitigation**	**baseline**	**mitigation**	**baseline**	**mitigation**	**baseline**	**mitigation**	**baseline**
MostPop	$\Delta\mathcal{V}_R$	-0.0004	-0.0428	0.0000	-0.0039	0.0000	-0.0353	0.0000	0.1230	0.0004	0.0268	0.0000	-0.0680
	$\Delta\mathcal{E}_R$	0.0000	-0.0112	0.0002	-0.0137	-0.0002	-0.0513	0.0000	0.0974	0.0000	-0.0019	0.0000	-0.0194
	$\Delta\mathcal{V}_C$	-0.0081	-0.0505	0.0000	-0.0556	0.0000	-0.0516	0.0000	0.1759	0.0081	0.0519	0.0000	-0.0702
	$\Delta\mathcal{E}_C$	0.0000	-0.0186	0.0000	-0.0252	0.0000	-0.0550	0.0000	0.1123	0.0000	0.0091	0.0000	-0.0228
RandomG	$\Delta\mathcal{V}_R$	0.0000	0.0196	0.0000	0.0728	0.0000	0.0200	0.0000	-0.1267	0.0000	-0.0009	0.0000	0.0152
	$\Delta\mathcal{E}_R$	0.0000	0.0094	0.0000	0.0246	0.0000	0.0085	0.0000	-0.0549	0.0000	0.0053	0.0000	0.0074
	$\Delta\mathcal{V}_C$	0.0000	0.0119	0.0000	0.0211	0.0000	0.0038	0.0000	-0.0739	0.0000	0.0242	0.0000	0.0129
	$\Delta\mathcal{E}_C$	0.0000	0.0062	0.0000	0.0116	0.0000	0.0019	0.0000	-0.0389	0.0000	0.0128	0.0000	0.0065
UserKNN	$\Delta\mathcal{V}_R$	0.0000	0.0303	0.0000	0.0060	0.0000	0.0416	-0.0006	0.0230	0.0007	0.0543	0.0000	0.0456
	$\Delta\mathcal{E}_R$	0.0000	-0.0079	0.0000	-0.0059	0.0000	-0.0157	0.0000	0.0315	0.0000	0.0035	0.0000	-0.0056
	$\Delta\mathcal{V}_C$	0.0000	-0.0021	0.0001	-0.0510	0.0000	0.0119	-0.0001	0.0230	0.0000	0.0030	0.0000	0.0153
	$\Delta\mathcal{E}_C$	0.0000	-0.0102	0.0000	-0.0092	0.0000	-0.0173	0.0000	0.0329	0.0000	0.0039	0.0000	-0.0001
ItemKNN	$\Delta\mathcal{V}_R$	0.0001	-0.0066	0.0000	-0.0114	0.0000	0.0101	0.0000	-0.0076	0.0000	0.0137	0.0000	0.0019
	$\Delta\mathcal{E}_R$	0.0000	-0.0021	0.0000	-0.0039	0.0000	0.0010	0.0000	-0.0021	0.0000	0.0080	0.0000	-0.0009
	$\Delta\mathcal{V}_C$	0.0000	-0.0143	0.0000	-0.0631	0.0000	-0.0062	0.0000	0.0452	0.0000	0.0388	0.0000	-0.0016
	$\Delta\mathcal{E}_C$	0.0000	-0.0035	0.0000	-0.0193	0.0000	-0.0040	0.0005	0.0177	-0.0005	0.0108	0.0000	-0.0003
BPR	$\Delta\mathcal{V}_R$	0.0003	-0.0029	0.0001	-0.0075	-0.0001	0.0059	-0.0002	-0.0135	-0.0001	0.0130	0.0000	0.0049
	$\Delta\mathcal{E}_R$	0.0000	-0.0002	0.0000	-0.0015	0.0000	-0.0034	0.0000	-0.0025	0.0000	0.0070	0.0000	0.0005
	$\Delta\mathcal{V}_C$	0.0000	-0.0106	0.0000	-0.0593	0.0000	-0.0104	0.0000	0.0393	0.0000	0.0381	0.0000	0.0027
	$\Delta\mathcal{E}_C$	0.0000	-0.0024	0.0040	-0.0184	0.0000	-0.0086	0.0000	0.0189	-0.0040	0.0104	0.0000	0.0000
BiasedMF	$\Delta\mathcal{V}_R$	0.0000	0.0431	0.0000	0.0006	0.0000	0.0366	0.0000	-0.0206	0.0000	-0.0630	0.0000	0.0031
	$\Delta\mathcal{E}_R$	0.0000	0.0325	0.0000	-0.0079	0.0000	0.0055	0.0000	-0.0280	0.0000	-0.0196	0.0000	0.0177
	$\Delta\mathcal{V}_C$	0.0000	0.0354	0.0000	-0.0511	0.0000	0.0203	0.0000	0.0323	0.0000	-0.0379	0.0000	0.0009
	$\Delta\mathcal{E}_C$	0.0000	0.0327	0.0000	-0.0007	0.0000	-0.0016	0.0000	-0.0093	0.0000	-0.0128	0.0000	0.0162
SVD++	$\Delta\mathcal{V}_R$	0.0000	0.0136	0.0000	-0.0041	0.0000	0.0323	0.0000	-0.0566	0.0000	-0.0212	0.0000	0.0361
	$\Delta\mathcal{E}_R$	0.0000	0.0093	0.0000	0.0012	0.0000	0.0081	0.0000	-0.0406	0.0000	-0.0040	0.0000	0.0260
	$\Delta\mathcal{V}_C$	0.0000	-0.0006	0.0000	-0.0673	0.0000	-0.0059	0.0000	0.0132	0.0000	0.0394	0.0000	0.0211
	$\Delta\mathcal{E}_C$	0.0000	0.0068	0.0000	-0.0166	0.0000	0.0027	0.0000	-0.0225	0.0000	0.0042	0.0000	0.0255

The reason why the original approach can only partially mitigate disparity is since an item of the group becomes more relevant than what it was predicted, whenever that group is not yet in the top-k. Once the group is included in the recommendation list, the items stop getting a boost. However, there is no guarantee that disparities are fully mitigated. On the contrary, our approach keeps injecting items in the top-k as long as disparities are fully mitigated.

> **Observation 7**. *Introducing provider fairness requires interventions at recommendation-list level. Mitigating by boosting predicted relevance for the disadvantaged groups does not provide guarantees of equity of visibility and exposure are fully mitigated. Disparities are only partially mitigated.*

5.6 Conclusions and Future Work

Accounting for provider fairness in the recommendation process is a central aspect to account for equity in the way recommendations are produced. In this paper, we considered a course recommendation scenario and assessed unfairness for demographic groups based on the continent of provenience of the teachers. We run state-of-the-art collaborative filtering approaches on real-world data coming from a MOOC platform, and observed disparities in the visibility and exposure at the expense of the smaller demographic groups. We mitigated these disparities with a novel re-ranking multi-class approach, which adjusted the final ranking based on the target visibility and exposure, thus enabling *cross-continent provider fairness* to teachers. Results have shown that the disparities in visibility and exposure can be overcome without affecting the recommendation effectiveness for learners.

While we have highlighted that mitigating disparities at the level of individual countries can be very challenging, it is still relevant to generate equity also at this granularity. Indeed, highly represented countries inside a continent (e.g., the United States in North America) can be over-exposed, thus maintaining unfairness. In future work, we plan to introduce a two-stage process to regulate the distribution of recommendations inside a continent and guarantee fairness for teachers also at this level.

At the moment, only the dataset we considered in this study is available to study these phenomena. In the future, we plan to enrich other existing educational datasets with synthetic demographic groups to validate our approach under different scenarios.

Finally, we plan to study our multi-class mitigation in different application scenarios, such as movies or books, to study the impact of Recommender Systems in the context of pure consumption items.

```
Input: $recList$: ranked list (records contain $user$, $item$, $prediction$, $exposure$, $continent$, $position$)
       $targetProportions$: list with the target proportions of each continent
Output: $reRankedList$: ranked list adjusted by visibility and exposure
define **optimizeVisibilityExposure** ($recList$, $targetProportions$)
begin
    $reRankedList \leftarrow$ **mitigation**($recList$, “visibility”, $targetProportions$) ; // mitigation to target the desired visibility
    $reRankedList \leftarrow$ **mitigation**($reRankedList$, “exposure”, $targetProportions$) ; // mitigation to regulate the exposure
    return $reRankedList$ ; // re-ranked list adjusted by visibility and exposure
end

define **mitigation** ($list$, $reRankingType$, $targetProportions$)
begin
    $itemsIn, itemsOut, possibleSwaps, continentList \leftarrow list(), list(), list(), list()$ ; // initializes 4 empty lists to store candidate items to add to the list, candidate items to remove, all possible swaps of items, and the disparities per continent, respectively
    $proportions \leftarrow initialProportions(list, reRankingType)$; // compute continents' proportions in the ranked list
    $continentList \leftarrow updateDisparity(proportions, targetProportions)$ ; // updates disparity of each continent
    foreach $user \in list$ do // for each user
        foreach $list.item \in top\text{-}n$ do // we loop over all items that belong to this user
            if checkPosition(list.item, itemsOut, reRankingType)==True and checkDisadvantagedGroup(list.continent,continentList)==False then
                itemsOut.add(list.item) ; // adds the item as possible candidate to move out if it belongs to an advantaged group and belongs to the top-k
            else if checkPosition(list.item, itemsOut, reRankingType)==False and checkDisadvantagedGroup(list.continent,continentList)==True then
                itemsIn.add(list.item) ; // adds the item as possible candidate to move in if it belongs to a disadvantaged group and it is not in the top-k
            end
        end
        while !itemsIn.empty() and !itemsOut.empty() do
            $itemIn \leftarrow itemsIn.pop(first)$; // item ranked higher in the top-n, outside the top-k
            $itemOut \leftarrow itemsOut.pop(last)$; // item ranked lower in the top-k
            $loss \leftarrow itemOut.prediction - itemIn.prediction$ ; // compute the loss if swapped the elements in the list
            possibleSwaps.add(id, user, itemOut, itemIn, loss); // add the possible swap
        end
    end
    sortByLoss(possibleSwaps); // sort the possible swaps by loss, from minor to major
    $i \leftarrow 0$;
    // do swaps until the target proportions are reached
    while $proportions < targetProportions$ and $i <$ len(possibleSwaps) do
        $elem \leftarrow possibleSwaps.get(i)$ ; // gets candidate swap of items with the minor loss
        if checkPosition(elem.id, elem.itemOut, reRankingType)==True and checkDisadvantagedGroup(elem.itemIn.continent,continentList)==False then
            $list \leftarrow$ **swap**$(list, elem.itemOut, elem.itemIn)$; // makes the swap of the candidate with the minor loss
            $exp \leftarrow itemOut.exposure - itemIn.exposure$; ; // computes the exposure difference of the swap performed
            $proportions \leftarrow updateProportions(elem.itemOut, reRankingType, exp, -1)$; // reduces continents' proportions
            $proportions \leftarrow updateProportions(elem.itemIn, reRankingType, exp, 1)$; // adds continents' proportions
            $continentList \leftarrow updateDisparity(proportions, targetProportions)$ ; // updates disparity of each continent
        $i \leftarrow i + 1$ ; // advances to the next possible swap with minor loss
    end
    return $list$ ; // re-ranked list
end
```

Algorithm 6: Visibility and exposure mitigation algorithm

```
define checkPosition($item, itemsOut, reRankingType$) // check the position of an item
begin
    if reRankingType == "visibility" then return $item.position < top\text{-}k$ ;
    else if reRankingType == "exposure" then return $item.position < itemsOut.last.position$ ;
end
define checkDisadvantagedGroup (continent, continentList) // check disadvantaged continent
begin
    for cont $\in$ continent do sumDeltas += continentList.get(cont) ; // adds the disparity of the continent
    return (sumDeltas $>$ 0);
end
define initialProportions($list, reRankingType$) // check initial continents' proportions
begin
    proportions $\leftarrow$ 0; // set up each continent' proportion to 0
    foreach user $\in$ list do // for each user
        foreach $list.item$ $\in$ top-k do // we loop over the top-k items that belong to this user
            if reRankingType == "visibility" then
                for cont $\in$ $list.continent$ do proportions[cont] += 1 ;
            else if reRankingType == "exposure" then
                for cont $\in$ $list.continent$ do proportions[cont] += list.exposure ;
            end
        end
    end
    return proportions
end
define updateProportions($item, reRankingType, exp, value$) // update proportions after a swap
begin
    if reRankingType == "visibility" then
        for cont $\in$ $item.continent$ do proportions[cont] += (1 $\times$ value) ;
    else if reRankingType == "exposure" then
        for cont $\in$ $item.continent$ do proportions[cont] += ( exp $\times$ value) ;
    end
    return proportions
end
define updateDisparity($proportions, targetProportions$) // update disparities after a swap
begin
    $continentList \leftarrow proportions - targetProportions$
    return continentList
end
```

Algorithm 7: Support methods for the main mitigation algorithm

Part III

FAIRNESS FROM MULTIPLE PERSPECTIVES

CHAPTER 6

Bringing Equity to Coarse and Fine-Grained Provider Groups in Recommender Systems

This chapter contains the paper entitled ”Bringing Equity to Coarse and Fine-Grained Provider Groups in Recommender Systems”, which presents the CONFIGRE approach published at the Conference on User Modeling, Adaptation and Personalization (UMAP 2024). This research presents an innovative approach to enhancing fairness in Recommender Systems by concentrating on the visibility of content providers across various demographic tiers, ranging from broader classifications like continents to more granular ones such as countries. The published research manuscript included in this chapter is the following:

- Elizabeth Gómez, David Contreras, Maria Salamo, and Ludovico Boratto. 2024. **Bringing Equity to Coarse and Fine-Grained Provider Groups in Recommender Systems**. In Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24). Association for Computing Machinery (ACM), New York, NY, USA, 18–23. https://doi.org/10.1145/3627043.3659552 — **Rank: B en CORE**.

Bringing Equity to Coarse and Fine-Grained Provider Groups in Recommender Systems

Provider fairness aims at regulating the recommendation lists, so that the items of different providers/provider groups are suggested by respecting notions of equity. When group fairness is among the goals of a system, a common way is to use coarse groups since the number of considered provider groups is usually small (e.g., two genders, or three/four age groups) and the number of items per group is large. From a practical point of view, having few groups makes it easier for a platform to manage the distribution of equity among them. Nevertheless, there are sensitive attributes, such as the age or the geographic provenance of the providers that can be characterized at a fine granularity (e.g., one might group providers at the country level, instead of the continent one), which increases the number of groups and decrements the number of items per group. In this study, we show that, in large demographic groups, when considering coarse-grained provider groups, the fine-grained provider groups are under-recommended by the state-of-the-art models. To overcome this issue, in this paper, we present an approach that brings equity to both coarse and fine-grained provider groups. Experiments on two real-world datasets show the effectiveness of our approach.

Keywords: Recommender Systems, Bias, Provider Fairness, Geographic Groups, Disparate Impact.

6.1 Introduction

Provider fairness in Recommender Systems aims at ensuring that (groups of) providers receive a visibility or an exposure in the recommendation lists that is proportional to the amount of interactions of the users with these providers. When group fairness is enabled by a system, usually a small number of groups is considered [17, 134, 28, 81]. Indeed, attributes such as gender are considered as binary, while age is divided into three or four groups. In the context of demographic groups, we refer to the terms *coarse* and *fine*-grained to describe the degree of specificity in the classification of populations. *Coarse*-grained involves a more generalized composition and broad categorization of demographic groups, whereas *fine*-grained involves a very specific and detailed categorization of demographic groups. Therefore, there are sensitive attributes that can be seen at different granularities, which can allow us to characterize demographic groups in different ways. An example of a sensitive attribute currently studied in the literature of provider fairness where large groups are considered is the *geographic provenance of the providers* [77, 79, 80, 81]. Under this paradigm, item providers are usually considered as belonging to one of the seven continents. However, this attribute can be declined at finer granularities, and one might consider the country, the region, or the city of provenance, looking at this attribute as a sort of a Russian doll. Clearly, the finer the granularity at which we consider a sensitive attribute, the smaller the demographic groups we consider. The question that emerges, when considering coarse demographic groups (as in the current fairness literature), is *how the fine-grained provider groups inside each demographic group are treated in terms of fairness*.

The focus of this paper is two-fold: (i) to enhance that the use of coarse demographic groups under-recommend fine-grained provider groups; (ii) to show that a provider fairness algorithm that considers both coarse and fine-grained provider groups enhances equity and augments coverage. Specifically, in this paper, we present CONFIGRE (i.e., a COarse aNd FIne Grained RE-ranking) that not only focuses on providing fairness to large demographic groups but also enables provider fairness to each fine-grained group (in our study, the country) inside a coarse demographic group (in our study, the continent). Concretely, thanks to the use of buckets associating items to their continent and country of provenance, we ensure that the providers of a given country are given enough *visibility*, meaning that they are recommended a number of times proportional to the interaction of the users with their items. Note that our approach can be easily generalized to any granularity of the provider groups, but in the datasets we used in this work it would not make sense to consider finer granularities (e.g., region/city), due to the small representation these groups would have in the data.

6.2 Related Work

In this paper, our focus is on *provider fairness* in Recommender Systems [54, 55], which explores how different providers, either individually or as members of protected groups, have their items included (or not) in the rankings generated by a recommender system [53]. In particular, a lot of work has been done in various scenarios [168, 75, 128, 81, 64, 51, 130, 51]. Different strategies can be employed when mitigating unfairness phenomena, including data pre-processing, in-processing modifications to the model, or post-processing of recommendation lists [53]. Recent advancements include the *CP-fairness* method [134], which integrates fairness constraints from both consumer and provider perspectives into an optimization-based re-ranking approach. Furthermore, Burke *et al.* [28] introduce the SCRUF-D framework, where providers and other stakeholders are presented as agents participating in the recommendation process through a two-stage social choice mechanism. Finally, Wu *et al.* [179] propose the Multi-FR framework, a multi-objective optimization approach for fairness-aware recommendations in multi-sided marketplaces, ensuring a Pareto optimal solution.

Concerning the mitigation strategy, various policies have been devised to examine the trade-offs between user relevance and fairness. Kamishima *et al.* [102] introduced the concept of recommendation independence and developed an objective function, which aims to minimize loss and maximize independence. Tahery *et al.* [172] expanded on this analysis by considering items belonging to more than two protected groups by an algorithm called FARGO. Assessing provider fairness typically involves metrics such as visibility and exposure. Visibility measures the frequency with which an item appears in the rankings [59, 192]. Exposure evaluates the ranking position of an item, specifically for users who receive recommendations for items from each provider [15, 191]. Mehrotra *et al.* [130] introduce a fairness metric that rewards diverse recommendation lists in terms of popularity bias. Other approaches, like the one presented by Karakolis *et al* [104], aim to provide fair recommendations across item providers by considering user diversity and coverage. Raj and Ekstrand [145] provide a comparative analysis of several recently introduced fairness metrics for measuring fair rankings, and Wu *et al.* [180] formulate a family of exposure fairness metrics based on the expected exposure metric, that address fairness concerns by considering group attributes of both users and items. Recently, Chen *et al.* [183] proposed a model called P-MMF, that aims to balance provider fairness and user preference.

Existing proposals in provider fairness primarily focused on ensuring sufficient visibility for providers or groups at a coarse granularity, without considering the impact for the fine-grained groups characterized by a given sensitive attribute. In contrast, our proposal –to the best of our knowledge– is the the first that provides guarantees that also the existing fine-grained of providers are not affected by disparities and receive a fair visibility in the recommendations.

6.3 Preliminaries

Traditional recommendation scenarios are defined by a set of users, $U = \{u_1, u_2, ..., u_n\}$, that interact with a set of items, $I = \{i_1, i_2, ..., i_j\}$. A totally ordered set of values, V, can be used to express a preference together with a special symbol $\perp$. Considering a rating domain V, the set of ratings results from a map $r : U \times I \to V$. If $r(u, i) = \perp$, then we say that user u

did not rate item i. To easy notation, we denote $r(u, i)$ by r_{ui}. We define the set of ratings as $R = \{(u, i, r_{ui}) \, : \, u \in U, \, i \in I, \, r_{ui} \neq \perp\}$ and they can directly feed an algorithm in the form of triplets (point-wise approaches) or shape user-item observations (pair-wise approaches). We consider a random split of the data, where a fixed percentage of the ratings of the users goes to the training, and the rest goes to the test set. The recommendation goal is to learn a function f that estimates the relevance ($\hat{r}_{ui}$) of the user-item pairs that do not appear in the training data (i.e., $r_{ui} = \perp$). The term $\hat{R} = \{(u, i, \hat{r}_{ui}) \, : \, u \in U, \, i \in I\}$ refers to the set of recommendations.

Let $C = \{c_1, c_2, ..., c_g\}$ denote the set of g coarse-grained groups (i.e., in our case study, the geographic continents) associated with the items, and let $D = \{d_1, d_2, ..., d_h\}$ denote the set of h fine-grained groups (i.e., the set is the geographic countries) associated with items. We denote as $C_i \subseteq C$ and $D_i \subseteq D$ the set of coarse and fine-grained groups, respectively, associated with an item i. Specifically, for the geographic provenance domain, note that since an item could be produced by more than one provider, several geographic continents/countries may appear in a item, and thus, $|C_i| \geq 1$, and $|D_i| \geq 1$. In case two providers belong to the same geographic continent/country, that continent/country appears only once; this choice was made since we are dealing with group fairness, so when a group of providers is associated with an item (either once or multiple times), we account for the presence of that group. We use the geographic continents/countries to shape *g*/*h* demographic groups, which can be defined to group the ratings of the items produced in a continent/country (we denote the items in I produced in a continent $c \in C$ as I_c, and the ones produced in a country $d \in D$ as I_d). In our analysis and experiments, we use two metrics: *group representation* and *disparate visibility*.

Provider-group Representation. We compute the representation of a demographic group in the data as the number of ratings for items associated with that group in the data. We define with $\mathcal{R}$ the *representation* of a group $c \in C$ (i.e., continent of items) or $d \in D$ (i.e, country or items) as follows:

$$\mathcal{R}_* = \frac{|\{r_{ui} \, : \, u \in U, \, i \in I_*\}|}{|R|} \tag{6.1}$$

where the '*' symbol can be substituted by c or d, in case we are modeling the representation of a continent or a country, respectively. Eq. (6.1), which ranges in [0,1], accounts for the proportion of ratings given to the items of a demographic group associated with a continent or country. We compute the representation of a group only considering the training set. Trivially, the sum of the representations of all groups is equal to 1.

Disparate Visibility. Given a group $c \in C$ or $d \in D$, the *disparate visibility* returned by a recommender system for that group is measured as the difference between the share of recommendations for items of that group and the representation of that group in the input data:

$$\Delta\mathcal{V}_* = \left(\frac{1}{|U|} \sum_{u \in U} \frac{|\{\hat{r}_{ui} : i \in I_*\}|}{|\hat{R}|} \right) - \mathcal{R}_* \tag{6.2}$$

where, again, '*' can be either c or d. The range of values for this score is $[-\mathcal{R}_*, 1-\mathcal{R}_*]$; specifically, it is 0 when the recommender system has no disparate visibility, while negative/positive values indicate that the group received a share of recommendations that is lower/higher than its representation. This metric is based on Fabbri *et al.* [59].

6.4 Disparate Impact Assessment

6.4.1 Experimental Setup

We consider four state-of-the-art Collaborative Filtering algorithms: **ItemKNN** [159], **UserKNN** [92], **BPR** [149], and **SVD** [107]). In our experiments, we will report the results of the original non-fairness-aware recommendation algorithm, denoted as **OR** and two baselines a provider fairness algorithm, **P-Fair** [80], and a consumer-provider approach, **CP-Fair** [135]. We have used the P-Fair algorithm to control fairness for providers coming from different continents and countries; we denote them as **P-Fair**$_c$ and **P-Fair**$_d$, respectively. We follow a similar methodology to that described in [135] to evaluate our experiments' performance of **CP-Fair**, where we consider users as consumers and items (i.e., songs and movies) as producers. Since this algorithm does not allow a multi-group setting, as it addresses fairness as a binary setting, we used the most represented continent group (North America) versus the rest of the continents.

To run the recommendation models presented previously and generate consistently formatted lists that can be fed to CONFIGRE, P-Fair, and CP-Fair, we used the *Cornac* framework [156]. We used two datasets (publicly available at https://github.com/davidcontrerasaguilar/CONFIGRE.git): **(1)** the **MovieLens-1M (Movies)** extended to integrate the continent and the country of production of each movie, it provides 1M ratings (range 1-5), provided by 6,040 users, to 3,600 movies, 54 countries, and 6 continents; **(2)** a new created dataset called **DataSongs (Songs)**, which contains 1,777,981 ratings (range 1-5), provided by 30,759 users, to 16,380 songs, 62 countries, and 6 continents. Both datasets were randomly separated into a test (20%) and training (80%) sets. For each user, we generated the top-1000 recommendations (denoted in the paper as the top-n) to then re-rank the top-k (set up to 10) through the proposed CONFIGRE algorithm. To evaluate recommendation effectiveness, we measure the ranking quality of the lists by measuring the NDCG [99].

6.4.2 Group Representation

Analyzing the training set in the **Songs** dataset, we observe that the North American (NA) continent has the highest representation, which is close to 63.64%. Note that the United States country represents nearly 89.90% of the continent, whereas only Canada, with a percentage of 7.49%, has a percentage above 1%; therefore, the remaining 2.61% is divided between 7 North American (NA) countries. Moreover, the United States has 57.21% of the representation of the whole dataset. Regarding the European (EU) continent, the United Kingdom has a 21.26% representation in the whole dataset and a 77.67% concerning the other suppliers from the EU continent, leaving only 22.33% for the other 24 countries. As it occurs in other continents, one country has more than half of the representation concerning other countries. Similarly, this is the case of South Korea in the Asia (AS) continent, South Africa in Africa (AF), Australia in Oceania (OC), and Brazil in South America (SA). Finally, it is also important to mention that, out of 54 countries, only 7 have a representation of more than 1%, equivalent to 91.11% of the total representation. On the other hand, the **Movies** dataset has a similar behavior, where the North American (NA) continent has 57.13% representation, and the United States has 53.71% of the whole dataset. Additionally, this country has 94.01% of representation concerning the NA continent, whereas only Canada, with 4.50%, has a representation higher than 1.48%, and the remaining 1% is distributed among the other 4 countries. Finally, it is essential to highlight that only 8 of 62 countries have a representation of more than 1%, equivalent to 89.59% of the total representation. **Observation 1.** *The countries inside a continent have very different representations, with a majority country that usually attracts most of the ratings. Hence, the question of how fairly the providers of the minority countries inside a continent are recommended emerges.*

6.5 CONFIGRE: a COarse aNd FIne Grained RE-ranking provider fairness algorithm

6.5.1 Algorithm

CONFIGRE's primary objective is to reach a provider group's target percentage observed in the training set, while minimizing the loss in user predictions. By continually adjusting the recommendation list, CONFIGRE, which focuses on re-ranking, ensures the provision of fairness for providers and guarantees equity. CONFIGRE works following three main steps: **Step 1**: We compute the representation $\mathcal{R}_*$ of each demographic group (i.e., the coarse groups, $c \in C$, and the fine-grained groups, $d \in D$) considering the ratings in the training set, as in Eq. (6.1). **Step 2**: Given the items predicted as relevant for a user by the recommender system, we create a bucket list considering each item-coarse group pair, (i.e., in the experiments, item-continent pair), which will store the predicted items. In the same manner, we create a bucket for each item-fine grained group pair (i.e., in the experiments item-country pair). Each bucket comes with an attribute, which is the representation of the coarse or fine-grained group. Specifically, the recommender system returns a list of top-n recommendations (where n is much larger than the cut-off value k, to be able to perform a re-ranking). Our starting point to fill a bucket is the relevance predicted for a user u and an item i, $\hat{r}_{ui}$. Each element in the bucket is a record that contains the user ID, the item ID, and the relevance predicted by the recommender system for that user, $\hat{r}_{ui}$. We sort each bucket by item relevance. **Step 3**: We perform the re-ranking on the basis of the created bucket lists. *The goal is to guarantee fairness for coarse-grained providers and to correctly distribute the recommendations among the different fine-grained groups.* This step includes three phases, the first one is the most constrained and the conditions for selection are relaxed in the second and third phases, so to complete the recommendation list of the user. First, in **Phase 1** we select items from the least represented fine-grained groups to the most represented ones in their corresponding buckets. The algorithm selects items that satisfy the following **conditions**: **(1)** the percentage of items in the recommendation list for a fine-grained group (i.e., country in our experiments) is lower or equal to its representation ($\mathcal{R}_*$); **(2)** the number of recommended items so far is lower than k. Second, in **Phase 2** we start this phase to include more items in the recommendation list when phase 1 finishes, but the top-k is incomplete. Specifically, the selection is made in the same way as in Phase 1. However, this time we do not care that the fine-grained percentage of items is exceeded, but that it belongs to the same coarse-grained group as the item without exceeding the percentage calculated in step 2, $\mathcal{R}_*$. In this phase, condition (2) is not applied. Again, the recommendation list, top-k, may be completed or not. If completed, the process finishes; otherwise, it is necessary to move to phase 3. Finally, in **Phase 3** we complete the recommendation list if the top-k recommendations cannot be reached due to the constraints in the previous phases. In this phase, we select the items that have the greater relevance for the user until we complete the top-k.

6.5.2 Assessment of the Impact of CONFIGRE

Table 6.1 shows the disparities of the algorithms in the Songs and Movies datasets. In the Songs dataset, the sum of absolute value of disparities in the OR models (i.e., negative represents lower visibility than expected and vice versa) show that BPR produces a country disparate visibility ($\Delta\mathcal{V}_d$) of 57.5%, whereas SVD, UserKNN and ItemKNN show lower values, being 37.5%, 13.2%, and 10.2%, respectively. The recommendation lists of the OR models generate a significant geographical imbalance, especially in BPR and SVD algorithms, which only recommend 2 and 3 countries, respectively, out of the 54 countries in the dataset. The continent disparity, $\Delta\mathcal{V}_c$, in the OR models obtains better results than $\Delta\mathcal{V}_d$ in all recommendation algorithms.

Analyzing $\Delta\mathcal{V}_d$ in P-Fair$_c$, P-Fair$_d$, and CP-Fair, we can see that in the case of P-Fair$_c$, BPR and SVD produce a $\Delta\mathcal{V}_d$ almost equal to the OR. However, in UserKNN, it is reduced to 10.4%, and in ItemKNN to 8.2%. Although the P-Fair$_c$ method obtains a slight improvement, the $\Delta\mathcal{V}_d$ and $\Delta\mathcal{V}_c$ are still very high. P-Fair$_d$, presents results that are very similar to those obtained by P-Fair$_c$, with SVD being the algorithm that obtained the greatest reduction in both types of disparities. Similar results are obtained for the visibility of continents, where all results are less or equal to P-Fair$_c$. However, the number of countries recommended has been enlarged in this approach. In the case of CP-Fair, BPR produces a lower $\Delta\mathcal{V}_d$ of 43.0%, but the rest of the algorithms increase it. BPR continues to recommend items from 2 countries, but it is the only one that reduces the $\Delta\mathcal{V}_d$ with respect to the OR. SVD continues to recommend 3 countries, but

Table 6.1 Sum of the absolute value of the disparities $\Delta\mathcal{V}_d$ and $\Delta\mathcal{V}_c$ on the algorithms with respect to the training set in both datasets. Results are shown in percentages. In parentheses is the number of countries/continents recommended by each algorithm. At each row, in bold font the best $\Delta\mathcal{V}_*$ and the second one is underlined. Recommendations accuracy (NDCG) of algorithms is also shown.

	Algorithm		OR	P-Fair$_c$	P-Fair$_d$	CP-Fair	CONFIGRE
SONGS	BPR	$\Delta\mathcal{V}_d$	57.5% (2)	56.8% (3)	56.8% (13)	43.0% (2)	**6.2% (16)**
		$\Delta\mathcal{V}_c$	45.3% (3)	44.6% (2)	38.0% (4)	18.0% (2)	**0.4% (5)**
		NDCG	0.0148	0.0134	0.0151	0.0121	0.0117
	SVD	$\Delta\mathcal{V}_d$	37.5% (3)	37.5% (3)	28.7% (20)	50.4% (3)	**3.2% (26)**
		$\Delta\mathcal{V}_c$	11.9% (3)	20.0% (3)	15.3% **(5)**	27.1% (2)	**0.2% (5)**
		NDCG	0.0128	0.0129	0.0109	0.0083	0.0110
	UserKNN	$\Delta\mathcal{V}_d$	13.2% (48)	10.4% (49)	9.5% (52)	33.8% (9)	**0.5% (53)**
		$\Delta\mathcal{V}_c$	10.8% **(6)**	7.9% **(6)**	7.5% **(6)**	17.1% (4)	**0.0% (6)**
		NDCG	0.0036	0.0035	0.0037	0.0024	0.0037
	ItemKNN	$\Delta\mathcal{V}_d$	10.2% **(54)**	8.2% (54)	7.6% **(54)**	23.4% (9)	**0.0% (54)**
		$\Delta\mathcal{V}_c$	8.1% **(6)**	5.9% **(6)**	5.9% **(6)**	5.7% (4)	**0.0% (6)**
		NDCG	0.0036	0.0035	0.0036	0.0029	0.0037
MOVIES	BPR	$\Delta\mathcal{V}_d$	47.4% (5)	54.7% (17)	42.1% (29)	63.9% (3)	**16.6% (44)**
		$\Delta\mathcal{V}_c$	15.0% (2)	14.6% (4)	13.7% (4)	57.2% (2)	**11.7% (5)**
		NDCG	0.1131	0.0733	0.1132	0.0811	0.1508
	SVD	$\Delta\mathcal{V}_d$	63.9% (4)	50.4% (7)	43.1% (34)	60.6% (5)	**18.9% (42)**
		$\Delta\mathcal{V}_c$	15.1% (2)	18.6% (4)	20.8% **(6)**	16.3% (2)	**9.8% (6)**
		NDCG	0.1872	0.1552	0.1400	0.1653	0.1795
	UserKNN	$\Delta\mathcal{V}_d$	41.9% (14)	46.1% (21)	40.1% (21)	63.1% (7)	**15.0% (43)**
		$\Delta\mathcal{V}_c$	14.1% (3)	16.5% (4)	13.0% (5)	30.5% (4)	**9.1% (6)**
		NDCG	0.0563	0.0618	0.0683	0.0046	0.0531
	ItemKNN	$\Delta\mathcal{V}_d$	36.9% (35)	35.4% (42)	36.6% (40)	31.2% (16)	**5.3% (61)**
		$\Delta\mathcal{V}_c$	15.0% **(6)**	17.3% **(6)**	15.7% **(6)**	11.6% (4)	**3.5% (6)**
		NDCG	0.0791	0.0761	0.0817	0.0015	0.0792

the third country changes to one with the lowest representation in the dataset. In UserKNN and ItemKNN, the same number of countries are recommended, being much fewer than in OR, a total of 9 countries, which coincides with being the most representative ones. Note that CP-Fair is better than OR, but P-Fair$_d$ outperforms it, except for BPR.

Finally, analyzing the results of the CONFIGRE method, BPR produces a $\Delta\mathcal{V}_d$ of 6.2% (i.e., 51% lower than the OR, 57.5%), in SVD the $\Delta\mathcal{V}_d$ is reduced from 37.5% to 3.2% (-34%), UserKNN goes from 13.2% to 0.5% (-13%), and ItemKNN reduces the disparity from 10.2% to 0.0% (-10%). Moreover, our method recommends more countries for each original algorithm. For example, BPR recommends 16 countries, the SVD algorithm improves from 3 to 26 countries, UserKNN from 48 to 51, and ItemKNN continues to recommend all 54 countries. Note that CONFIGRE is the algorithm that reduces the most $\Delta\mathcal{V}_c$, being able to include more countries/continents in the recommendation lists.

In the Movies dataset, Table 6.1 shows that SVD has the highest $\Delta\mathcal{V}_d$ at 63.9%, BPR at

47.4%, UserKNN at 41.9%, and ItemKNN at 36.9%. Moreover, for these OR models, BPR recommended items from 5 different most represented countries in the dataset and, in the case of SVD, UserKNN, and ItemKNN 4, 14, and 35 countries, respectively. Using the P-Fair$_c$ approach, we obtain that SVD and ItemKNN drop the $\Delta\mathcal{V}_d$ to 50.4% and 35.4%, while BPR and UserKNN rise to 54.7% and 46.1%, respectively. Although this approach does not reduce $\Delta\mathcal{V}_d$, it increases the coverage of providers from different countries (e.g., in BPR to 17 countries). P-Fair$_d$, reduces $\Delta\mathcal{V}_d$ and increases the number of recommended countries. In CP-Fair, BPR increases $\Delta\mathcal{V}_d$ to 63.9% and UserKNN augments to 63.1%, whereas SVD reduces to 60.6% and ItemKNN to 31.2%. Moreover, considering the supplier's coverage, BPR reduces the recommended countries to 3, while SVD increases to 5.

Finally, the CONFIGRE method presents the best results, being the algorithm that better reduces the country/continent disparate visibility concerning the original algorithms. For example, in BPR, we obtain an improvement of $\Delta\mathcal{V}_d$ of 31%, reducing the country disparate visibility from 47.4% to 16.6%. In SVD, the $\Delta\mathcal{V}_d$ achieves an 18.9% (-45% w.r.t. the OR method), while in UserKNN, disparities are reduced from 41.9% to 15.0% (-27%), and in ItemKNN, from 36.9% to 5.3% (-32%). Therefore, the sum of the absolute values of the disparities shows a substantial improvement compared to the original results. Note that the coverage of countries in BPR increased from 5 to 44, SVD from 4 to 42, UserKNN from 14 to 43, and ItemKNN improved the coverage from 35 to 61 countries. Finally, it is important to note that similar to the Songs dataset, CONFIGRE is also the algorithm that reduces the most $\Delta\mathcal{V}_c$ compared to the OR and baselines models.

Additionally, we analyze how models impact the quality of recommendations using the NDCG metric. Analyzing the results obtained in the Songs dataset, we can observe that our method shows a better recommendation quality in contrast to the OR and P-Fair methods in UserKNN and ItemKNN. In the case of BRP and SVD, CONFIGRE is slightly smaller than the OR and P-Fair methods. Moreover, CONFIGRE is also better than CP-Fair in all methods except for the BPR algorithm. Regarding the Movies dataset, CONFIGRE obtains better recommendation quality concerning OR (in the case of BPR and ItemKNN) and P-Fair methods (in the case of BPR, SVD, and ItemKNN). In contrast, CONFIGRE obtained slightly lower quality recommendations than OR (UserKNN) and F-Pair (UserKNN). Finally, CONFIGRE obtained better quality recommendations than the CP-Fair method for all algorithms. **Observation 2.** *CONFIGRE can reduce both disparities (coarse and fine-grained) in the recommendations with a low impact on the recommendation quality while expanding the coverage for providers according to their country of origin.*

6.6 Conclusions

To facilitate the assessment of (un)fairness in Recommender Systems, a few, large, demographic groups are usually considered. In this study, we assessed the impact that this way of enabling fairness might have on the fine-grained groups. Specifically, we had provider fairness as a reference and studied the visibility of each coarse and fine-grained provider group in the recommendations. Our results show that the state-of-the-art approaches to regulating unfairness still bring disparities to the fine-grained groups within a demographic group. To overcome this issue, we presented an approach capable of regulating fairness for both coarse and fine-grained groups via a post-processing approach. Extensive experiments on novel datasets and against state-of-the-art baselines show that our solution can enable provider fairness at different granularities, with a negligible impact on the recommendation effectiveness.

CHAPTER 7

MOReGIn: Multi-Objective Recommendation at the Global and Individual Levels

This chapter contains the paper entitled ”MOReGIn: Multi-Objective Recommendation at the Global and Individual Levels”, which presents the MOReGIn approach published at the Conference on User Modeling, Adaptation and Personalization (ECIR 2024). The paper presents a novel approach for addressing multiple objectives in Recommender Systems, with an emphasis on achieving a balance between global fairness and individual fairness within the recommendations. The published research manuscript included in this chapter is the following:

- Elizabeth Gómez, David Contreras, Ludovico Boratto, Maria Salamó. 2024. **MOReGIn: Multi-Objective Recommendation at the Global and Individual Levels**. In: Goharian, N., et al. Advances in Information Retrieval. ECIR 2024. Lecture Notes in Computer Science, vol 14608. Springer, Cham. https://doi.org/10.1007/978-3-031-56027-9_2 — **Rank: A en CORE**.

MOREGIN: Multi-Objective Recommendation at the Global and Individual Levels

Multi-Objective Recommender Systems (MORSs) emerged as a paradigm to guarantee multiple (often conflicting) goals. Besides accuracy, a MORS can operate at the *global* level, where additional beyond-accuracy goals are met for the system as a whole, or at the *individual* level, meaning that the recommendations are tailored to the needs of each user. The state-of-the-art MORSs either operate at the global or individual level, without assuming the co-existence of the two perspectives. In this study, we show that when global and individual objectives co-exist, MORSs are not able to meet both types of goals. To overcome this issue, we present an approach that regulates the recommendation lists so as to guarantee both global and individual perspectives, while preserving its effectiveness. Specifically, as individual perspective, we tackle genre calibration and, as global perspective, provider fairness. We validate our approach on two real-world datasets, publicly released with this paper[1].

Keywords: Multi-Objective Recommendation, Calibration, Provider Fairness.

7.1 Introduction

Motivation. Since the goal of Recommender Systems is to provide relevant suggestions for the users, the main focus has been the effectiveness of the results [154]. Nevertheless, users might be interested in properties of the items besides their effectiveness, and there are other stakeholders who can benefit from how recommendations are produced (e.g., content providers). Hence, beyond-accuracy perspectives are central to the generation and evaluation of recommendations.

Multi-Objective Recommender Systems (MORSs) support the provision of perspectives that go beyond item relevance, such as, e.g., diversity, calibration, and fairness [196]. The optimization for these objectives can happen at the *global* (*aggregate*) level, thus ensuring that the system as a whole can guarantee certain properties (e.g., all providers receive a certain exposure in the recommendation lists). In alternative, a MORS can operate at the *individual* (*local*) level, and shape results that are consider the prominence of individual users towards the different goals (e.g., each user can receive a different level of diversity or the recommended genres can be calibrated to the preferences in the training set) [98].

When analyzing the current literature, a MORS either operates at the global level [72, 115, 119, 134, 179] or at the local level [141, 46, 47].

Open issues. There might be scenarios in which both global and individual objectives co-exist. Indeed, a platform might decide that, as a whole, the recommendations should offer certain properties (e.g., be fair to providers of different demographic groups, or enable a certain level of novelty). Moreover, specific goals might be set for the individual users (e.g., the calibration of the genres or the diversity of the recommended items might need to follow what is observed in the training set of each user). As we show in Section 7.6, when a MORS tackles *only* global or individual perspectives, the other perspective trivially remains under-considered and cannot be guaranteed by the system.

Our contributions. To overcome the aforementioned challenges, in this paper, we present a MORS that produces recommendations with both global and individual objectives. As a use case, we consider, as a global objective, *provider fairness* and, as an individual one, *calibrated recommendations*. This aligns our study with the rest of the MORS literature, where two beyond-accuracy objectives are considered. For the sake of clarity, we will talk about *provider-fair and calibrated recommendations* but, as we discuss in Section 7.3.2, **our approach can be generalized to any global or individual objective**.

Besides accounting for beyond-accuracy perspectives involving both global and individual objectives, the problem of providing provider-fair and calibrated recommendations becomes interesting also from a practical point of view. As we will show in Section 7.4, users tend to rate items of certain genres and that are produced in certain geographic areas, suggesting that we can account for both perspectives at the same time when generating the recommendations. Hence, at the technical level, we would need a unique solution that (i) produces effective results for the users, (ii) can provide fairness for providers belonging to different groups at the *global* level, i.e., by distributing, over the entire user base, the recommendation of items belonging to different provider groups in equitable ways, and (iii) can calibrate the recommendation lists of each *individual* user.

Our approach involves a post-processing strategy. To enable a form of provider fairness that can consider demographic groups that are not necessarily characterized by a binary group (e.g., males and females), we consider, as a sensitive attribute, the geographic provenance of the providers and have the different continents as the granularity with which we split the groups; this is aligned with recent literature on provider fairness [80, 81]. As in classic calibrated fashion, we distribute the recommendations according to the item genre. Based on this characterization of the data, we present an approach that makes use of buckets to associate the continents in which the items are produced and the genre of the items. We use these buckets to post-process the recommendation lists (we will later discuss that this is the best way to regulate both aggregate- and individual-level properties) and regulate how the recommendations are distributed across the users. Thanks to the fact that each bucket contains (i) the continent in which the item is produced, to regulate provider fairness, and (ii) the genre of the item, to regulate calibration, both global and individual perspectives are captured at once by our approach. To validate our proposal, we apply it to the recommendations produced by five algorithms, and study the effectiveness of our approach on two datasets (including a novel one, released with this study), and against state-of-the-art approaches for calibrated recommendation and provider fairness.

Concretely, our contributions can be summarized as follows:

- After the identification of the research gaps (Section 7.2) and characterization of our setting (Section 7.3), we provide the foundations to our use case, by showing that calibration and provider fairness are related problems, since the genres of the items and their continent of production are connected (Section 7.4);

- We present an approach to post-process the recommendation lists to meet both global and individual goals. We calibrate the results for the individual users in terms of genres, and are fair towards providers (Section 7.5);

- We face the limitation of evaluating this problem, due to the scarcity of data offering both the category of the items and the sensitive attributes of the providers, so we i) extend the MoviLens-1M dataset, to integrate the continent of production of each item, and ii) we collect and present (in Section 7.3) a novel dataset. Both resources are publicly available here[1];

- We perform experiments (Section 7.6) to validate our proposal when applied to the recommendation produced by five algorithms, covering both memory- and model-based approaches, and point-wise and pair-wise approaches. To evaluate its effectiveness in dif-

[1] https://tinyurl.com/yc6nnx5v

ferent domains, we consider movie and song recommendation as application scenarios. Based on our outcomes, we highlight possible research paths that might emerge from it (Section 7.7).

7.2 Related Work

MORSs. Recent literature has studied how to account for multi-objective goals from different angles. The user perspective was tackled by Li *et al* [115], which balance recommendation accuracy for users with different levels of activities. From an item perspective, Ge *et al.* [72] proposed an approach to balance item relevance and exposure. Considering both the user and item perspectives, Naghiaei *et al.* [134] propose a re-ranking approach to account for consumer and provider fairness. Other studies blend the multiple objectives into a single function, in order to obtain a Pareto-optimal solution [119, 179]. Recent advances have also proposed MORSs in sequential settings, by optimizing the results for accuracy, diversity, and novelty [169]. MORS that operate at the individual level have optimized the recommendation process mainly via online interactions, such as conversational approaches [112] or via critiquing [178, 57], but approaches aiming at learning individual propensities from past interactions also exist, e.g., [101, 141, 46, 47].

Calibrated recommendation. *Calibration* is a well-studied technique commonly used to solve the problem of unfair output [137, 190, 170] in Recommender Systems. Seymen *et al.* [165] address the problem of providing calibration in the recommendations from a constrained optimization perspective. Abdollahpouri *et al.* [4] study the connection between popularity bias, calibration, and consumer fairness in recommendation. Recently, Rojas *et al.* [155] analyze how the calibration method in [170] deals with the bias in different recommendation models. Other studies focus on analyzing user profiles to mitigate miscalibrated recommendations [118] or to mitigate popularity bias from the user's perspective [35]. Existing metrics have some limitations when applying a user-centered approach to evaluate popularity bias and calibrated recommendations. To address these limitations, Abdollahpouri *et al.* [5] present a new metric.

Provider fairness. *Provider fairness* [53] has been studied in many common scenarios, e.g., [128, 81, 51, 51, 75, 64, 130]. It is usually assessed by considering metrics such as the visibility and the exposure that respectively assess the amount of times an item is present in the rankings [59, 192] and *where* an item is ranked [15, 191], for users to whom each provider's items are recommended. Other approaches, such as that by Karakolis *et al.* [104], consider diversity and coverage for users. Raj *et al.* [145] present a comparative analysis among several fairness metrics recently introduced to measure fair ranking. Wu *et al.* [180] formalize a family of exposure fairness metrics that model the problem of fairness jointly from the perspective of both types of stakeholders.

Contextualizing our work. No MORS can address both calibrated recommendation lists for the users and provider fairness. Our algorithm's aims are to provide i) each user with calibrated recommendations, ii) fair recommendations for the providers, iii) aiming at a minimum loss in effectiveness.

7.3 Preliminaries

7.3.1 Recommendation Scenario

Let $U = \{u_1, u_2, ..., u_n\}$ be a set of users, $I = \{i_1, i_2, ..., i_j\}$ be a set of items, and V be a totally ordered set of values that can be used to express a preference together with a special symbol $\perp$. The set of ratings results from a map $r : U \times I \rightarrow V$, where V is the rating domain. If $r(u, i) = \perp$, then we say that u did not rate i. To easy notation, we denote $r(u, i)$ by r_{ui}. We define the set of ratings as $R = \{(u, i, r_{ui}) : u \in U, i \in I, r_{ui} \neq \perp\}$ and they can directly feed an algorithm in the form of triplets (point-wise approaches) or shape user-item observations (pair-wise approaches). We denote with R_u the ratings associated with a user $u \in U$. We consider a temporal split of the data, where a fixed percentage of the ratings of the users (ordered by timestamp) goes to the training and the rest goes to the test set [12]. The goal is to learn a function f that estimates the relevance ($\hat{r}_{ui}$) of the user-item pairs that do not appear in the training data (i.e., $r_{ui} = \perp$). We denote as $\hat{R}$ the set of recommendations.

Let C denote the set of geographic continents in which items are organized. We consider a geographic continent as the provenance of an item provider. We denote as C_i the set of geographic continents associated with an item i. Note that, since an item could be produced by more than one provider, it might be associated with several geographic continents, and thus, $|C_i| \geq 1$ and $C_i \subseteq C$. In case two providers belong to the same geographic continent, that continent appears only once; indeed, we are dealing with group fairness so, when a group of providers is associated with an item (once or multiple times), we account for its presence. We use the geographic continents to shape demographic groups, which can be defined to group the ratings of the items produced in a continent (we denote the items in I produced in a continent $c \in C$ as I_c, where $I_c \subseteq I$).

Let G denote the set of genres in which items are organized. We denote as G_i the set of genres associated with an item i. Note that, an item can be of one or more genres, and thus, $|G_i| \geq 1$ and $G_i \subseteq G$. We denote the items in I that have a genre $g \in G$ as I_g, where $I_g \subseteq I$.

7.3.2 Metrics

Provider-group Representation. In order to enable provider fairness, we should understand the attention received by a provider group in the training data. For this reason, we compute the representation of a demographic group in the data as the number of ratings for items associated with that group in the data. We define with $\mathcal{R}$ the *representation* of a group $c \in C$ as follows:

$$\mathcal{R}_c = \frac{|\{r_{ui} \,:\, u \in U,\, i \in I_c\}|}{|R|} \tag{7.1}$$

Eq. (7.1) accounts for the proportion of ratings given to the items of a demographic group associated with a continent. This metric ranges between 0 and 1. We compute the representation of a group only considering the training set. Trivially, the sum of the representations of all groups is equal to 1.

User-based genre propensity. In order to calibrate the results for the users, we need to understand how the preferences for the different item genres are distributed. For this reason, we define with $\mathcal{P}$ the *propensity* of a user of $u \in U$ to rate items of a genre $g \in G$, as follows:

$$\mathcal{P}_{ug} = \frac{|\{r_{ui} \,:\, g \in G_i\}|}{|R_u|} \tag{7.2}$$

Eq. (7.2) accounts for the proportion of ratings associated with a genre for a given user. This metric ranges between 0 and 1. Trivially, the sum of the propensities of all genres for a user is equal to 1. This metric is equivalent to the distribution $p(g|u)$ [170].

Disparate Impact. We assess unfairness with the notion of *disparate impact* generated by a recommender system. Specifically, we assess disparate visibility.

Definition 7.3.9 (Disparate visibility)**.** *Given a group $c \in C$, the* disparate visibility *returned by a recommender system for that group is measured as the difference between the share of recommendations for items of that group and the representation of that group in the input data:*

$$\Delta\mathcal{V}_c = \left(\frac{1}{|U|}\sum_{u \in U}\frac{|\{\hat{r}_{ui} : i \in I_c\}|}{|\hat{R}|}\right) - \mathcal{R}_c \tag{7.3}$$

The range of values for this score is $[-\mathcal{R}_c, 1 - \mathcal{R}_c]$; specifically, it is 0 when the recommender system has no disparate visibility, while negative/positive values indicate that the group received a share of recommendations that is lower/higher than its representation. This metric is based on that defined by Fabbri *et al.* [59].

Miscalibration. We assess the tendency of a system to recommend a user items whose genres are distributed differently from those they prefer via *miscalibration.*

Definition 7.3.10 (Miscalibration)**.** *Given a user $u \in U$ and a genre $g \in G$, the miscalibration returned by a recommender system for that user is measured as the difference between the share of recommendations for items of that genre and the propensity of the user for that genre in the training data:*

$$\Delta\mathcal{M}_{ug} = \frac{|\{\hat{r}_{ui} : i \in I_g\}|}{|\hat{R}_u|} - \mathcal{P}_{ug} \tag{7.4}$$

Generalizability. The rest of our paper will consider disparate visibility ($\Delta\mathcal{V}_c$) as the *global* perspective and miscalibration ($\Delta\mathcal{M}_{ug}$) as the *individual* perspective our MORS considers.

Nevertheless, our approach can be generalized to *any* metric that assesses the difference between (i) the distribution of the recommendations and (ii) what can be observed in the training set or an objective set by the platform via a policy (e.g., a given amount of content novelty or diversity).

7.4 Matching Item Providers and Genre Propensity

7.4.1 Real-world Datasets

First, we extended the MovieLens-1M dataset, so as to integrate the continent of production of each movie. Second, a domain that fits our problem is song recommendation. However, existing music datasets, such as LastFM-2B [131], do not contain song genres and sensitive attributes of the artists, so they do not fit our problem. Thus, we collected a dataset from an online music platform.

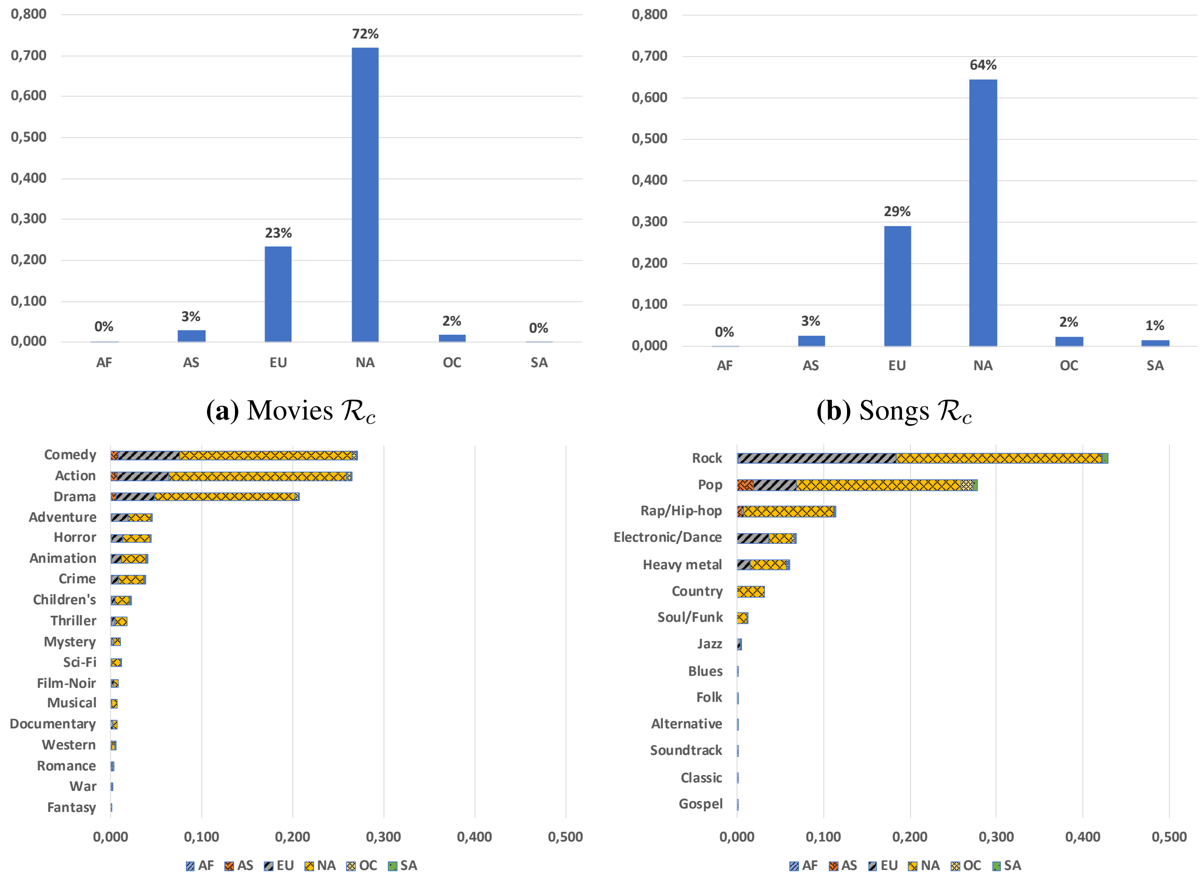
In particular, the **MovieLens-1M (Movies)** dataset comprises 1M ratings (range 1-5), from 6,040 users for 3,600 movies across 18 genres. The dataset provides its IMDB ID, which allowed us to associate it to its continent of production, thanks to the OMDB APIs (`http://www.omdbapi.com/`). Keep in mind that *a movie may be produced on more than one continent*. On the other hand, **BeyondSongs (Songs)** contains 1,777,981 ratings (range 1-5), provided by 30,759 users, to 16,380 songs. For each song, we collected the continent of provenance of the artist, and 14 music genres. Both resources are available online[1].

7.4.2 Characterizing Group Representation and Genre Propensity

We consider the temporal split of the data, where 80% of the ratings are considered for the training set and have been used to measure $\mathcal{R}_c$ and $\mathcal{P}_{ug}$. Note that, while the representation of a demographic group covers the entire training set, the propensity is measured at the user level. Hence, to characterize the link between the two phenomena we aggregate the propensity of all the users for a given genre by summing their values.

Figures 7.1a and 7.1b show the $\mathcal{R}_c$ for Movies and Songs, respectively. Both datasets depict a similar representation by continents, where the highest representation is of items from NA providers (72% in movies and 64% in songs) and the second place is for EU providers (23% and 29%). In the rest of the continents, for both datasets, it is less than 10%. Figures 7.1c and 7.1d show the $\mathcal{P}_{ug}$ in both datasets; three genres attract most of the ratings by users.

We can also observe that ratings seem to be clustered between certain genre-continent pairs. In other words, different genres are distributed differently across continents. In the Movies data (Fig. 7.1c), Comedy movies are largely preferred when produced by EU producers, just as Action attracted the majority of ratings for movies by NA producers. In the Songs data (Fig. 7.1d), the Electronic/Dance genre was consumed much more heavily when produced by EU artists than by those in the rest of the world, and Heavy metal songs are mostly consumed



(a) Movies $\mathcal{R}_c$

(b) Songs $\mathcal{R}_c$

(c) Movies $\mathcal{P}_{ug}$

(d) Songs $\mathcal{P}_{ug}$

Fig. 7.1 Group representation (a and b) and genre propensity (c and d) in the Movies and Songs data. Acronyms stand for AF: Africa, AS: Asia, EU: Europe, NA: North America, OC: Oceania, SA: South America.

when they come from NA. In both datasets, users' preferences for the minority provider groups (AF, AS, OC, and SA) are also concentrated on a few selected genres, confirming this rating aggregation in certain genre-continent pairs.

> **Observation 1**. *Users have the propensity to rate items of certain genres and that are produced by certain geographic groups (i.e., in certain continents). Calibration and provider fairness are related problems so, when producing recommendation lists, both perspectives should be accounted at the same time, in MORS fashion.*

7.5 Individually Calibrated and P-Fair Recommendation

7.5.1 Algorithm

MOREGIN adjusts the recommendations according to the continent of the providers and the representation of each demographic group and seeks to make a calibration at the individual

```
Input: $recList$: ranked list (records contain $user$, $item$, $rating$, $position$, $genre$ ,
       $continent$), which arrives sorted by user and rating and contains $topn$
       recommendations to the user.
       $trainList$: list with the training set (records contain $user$, $item$, $rating$, $genre$,
       $continent$), which is sorted by user and rating.
       $topk$: top $k$ recommendations, we set up $k = 10$.
       $topn$: top $n$ recommendations, we set up $n$= 1000.
Output: $reRankedList$: ranked list with Individually Calibrated and P-Fair Recommendation.
define MOReGIn ($recList$, $trainList$, $topk$, $topn$)
begin
    // Step 1.  Compute $R_c$
    $recBucketRep \leftarrow$ computeRepresentation($topk$, $recList$, $trainList$);
    // Step 2.  Compute $\mathcal{P}_{ug}$
    $recBucketUserProp \leftarrow$ computePropensity($topk$, $recList$, $trainList$);
    // Step 3.  Create a bucket list
    $joinBucket \leftarrow recList + recBucketRep + recBucketUserProp$;
    $joinBucket \leftarrow$ sort(joinBucket);
    // Step 4.  Perform selection of items with three phases
    $userCounts$, $userGenCounts$ , $contCounts \leftarrow \emptyset$;
    $joinBucket \leftarrow$ selectWithHardConstraints($joinBucket$, $recBucketRep$,
      $recBucketUserProp$, $userCounts$, $userGenCounts$, $contCounts$) ; // Phase 1
    $joinBucket \leftarrow$ selectWithSoftConstraints($joinBucket$, $recBucketRep$,
      $recBucketUserProp$, 2, $userCounts$, $userGenCounts$, $contCounts$) ; // Phase 2
    $joinBucket \leftarrow$ selectWithSoftConstraints($joinBucket$, $recBucketRep$,
      $recBucketUserProp$, 3, $userCounts$, $userGenCounts$, $contCounts$) ; // Phase 3
    $reRankedList \leftarrow$ chooseSelectedItems(joinBucket);
    $reRankedList \leftarrow$ sort(reRankedList) ; // sort by user and rating
    return $reRankedList$;
end
```

Algorithm 8: Pseudocode of MOReGIn algorithm

level, following the propensity of each user to rate items of a given genre. Formally, MOReGIn (see Algorithm 8) works following four main steps. **Steps 1 and 2** are devoted to compute $\mathcal{R}_c$ and $\mathcal{P}_{ug}$, considering the ratings in the training set. **Step 3** computes the items that were predicted as relevant for a user by the recommender system and creates a bucket list, $joinBucket$, considering each continent-genre pair, which will store the predicted items. Each bucket comes with two attributes: $\mathcal{R}_c$ and $\mathcal{P}_{ug}$. Specifically, the recommender system returns a list of top-n recommendations (where n is much larger than the cut-off value k, so as to be able to perform a re-ranking). Our starting point to fill a bucket is the relevance predicted for a user u and an item i, $\hat{r}_{ui}$. That item will be stored in the buckets associated with each genre $g \in G_i$ and each continent $c \in C_i$ (even though an item may appear in more than one bucket, it can only be recommended only once). Each element in the bucket is a record that contains the item ID and the $\hat{r}_{ui}$. We sort each bucket considering three values. We sort out $\mathcal{R}_c$ and $\mathcal{P}_{ug}$, in ascending order to ensure the inclusion in the recommendation lists of items from genres and continents that are less represented in the dataset, and we sort in descending order by rating to enhance those products that are relevant to the user. Finally, **Step 4** performs a three-phase re-ranking based

```
define **selectWithHardConstraints**($joinBucket$, $recBucketRep$, $recBucketUserProp$,
  $userCounts$, $uGenCounts$, $contCounts$) **begin**
    $expectedRecordsCont \leftarrow getExpectedRecordsCont(recBucketRep)$;
    $expectedRecUserGen \leftarrow getRecordsUserGen(recBucketUserProp)$;
    **foreach** $rec \in joinBucket$ **do** // for each record
        $userGen \leftarrow rec.user + " - " + rec.genre$;
        **if** $userGen \in expectedRecUserGen$ and $rec.cont \in expectedRecordsCont$ **then**
            $userCounts[rec.user] \leftarrow userCounts[rec.user] + 1$;
            $uGenCounts[rec.userGen] \leftarrow uGenCounts[rec.userGen] + 1$;
            $contCounts[rec.cont] \leftarrow contCounts[rec.cont] + 1$;
            **if** *expectedRecUserGen[rec.userGen]* $\geq$ *userGenCounts and*
              *expectedRecordsCont[rec.cont]* $\geq$ *contCounts and topk* $\geq$ *userCounts[rec.user]*
              **then**
                rec.phase $\leftarrow$ 1; // selects element in phase 1
                $joinBucket.update(rec)$; // updates the element
            **end**
        **end**
    **end**
    **return** joinBucket
**end**
define **selectWithSoftConstraints**($joinBucket$, $recBucketRep$, $recBucketUserProp$,
  $phaseMOReGIn$, $userCounts$, $userGenCounts$, $contCounts$) **begin**
    $expectedRecordsCont \leftarrow getExpectedRecordsCont(recBucketRep)$;
    **foreach** $rec \in joinBucket$ **do** // for each record
        **if** $rec.cont \in expectedRecordsCont$ **then**
            **if** $phaseMOReGIn == 2$ **then**
                **if** *expectedRecordsCont[rec.cont]* $\geq$ *contCounts and topk* $\geq$
                  *userCounts[rec.user]* **then**
                    $userCounts[rec.user] \leftarrow userCounts[rec.user] + 1$;
                    $contCounts[rec.cont] \leftarrow contCounts[rec.cont] + 1$;
                    rec.phase $\leftarrow$ 2; // selects element in phase 2
                    $joinBucket.update(rec)$; // updates the element
                **end**
            **end**
            **if** $phaseMOReGIn == 3$ **then**
                **if** *topk* $\geq$ *userCounts[rec.user]* **then**
                    $contCounts[rec.cont] \leftarrow contCounts[rec.cont] + 1$;
                    rec.phase $\leftarrow$ 3; // selects element in phase 3
                    $joinBucket.update(rec)$; // updates the element
                **end**
            **end**
        **end**
    **end**
    **return** joinBucket
**end**
```

Algorithm 9: Selection methods for the MOReGIn algorithm

on the generated bucket lists. Phase 1 is where we begin, and subsequent phases occur until the top-k is complete. In detail, **Phase 1** selects items starting from the least represented continents to the most represented ones in their corresponding buckets. The algorithm selects items with these conditions: (1) the percentage of items in the recommendation list for a continent is lower or equal to the representation of the continent ($\mathcal{R}_c$); (2) the percentage of items of a given genre in the top-k is lower or equal than $\mathcal{P}_{ug} \cdot k$; and (3) the number of recommended items so far is lower than k. **Phase 2** relaxes the restrictions of phase 1 and here condition 2 is not applied. **Phase 3** selects the items that have the greater relevance for the user, until we complete the top-k. That is, conditions 1 and 2 are not considered.

7.6 Experimental Evaluation

7.6.1 Experimental Methodology

In this work, we focus on well-known state-of-the-art Collaborative Filtering algorithms: **ItemKNN** [159], **UserKNN** [92], **BPRMF** [149], **SVDpp** [107], and **NeuMF** [90]). We will report the results of the original recommendation algorithm (denoted as **OR**). We also consider two comparison baselines: (i) a greedy calibration algorithm [170] (denoted as **CL**) with a λ value of 0.99 (setup defined in [170]), which post-processes the recommendation lists generated by traditional Recommender Systems; and (ii) a provider fairness algorithm [80] (denoted as **PF**) that considers the providers' continent provenance as a sensitive attribute, with a re-ranking approach that regulates the share of recommendations given to the items produced in a continent (visibility) and the positions in which items are ranked in the recommendation list (exposure).

To run the recommendation models, we used the *Elliot* framework [7], which generated the recommendations for each user that fed the input of MOREGIN. As noted in Section 7.4.2, the dataset was divided into two sets, one for training (80%) and the other for testing with the most recent ratings of each user (20%).

For each user, we generated the top-1000 recommendations (denoted in the paper as the top-n; we remind the reader that these $n = 1000$ results are not shown to the users, they are only used internally by our algorithm) to then re-rank the top-k (set up to 10) through the proposed MOREGIN algorithm. We performed a grid search of the hyper-parameters for each model in the two datasets. For ItemKNN and UserKNN, in both datasets, we use 50 *neighbors*, a cosine *similarity*, and the classical *implementation*. For BPRMF, SVDpp, and NeuMf we defined 10 *epochs* and 10 *factors* on each dataset, except NeuMF in Movies that uses 12 *factors*. The *batch size* is 512 for SVDpp and NeuMF and is 1 for BPRMF on both datasets. Moreover, for BPRMF in Movies$\sim$Songs, *learning rate*=0.1$\sim$1.346, *bias regularization*=0$\sim$1.236, *user regularization*=0.01$\sim$1.575, *positive item regularization*=0.01$\sim$1.376, and *negative item regularization*=0.01$\sim$1.624; for SVDpp in Movies$\sim$Songs, *learning rate*=0.01$\sim$0.001,*factors regularization*=0.1$\sim$0.001, and *bias regularization*=0.001 in both datasets; NeuMF in Movies$\sim$Songs, the *multi-layer perceptron*=10 in both, *learning rate*=0.0025 in both, and *factors regularization*=0.1$\sim$0.001.

7.6.2 Assessment of Disparities and Mitigation

Table 7.1 Results of disparity mitigation of continents in the Movies and Songs datasets. Each value represents the sum of disparities, $\Delta Total$.

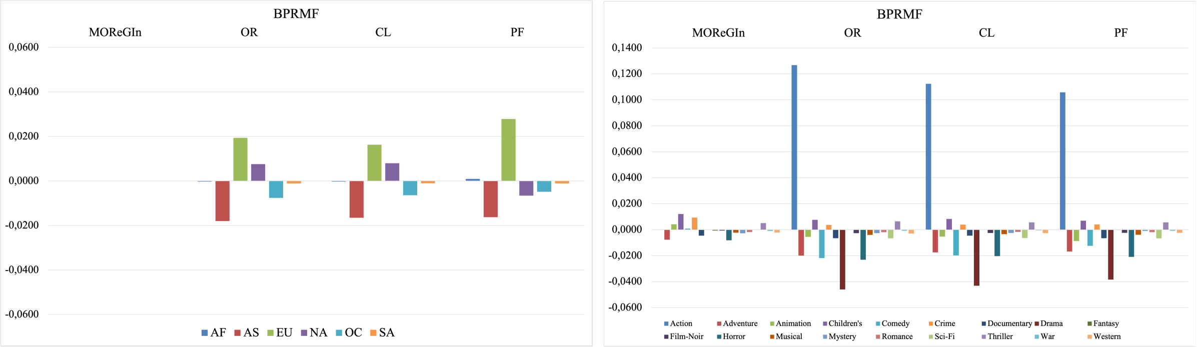
	MOVIES				SONGS			
	OR	**CL**	**PF**	**MOReGIn**	**OR**	**CL**	**PF**	**MOReGIn**
BPRMF	0.0539	0.0485	0.0576	**0.0000**	0.2637	0.0840	0.2628	**0.0000**
SVDpp	0.1154	0.1085	0.1059	**0.0000**	0.2678	0.1063	0.2445	**0.0000**
NeuMF	0.0395	0.0421	0.0638	**0.0000**	0.4434	0.4516	0.3990	**0.0000**
UserKNN	0.0345	0.0327	0.0328	**0.0000**	0.0361	0.0575	0.0370	**0.0000**
ItemKNN	0.0431	0.0418	0.0412	**0.0000**	0.0392	0.0583	0.0420	**0.0000**

Table 7.2 Results of miscalibration of genres in the Movies and Songs datasets. Each value represents the sum of miscalibrations, $\Delta Genre$.

	MOVIES				SONGS			
	OR	**CL**	**PF**	**MOReGIn**	**OR**	**CL**	**PF**	**MOReGIn**
BPRMF	0.2892	0.2606	0.2454	**0.0634**	5.5107	0.0772	0.4610	**0.0289**
SVDpp	0.5792	0.5026	0.5694	**0.1184**	0.5773	0.1031	0.5029	**0.0256**
NeuMF	0.4596	0.3962	0.3735	**0.2901**	1.2886	0.7494	1.2202	**0.0787**
UserKNN	0.0743	0.0862	0.0580	**0.0392**	0.0298	0.0989	0.0291	**0.0208**
ItemKNN	0.2102	0.1966	0.1954	**0.0559**	0.0890	0.0601	0.0879	**0.0205**

Table 7.1 compares MOReGIn with the baselines in terms of the overall disparate visibility, $\Delta Total$, for each continent. It is computed as $\forall c \in C$, $\Delta Total = \sum \Delta \mathcal{V}_c$. MOReGIn almost entirely reduces the disparities in both movies and song datasets, where most results are $\Delta Total = 0.0000$. Although there is a little difference in the $\Delta Total$ between some approaches, these differences are more explicit, considering the provider provenance. For example, in the movie domain with the BPRMF algorithm, the $\Delta Total$ value in the OR approach is similar to that of PF. However, in a more detailed analysis of more representative continents such as NA and EU, there are notorious differences between the two approaches (i.e., 0.0075 for OR in contrast to -0.0066 for PF in the NA continent, see the example shown in Figure 7.2a). It is important to highlight that in both domains, our proposal mitigates the disparity regardless of the provenance of the provider, in contrast to the other algorithms that show a clear dependence on the data (i.e., the continent attribute).

Regarding the item genres, Table 7.2 compares MOReGIn with the baselines in terms of the overall miscalibration, $\Delta Genre$, for each continent. It is computed as $\forall g \in G$ and each user $u \in U$, $\Delta Genre = \sum \Delta \mathcal{M}_{ug}$. For both datasets, MOReGIn obtained the best $\Delta Genre$ (i.e., lowest miscalibration) in all the recommendation models. An analysis of how the algorithms behave with the different genres is shown in Fig. 7.2b. Although miscalibration never reaches values of $\Delta Genre$ equal to zero, our proposal always calibrates better than the baselines.



(a) Movies $\Delta\mathcal{V}_c$

(b) Movies $\Delta\mathcal{M}_{ug}$

Fig. 7.2 Disparity mitigation per continent (a) and miscalibration per genre (b) in BPRMF.

> **Observation 2**. MOReGIn, *by taking action on the distribution of the items per genre at the user level and provider provenance at the same time, can both calibrate and be fair to the providers. This joint effort allows us to improve the capability to calibrate the results and to be fair to providers with respect to baselines devoted solely to these purposes.*

Table 7.3 NDCG for each approach and recommendation algorithm.

	MOVIES				SONGS			
	OR	**CL**	**PF**	**MOReGIn**	**OR**	**CL**	**PF**	**MOReGIn**
BPRMF	**0.3204**	0.3144	0.3195	0.3057	0.0034	**0.0067**	0.0031	0.0055
SVDpp	0.0830	0.0888	0.0812	**0.1024**	0.0050	0.0103	0.0051	**0.0138**
NeuMF	0.1963	0.1931	0.1956	**0.2050**	0.0183	0.0098	0.0179	**0.0314**
UserKNN	0.3051	0.2954	0.3030	**0.3053**	**0.3760**	0.1925	0.3759	0.2648
ItemKNN	**0.3229**	0.3145	0.3211	0.3131	**0.3860**	0.1668	0.3857	0.2864

7.6.3 Impact on the Quality of Recommendations

We evaluate the accuracy for the different approaches via the NDCG metric.

Table 7.3 shows its values for MOReGIn and the rest of the baselines, in all the recommendation algorithms, for Movies and Songs. MOReGIn obtained a better NDCG than the PF model, except for BPRMF and ItemKNN in the Movies dataset, and UserKNN and ItemKNN in the Songs dataset. Similar results are obtained with the CL method. Comparing MOReGIn to a non-fair approach, MOReGIn outperforms OR models, with the exception of BPRMF and ItemKNN in the Movie domain. Except for UserKNN and ItemKNN, MOReGIn also outperforms the OR model in the Songs domain.

All recommendation quality results show that the need for fairer and calibrated recommendations impacts the recommendation quality. However, beyond-accuracy perspectives, such as those offered by MOReGIn allows for compensating for the minimal loss in quality with more unbiased recommendations.

7.7 Conclusions and Future Work

Global and individual objectives in MORs have never been studied jointly. To study this problem, we provided data, by i) extending the MovieLens-1M dataset and ii) collecting a new dataset for song recommendation. The analysis of this data showed that when users rate items of a given genre, the geographic provenance of that item matters. Based on these insights, we proposed a new post-processing approach, named MOREGIN, that aggregates the recommended items into buckets, pairing item genres and their continent of production. Results show that MOREGIN outperforms the existing approaches at producing effective, calibrated, and provider-fair recommendations. Future work will explore different strategies to generate recommendation lists given the generated buckets. Moreover, we will consider consumer fairness as a global perspective.

CHAPTER 8

AMBAR: A dataset for Assessing Multiple Beyond-Accuracy Recommenders

This chapter contains the paper entitled ”AMBAR: A dataset for Assessing Multiple Beyond-Accuracy Recommenders”, which has been accepted and will be presented at the 18th ACM Conference on Recommender Systems (RecSys ’24), Bari, Italy, October 14–18, 2024. We introduce a new dataset in the music domain that includes various sensitive attributes with multiple levels of granularity from different perspectives: user, item, and subject.

- Elizabeth Gómez, David Contreras, Ludovico Boratto, Maria Salamó. (in press). **AMBAR: A dataset for Assessing Multiple Beyond-Accuracy Recommenders**. In: 18th ACM Conference on Recommender Systems. RecSys ’24. Bari, Italy, October 14–18, 2024. https://doi.org/10.1145/3640457.3688067 — **Rank: B in CORE**.

AMBAR: A dataset for Assessing Multiple Beyond-Accuracy Recommender

Nowadays a recommendation model should exploit additional information from both the user and item perspectives, in addition to utilizing user-item interaction data. Datasets are central in offering the required information for evaluating new models or algorithms. Although there are many datasets in the literature with user and item properties, there are several issues not covered yet: *(i)* it is difficult to perform cross-analysis of properties at user and item level as they are not related in most cases; and *(ii)* on top of that, in many occasions datasets do not allow analysis at different granularity levels. In this paper, we propose a new dataset in the music domain, named *AMBAR*, that tackles the above-mentioned issues. Besides detailing in depth the structure of the new dataset, we also show its application in contexts (i.e., multi-objective, fair, and calibrated recommendations) where both the effectiveness and the beyond-accuracy perspectives of recommendation are assessed.

Keywords: Dataset, Fairness, Music Recommendation, Music Information Retrieval.

8.1 Introduction

Many studies have pointed out that aspects beyond accuracy –such as the *diversity*, *fairness*, or *novelty* of the recommended items– are as important as accuracy in making a satisfactory recommendation [129, 87, 70]. Although *beyond-accuracy* properties have attracted much research in the literature in the last decade, there is still a long way to go. For example, the cross-analysis of multiple beyond-accuracy properties, or the analysis of models at different granularities of a property. The absence of suitable data sets for study is one of the main obstacles to scientific advances.

In the field of Recommender Systems, there are well-known datasets, such as MovieLens [88], Book-Crossing [201], or The Million Song [13] datasets, among other datasets (detailed in Section 8.2). Although these datasets provide beyond-accuracy properties, as far we know, some issues persist; no dataset available allows for embracing beyond-accuracy properties –from both the user and item side– that are related in both sides and with different levels of granularity of these properties.

In this paper, inspired by the need to evaluate new models and algorithms from the different stakeholders' points of view and with the idea of providing recommendations that offer beyond-accuracy properties, we propose a novel dataset in the music domain, named *AMBAR*. It offers both user-item interactions and additional information on users, items, and subjects (e.g., users by considering sensitive attributes such as their gender or geographic provenance and, items and subjects by considering the category styles or the providers' sensitive attributes such as their gender or geographic provenance). Another novelty of this dataset is that it enables to study systems that offer fair recommendations by both binary and multi-class strategies with different granularities. AMBAR is a real dataset that was extracted to contain reliable and precise data from a popular music platform. For the sake of protecting user privacy and the platform's business, the data has been anonymized. The dataset is available for the community and it has been released at the following **Link**[1]. The ultimate goal of AMBAR is to provide a common source of data for benchmarking algorithms that cover equity from the user and item side.

This dataset makes it possible to conduct new analyses, such as multi-objective recommendation [97], fair [22, 23], calibrated recommendation [170], or a cross-analysis of properties at user and item level. We show the application of the proposed dataset in several contexts (consumer fairness [135], provider fairness [81], consumer-provider fairness [135], and a cross-analysis with calibration and multi-objective [82] recommendation).

In summary, as far as we know, this is the first dataset that provides several sensitive attributes –with different levels of granularity– from several perspectives: the user, the item, and the subject side.

The paper is structured as follows. First, in Section 8.2, we briefly present an overview of

[1] https://github.com/davidcontrerasaguilar/AMBAR

beyond-accuracy perspectives of Recommender Systems and detail the related datasets. Then, in Section 8.3, we provide information about the acquisition of the dataset and its content, as well as some statistical analysis and possible uses of the dataset. In Section 8.4, we define several tasks and analyze the result of different fairness and non fairness-aware algorithms on the proposed dataset. Finally, in Section 8.5, we summarize the work and outline possibilities of further exploiting the dataset.

8.2 Related Work

Overview of Beyond-Accuracy Perspectives of Recommender Systems There is currently a growing interest in equity and non-discrimination in RSs. Several studies seek to mitigate popularity bias [105, 122, 34], unfairness [22, 23, 77, 79, 81] and miscalibrated recommendations [170]. In Abdollahpouri *et al.* [4], the authors show a connection between how different user groups are affected by popularity *bias* and how it leads to poorly *calibrated* recommendations. Steck [170] analyzes the problem when the recommendations produced by a model differ from the users' play history and proposes a *calibration* algorithm. In recent research, Seyment *et al.* [165] propose an optimization model that combines both accuracy and calibration. Some previous studies related to the concern of *fairness* in RSs [40, 41] distinguish between *consumer* fairness, *provider* fairness, and *subject* fairness [54]. The concept of *consumer* fairness regards how Recommender Systems may particularly affect those who receive the recommendations [23]. On the other hand, *provider* fairness relates to the impact of the generated recommendations on the item providers. It guarantees that the providers of the recommended objects that belong to different groups are similar at the individual level and will get recommended according to their representation in the data [79, 80]. Finally, *subject* fairness refers to who receives assignment in the allocation process as fairness subject, which may be items, users, or both (item-user) [54]. A detailed introduction concerning fairness in RSs is described in [142, 56, 40].

Related Datasets The literature has a large number of datasets, however not all of them have sensitive features like COCO [43]. In our opinion, sensitive features are crucial to assess *beyond-accuracy* perspectives of recommendations. MovieLens [88] is likely the most well-known RSs dataset. It offers information about users' preferences for movies represented by ratings and it is available in several sizes of users' ratings from 100K to 20M. Another one is the Book-Crossing [201] dataset. Both include the category of items (i.e., the genre of books and movies). However, geographic information of providers or attributes for consumer fairness are unavailable.

Datasets associated with music information retrieval (MIR) are also popular. The Million Song Dataset (MSD) [13] contains audio information but, unfortunately, the users' preferences (i.e., users' ratings) are unavailable. Therefore, combining it with other rating-based datasets is required, such as the Yahoo Music Rating Dataset (*http://webscope.sandbox.yahoo.com/*). Another one is the Million Musical Tweet Dataset (MMTD) [89], which contains listening histories inferred from microblogs. Each listening event is composed of temporal (e.g., date and time), spatial (e.g., longitude, latitude, and country), and contextual information. Recently, two datasets have been collected from the well-known *Last.fm* platform [32] with users' information (i.e., gender, age, country, and date of registering in Last.fm). The first one only includes artists that 360 thousand users most frequently listened to. The second one contains around a thousand users with information on the artist, track name, and timestamp for each listening event. In addition, two bigger *Last.fm* datasets have been presented LFM-1b [163] and LFM-2b [131]. Both datasets are an extensive collection of music listening events enriched by users' demographic information (i.e., users' age, country, and gender), music-related metadata (e.g., artist and track names), and timestamps (a specific time when a track was listened to by a user). Although these datasets contain demographic and gender information for users registered in Last.fm (consumers), the demographic and gender information for artists (providers) is unavailable neither the style of music (i.e., category of item). Table 8.1 describes datasets commonly used in existing fairness research. Due to space constraints, the list is limited to the most referred ones in the literature; see [41, 100] for further datasets.

Contextualizing our work Any of the datasets described previously in the music domain contain entire attributes for analyzing *consumer*, *provider*, or *subject* fairness. Concretely, our dataset offers the following features: *(i)* users' preferences in a rating Likert scale based on the listening events; *(ii)* demographic users' information in an anonymous way (e.g., sensitive attributes of gender and geographic provenance); *(iii)* information of artists, including demographic attributes (e.g., gender and geographic provenance) and music category (i.e., the genre or style of music); And *(iv)* information of tracks, including sensitive attributes such as the styles of music. While music style may not be typically viewed as a sensitive attribute, it can still play a role in bias, equity, and fairness for recommendations. This is especially true when considering minority music styles that might face disadvantages in comparison to more popular styles.

8.3 The AMBAR Dataset

8.3.1 Data Collection

It is difficult to gather a dataset with sensitive attributes since most internet platforms only allow the downloading of data, not the sensitive attributes. Because of this, there are multiple stages involved in the data collection process.

Table 8.1 Description of datasets used in existing fairness research in RSs and our AMBAR dataset

Datasets	Fairness-related User Attributes	Fairness-related Item Attributes	Fairness-related Subject Attributes	Fine-grained	Coarse-grained	Interactions	Domain	References
Amazon	gender	categories, gender of model	-	-	users (gender) item (categories, gender of model)	143.6M	Shopping	[51, 111, 39, 66]
Ciao	-	popularity	-	-	item (popularity)	484K	Shopping	[160]
Book-Crossing	-	category	-	-	-	1M	Books	[201]
Ctrip Flight	-	airline	-	-	item (airline)	25.1K	Travels	[182]
Google Local	-	category of business	-	-	item (category of business)	11.4M	Shopping	[139, 182]
Insurance	-	gender, marital status, occupation	-	-	item (gender, marital status, occupation)	5.3K	Insurances	[116]
Last.FM 1K	gender, age	-	-	-	user (gender, age)	904.6K	Music	[50]
Last.FM 360K	gender, age	-	-	-	user (gender, age)	17.5M	Music	[139, 32]
Million Song Dataset	-	-	style of track	-	style of track	1M (tracks no ratings)	Music	[13]
Million Musical Tweets Dataset	country, postal code	-	-	-	user (country, postal code)	1M (tweets no ratings)	Music	[89]
ModCloth	body shape	product size		-	user (body shape) item (product size)	99.8K	Clothes	[177]
Movielens 100K	-	popularity, provider, year of movie	-	-	item (popularity, provider, year of movie)	100K	Movies	[88, 71, 73, 121, 198]
Movielens 1M (Movies)	gender, age, occupation	genres, popularity	-	-	user (gender, age, occupation) item (genres, popularity)	1M	Movies	[88, 71, 96, 116] [124, 181, 200]
Movielens 20M	-	product company, genres	-	-	item (product company, genres)	20M	Movies	[24, 133, 170] [171, 200]
Sushi	gender, age	seafood or not	-	-	user (gender, age) item (seafood or not)	50K	Food	[102]
Xing	premium/standard	membership, education degree, working country	-	-	user (premium/standard) item (membership, education degree, working country)	8.1M	People	[39, 114, 200]
Yelp	-	food genres	-	-	item (food genres)	8.6M	Shopping	[123, 164, 199]
Tenrec	gender, age	category	-	-	user (gender, age) item (category)	140M	Shopping	[189]
AMBAR	gender, country, continent	- tracks (styles, category styles) - artists (gender, country, continent, styles, category styles)	tracks (styles, category styles)	- users (country) - artists (gender, country, styles) - tracks (styles)	- users (gender, continent) - artists (continent, category styles) - tracks (category styles)	3.3M	Music	

Stage 1. Information Acquisition This stage is responsible for obtaining information on music preferences from a popular and well-known entertainment platform[2]. In particular, we download the music information consisting of users, providers, items, and users' music preferences information. First, we select a top providers' list and select for each provider a unique list of users that have interacted with the provider. Second, for each user, we obtain a subset of their top listening histories (i.e., 100 top songs). Later, we search close users for a randomly selected user from the previous unique list. To avoid ending up with a prevalent subset of users or items, this random process is repeated n times. For each user, we obtain a subset of their top

[2] We used the API provided by the platform and had permission to download the non-sensitive data. However, to complete the dataset, we made inferences about several sensitive attributes. It is worth noting that due to privacy and ethical considerations, the platform in question has chosen to remain anonymous as the primary source of information for the dataset. In addition, the platform permitted us to publish the dataset, with the condition of anonymizing the dataset. Despite these limitations, the AMBAR dataset currently includes all the necessary information to perform RSs and a more extensive set of sensitive and non-sensitive attributes compared to data sources available in the literature.

interacted items; for these top items, we download the item and the provider item information.

Stage 2. User's Preferences Generation Commonly, users' music preferences are represented by a rating. However, this information is not available on the platform we used in this study, where only the *playcounts* attribute, which represents the number of times a user has listened to a song, is available. Therefore, we create a user play history dataset that contains the user, artist, track, and play counts. Then, we transformed the implicit feedback embedded in people's listening histories – represented by the number of *playcounts*– into ratings (i.e. a 1-5 Likert scale value), following the approach used in [175]. Concretely, we computed the complementary cumulative distribution of the listening histories per listener, then assigned a score of five to music items (i.e., the track) located in the first quintile, four to the ones in the second quintile, and so forth. It is important to note that this mapping is based on the assumption that people who have listened more to a particular music item prefer it.

Stage 3. Styles of Music Filtering In this stage, two strategies are applied to obtain the style of music of artists and tracks. First, for each track in the users' play histories, we have the associated tags that users have defined on the platform. For example, a tag may be the name of the artist, the name of the track, the artist's geographic provenance, the style of music also named genres (e.g., rock, pop, classical, or other), or other attributes defined by a user. In particular, using a list of music genres obtained from the Music Map web[3], we extracted the genres of music from the list of tags in the track. Then, we define a new genre attribute that contains a list of genres for each track. We obtained more than 280 styles defined by users. Due to the sheer number of styles and to aid in the analysis of music style data, we defined a new (more general) genre attribute category containing a reduced list of categories of music (i.e., 14 categories of music styles), based on the study presented by Rentfrow *et al.* [150]. Second, we also used the Spotify API[4] to complement and validate the list of styles. Concretely, we obtained a list of styles for each artist, and later we used these styles in their corresponding tracks.

Stage 4. Geo Provider/User Filtering First, in this stage, we used the Wikipedia[5] platform to obtain the geographic provenance of the artists (i.e., the providers). In particular, we seek information on the place of birth (i.e., the *Born* tag) in the case of single artists and the *Origin* for music bands. On the other hand, the geographic provenance of users was downloaded from the API provided by the platform. Moreover, users without a country were removed.

Stage 5. Users' Gender To enable an eventual analysis of customer fairness in future research, we also include the gender of users (i.e., female or male) in our dataset. To get the

[3] https://musicmap.info/

[4] https://developer.spotify.com/

[5] http://www.wikipedia.org

gender of users, we used a list of names for male and female genders obtained from platforms like Wikipedia and the Gender API[6]. There are names that can be masculine or feminine depending on the country. We used the origin of the user to disambiguate it as male or female. If the user's country was unknown, the user with an ambiguous name was removed from the dataset. It is important to note that users without real names were also removed and the nickname attribute was not used in this module. Moreover, users with ambiguity in gender inference were eliminated to reduce the noise that could be generated by data inference.

Stage 6. Artists' Gender This stage is responsible for obtaining the genders of artists. In particular, we used the *SPARQLwrapper* python package to query the Wikidata[7] platform using the artists' names. When an artist is a single person, the gender is clearly inferred (male or female), however, in the case of bands, several artists are involved and, in this case, we cannot infer a unique genre and have defined it as undefined in the dataset.

8.3.2 Data Content

The complete dataset occupies 95 MB and is available at the following **Link**[1]. The terms of use of the dataset are for research purposes only and are strictly non-commercial. All files are stored in *comma-separated value* format to ensure better compatibility. Since AMBAR includes sensitive attributes, the dataset has been anonymized to preserve the privacy of any interested parties: *users*, *tracks*, and *artists*. Specifically, AMBAR consists of 3,311,462 ratings, from 31,013 users, for 443,921 songs and 30,667 artists, described in four files.

- *users_info.csv*. This file contains specific users' information depicted by the attributes user_id (i.e., the dataset index), country, continent, and gender.
- *tracks_info.csv*. This file contains information on the music items (i.e., tracks or songs). The attributes in this file are track_id, artist_id, duration, styles (i.e., the style of music), and category_styles (i.e., music style categories).
- *artists_info.csv*. This file contains artists' information. The attributes in this file are artist_id, gender, country, continent, styles (i.e., the style of music), and category_styles (i.e., music style categories).
- *ratings_info.csv*. This file contains the users' music preferences represented by a rating. In particular, the attributes in each tuple are user_id, track_id, and rating.

[6] https://gender-api.com/

[7] https://www.wikidata.org/

Table 8.2 Gender Distributions for Users and Artists

	Users		Artists	
Gender	**Number**	**%**	**Number**	**%**
N/A*	0	0.000%	21253	58.162%
Male	19633	63.306%	6528	17.865%
Female	11380	36.694%	2831	7.747%
Non-binary	0	0.000%	27	0.074%
Trans woman	0	0.000%	14	0.038%
Trans man	0	0.000%	2	0.005%
Genderfluid	0	0.000%	7	0.019%
Genderqueer	0	0.000%	2	0.005%
Bigender	0	0.000%	1	0.003%
Agender	0	0.000%	1	0.003%
Transgender	0	0.000%	1	0.003%
Total sum	**31013**	**100%**	**30667**	**100%**

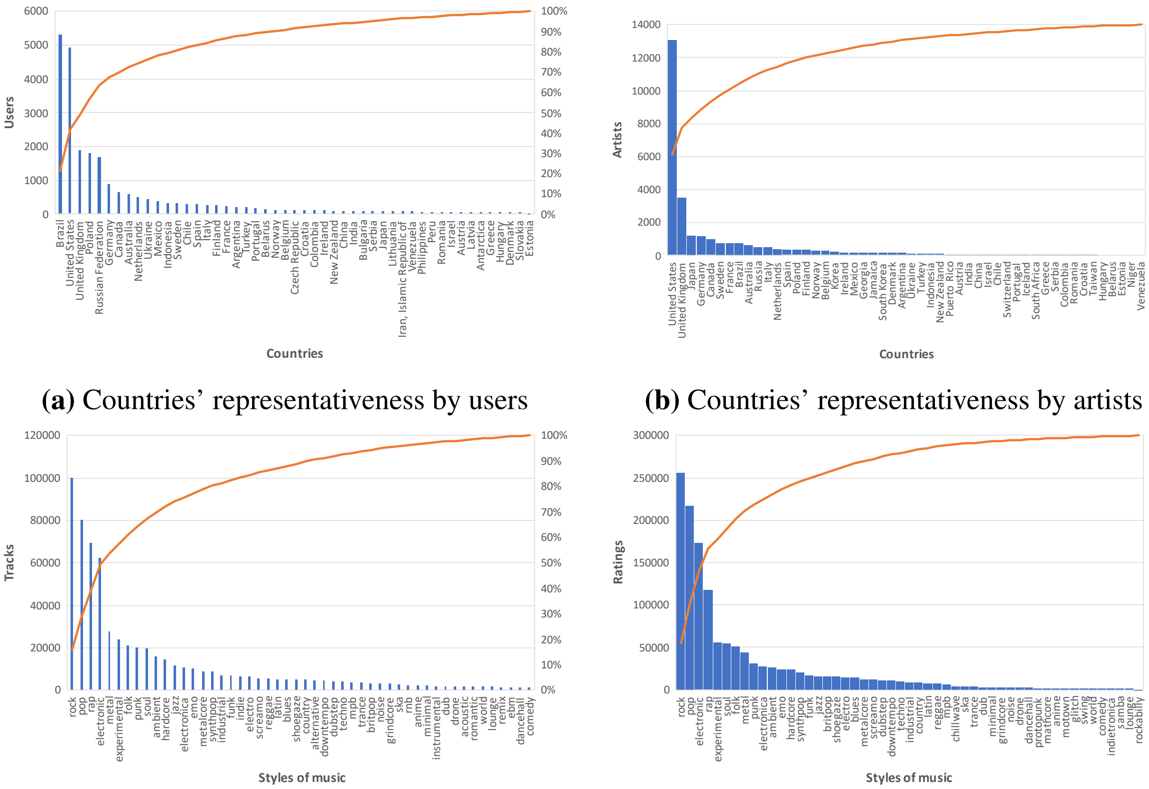
8.3.3 Data Statistics

As previously stated, AMBAR contains sensitive attributes. Nevertheless, as occurs in the majority of real-life situations, there are imbalances in the representation of these attributes. This is a perfect scenario for applying our dataset in fairness or calibrated algorithms to reduce inequalities in making recommendations.

One of the sensitive attributes is the gender of users and artists, whose distribution within the dataset is shown in Table 8.2. The male gender is the more representative one with 63.30% for users and 17.87% in the case of artists. It is important to note that there is 58.16% of artists without gender (i.e., denoted as N/A), which in most cases corresponds to music bands. In addition, the male and female genders represent around 26% of the total number of artists. Note that in users there is no gender variability, other than male and female, while in artists there is much more variability, although with low representation. This is due to the different ways the gender attribute has been collected for users and artists.

Another sensitive attribute available in AMBAR dataset is the geographic provenance of users and artists. The distribution of this attribute is depicted in Figures 8.1a and 8.1b. For a better display, we only show the top_50 countries in plots.

Figure 8.1a shows the representation of each country by users, where the top_4 countries (with around 45% of the total users) are Brazil, the United States, the United Kingdom, and Poland. On the other hand, the representation of each country by artists is depicted in Figure 8.1b with around 62% of artists from the United States, the United Kingdom, Japan, and Germany. In total, the whole dataset includes 134 distinct countries for artists and 200 distinct countries for users. Since the dataset includes a large number of countries, it is useful to analyze fairness with a coarse (i.e., reduced) granularity. For this reason, the dataset also includes the continent of the users and providers. Table 8.3 describes the group representation in AMBAR



(a) Countries' representativeness by users

(b) Countries' representativeness by artists

(c) Styles' representativeness by tracks

(d) Styles' representativeness by ratings

Fig. 8.1 Country and Music styles representation

dataset, being the majority groups North America and Europe for the artists and tracks, and the majority groups are Europe and South America for the users.

Figure 8.1c depicts the distribution of music styles concerning tracks. The most popular style is the rock genre with 14.76% representation, and the *top*_4 styles are rock, pop, rap, and electronic (achieving a total of 46%). Note that the total amount of music styles by tracks in the dataset is 282 and for a better display, we only show the *top*_50 styles in plots. The music styles have been categorized into 14 main categories (i.e., blues, jazz, classical, folk, rock, alternative, heavy metal, country, soundtracks, religious, pop, rap/hip-hop, soul/funk, and electronic/dance).

Finally, Figure 8.1d shows the representation of each style of music considering the number of ratings that it attracted by users. The *top*_4 are shared between rock, pop, electronic, and rap, which account around 54% of the ratings, and the percentage share of styles decreases while going down the ranking.

8.3.4 Dataset Use

The AMBAR dataset can be used for training and evaluating traditional Recommender Systems algorithms, such as Collaborative Filtering [125]. However, our dataset's main strength is its sensitive attribute content (i.e., gender of artists, gender of users, countries of artists, and

Table 8.3 Group Representation of the Users and Artists per continent in AMBAR, in alphabetical order by continent

Continent	Users	Artists	Tracks
Africa	0.0074	0.0075	0.0016
Asia	0.0653	0.0896	0.0421
Europe	0.4156	0.3599	0.3036
North America	0.2391	0.4795	0.6052
Oceania	0.0294	0.0261	0.0231
South America	0.2432	0.0374	0.0243

countries of users, continent of users, continent of artists). Moreover, apart from the wide range of sensitive attributes, it includes different levels of granularity (i.e., music styles, category of styles) at several perspectives: the user, the item, and the subject side. Since the data set comprises more than one sensitive attribute and with different granularities (i.e., fine- and coarse-grained), it opens up a wide range of possibilities for researchers to analyze the data to minimize inequalities: analyze from the consumer side, analyze the provider side, analyze both consumers and providers, analyze several sensitive attributes at the same time, and perform cross-analysis of properties at user and item level.

In summary, it enables new analyses, including multi-objective recommendation [97], fair [22, 23], calibrated recommendation [170], and a cross-analysis of properties at the user and item level, such as MOReGiN [82] that analyzes the interplay between provider fairness and calibration. In Section 8.4, the tests performed demonstrate each of these uses.

8.4 Tasks and Experiments

In this section, we apply four state-of-the-art recommendation algorithms to our dataset and analyze how different fairness algorithms can be applied: **CPFair** [135], **PFair** [79, 80], and **MOReGIn**[82].

8.4.1 Experimental Setup

Note that the AMBAR dataset contains attributes with two types of perspectives: binary and graded, for example, the binary one with the gender and the graded one with the country. As we mentioned earlier, the dataset is available in the following **Link**[1].

8.4.1.1 State-of-the-art Recommendation Algorithms

To analyze the use of our AMBAR in traditional RSs (i.e., using AMBAR with only non-sensitive attributes), we used four well-known algorithms based on latent factors models: **MF**[108], **WMF**[95], **SVD**[106], and **VAECF**[117]. To evaluate the performance of algorithms, we use: **NDCG@k**, **Recall@k**, and **Precision@k**, being $k = 50$.

To run these state-of-the-art models and evaluate them according to the aforementioned, we used the **Cornac** framework [156]. Note that the list of recommendations for each user, generated by Cornac, will be used by the fairness-aware algorithms described in Section 8.4.1.2, 8.4.1.3, and 8.4.1.4. In addition, the dataset was randomly separated into a test data set containing the 20% of ratings of each user and a training data set containing the remaining 80% (i.e., we can not consider a temporal split of the data because the user-item interactions based on transaction timestamp is unavailable in the dataset and also in the source of information of AMBAR).

Each algorithm was run with the following hyper-parameters:

- **MF**: *epochs* = 12, *factors* = 10, *learning rate* = 0.0001, *regularization* = 0.1
- **WMF**: *epochs* = 15, *batch size* = 512, *factors* = 12, *learning rate* = 0.001, *regularization* = 0.1
- **SVD**: *epochs* = 10, *factors* = 10, *learning rate*= 0.0001, *regularization* = 0.1
- **VAECF**: *epochs* = 12, *batch size* = 512, *factors* = 10, *learning rate* = 0.001, *regularization* = 0.1

8.4.1.2 Accounting for Beyond-accuracy Perspectives

To evaluate AMBAR under multiple perspectives that involve the sensitive attributes it offers, we used the sensitive attribute *gender* (male/female) present in users' and artists' information to assess consumer and provider fairness. To do the mitigation of inequalities we applied a new optimization-based re-ranking algorithm [135], named CPFair, that integrates fairness constraints from both the consumer and producer side. CPFair focuses exclusively on the (binary) exposure dimension, leaving the graded option as future work, as denoted in [135]. In these experiments, we consider that users are the consumer side and tracks (i.e., songs) are the producer side. Additionally, recommendations are generated by the following algorithms: MF, WMF, SVD, and VAECF methods, as detailed above.

In particular, we follow a similar methodology to that described in [135]. Concretely, to evaluate our experiments' performance, we randomly divided the AMBAR into two parts: train and test sets, with proportions of 80% and 20%, respectively. First, all models were trained and evaluated on the same datasets. Next, we re-ranked the experiment by classifying users and tracks into two groups each. For constraint on the consumer side, we divided users into male and female groups. For constraint on the producer side, we remove tracks of artists without gender (e.g., musical bands), and next, we divide tracks into two groups by gender of the artist.

Finally, to measure the overall fairness of models concerning producers and consumers, we apply the Consumer-Producer fairness evaluation ($mCPF$) metric, the **Deviation of Provider Fairness** (DPF), and the **Deviation of Consumer Fairness** (DCF). In particular, the fairest model is the one that has the lowest $mCPF$ value. All metrics used in these experiments are defined in [135] and the hyper parameters $\lambda 1$ and $\lambda 2$ were set to 0.000005. Besides this, the CPFair framework allows us to assess other beyond-accuracy perspectives, such as **novelty** (i.e., it measures the capacity of the recommender system to generate novel and unexpected results to suggest objects a user is unlikely to know about already, we use the novelty measure as defined in [197]), and **coverage** (i.e., it measures the percentage of the available items which effectively are ever recommended to a user, specifically, we used the *catalog coverage* measure as defined in [69]).

8.4.1.3 Provider Fairness Algorithm

The second model used to evaluate AMBAR was the one proposed by Gómez *et al.* [79, 81], which assess provider fairness with a multi-class perspective. That is, it works with graded attributes. We follow the approach used in [79, 81], where the countries were grouped into continents (i.e., a coarse-graded approach). This multi-group approach divides the different item providers into groups according to the continent of origin. Concretely, the model seeks to perform a re-ranking to reduce the disparity between the visibility and exposure given to providers concerning the representation of the same in the dataset, seeking a minimum loss in the effectiveness of the recommendations. Concretely, the visibility and the exposure respectively assess the amount of times an item is present in the rankings [59, 192] and *where* an item is ranked [15, 191], for users to whom each provider's items are recommended. The experiments start from a top 1000 recommendations in 4 different models (MF, WMF, SVD, and VAECF). Later, recommendations are redistributed, obtaining a top 10, with visibility and exposure adjusted to the users' preferences and, at the same time, balanced to the geographical origin of the providers according to the training data.

The metrics used in our analysis and experiments are: the *representation of a group*, *disparate visibility*, and *disparate exposure*. All these metrics are defined in [81]. First, the **representation of a group** C ($\mathcal{R}_C$) is the amount of times that group appears in the data, that is the amount of ratings collected for that group. This metric is between 0 and 1. We compute the representation of a group only considering the training set. Trivially, the sum of the representations of all groups is equal to 1, $\sum_{i=1}^{k} \mathcal{R}_{C_i} = 1$. Second, given a group C, the **disparate visibility** ($\Delta\mathcal{V}(C)$) returned by a recommender system for that group is measured as the difference between the share of recommendations for items of that group and the representation of that group in the input data: Finally, given a group C, the **disparate exposure** ($\Delta\mathcal{E}(C)$) returned by a recommender system for that group is measured as the difference between the exposure given to that group in the recommendation lists [167] and its representation.

The range of values for these scores is $[-\mathcal{R}_{(C)}, 1 - \mathcal{R}_{(C)}]$; specifically, it is 0 when the recommender system has no disparate visibility/exposure, while negative/positive values indicate that the group received a share of recommendations that is lower/higher than its representation.

8.4.1.4 Multi-Objective Recommendation at the Global and Individual Levels

The third algorithm that we used to evaluate our AMBAR dataset, is a Multi-Objective Recommendation at the Global and Individual Levels, called MOReGIn [82]. MOREGIN adjusts the recommendations according to the geographic provenance (i.e., the continent of providers) of the groups and the representation of each demographic group and seeks to make a calibration at the individual level, following the propensity of each user to rate items of a given genre. In our experiments, we assess unfairness in the provider groups by the **disparate visibility** measure $\Delta\mathcal{V}(C)$, defined in [81].

On the other hand, we asses the tendency of a system to recommend a user items whose genres are distributed differently from those they prefer via **miscalibration** ($\Delta\mathcal{M}_{ug}$) [82]. In particular, given a user $u \in U$ and a genre $g \in G$, the miscalibration returned by a recommender system for that user is measured as the difference between the share of recommendations for items of that genre and the propensity of the user for that genre in the training data.

8.4.2 Results

8.4.2.1 State-of-the-art Recommendation Algorithms

The results are depicted in Table 8.4. First, when we apply latent factor models to the AMBAR dataset, the VAECF is the best model for NDCG@50, Recall@50, and Precision@50. The second best model is WMF, being MF and SVD the ones that reached the worst results. This evaluation's objective is to determine how much the AMBAR dataset can be used to study cutting-edge algorithms that take user-item interactions into account. Our analysis shows that there are no limitations and that it is simple to apply to any kind of RS.

Table 8.4 Results obtained for the state-of-the-art RSs.

Algorithm	NDCG@50	Recall@50	Precision@50
MF	0.1381	0.0227	0.0065
WMF	0.0349	0.0428	0.0122
SVD	0.0211	0.0204	0.0058
VAECF	**0.1417**	**0.0685**	**0.0191**

8.4.2.2 Assessment of a Binary Perspective

Here, we show the results of the CPFair re-ranking method [135] using data from AMBAR. Table 8.5 depicts the performance in making recommendations, where the evaluation metrics are calculated based on the top-10 predictions in the test set considering *consumer-side* (**User Relevance** column), *producer-side* (**Tracks Exposure** column), and *consumer-provider* side (**Both** column, measured with the $mCPF$ metric) approaches. The description of the DCF, DPF, and $mCPF$ metrics are detailed in Section 8.4.1.2. The last column $\Delta(\%)$ denotes the percentage of relative improvement in $mCPF$ compared to OR. Note that all re-ranking results

Table 8.5 Performance of recommendations using the CPFair algorithm with MF, WMF, SVD, and VAECF algorithms on AMBAR. In the type, OR is fairness-unaware, C consumers, P providers, and CP are consumer-provider fairness approaches.

Model	**Type**	*Consumer-side*				*Provider-side*					*Consumer-provider side*	
		User Relevance (nDCG)				**Track Exposure**					**Both**	
		All	*Male*	*Female*	*DCF* ↓ —	*Nov.*	*Cov.*	*Male*	*Female*	*DPF* ↓ —	*mCPF* ↓	Δ(%)
MF	OR	0.0101	0.0104	0.0095	0.0009	**12.0328**	2.7100	0.6400	0.3600	0.2800	0.1405	0.0
MF	C	0.0102	0.0111	**0.0107**	**0.0004**	12.0287	**2.7100**	0.6399	0.3601	0.2798	0.1401	0.27
MF	P	0.0104	0.0112	0.0090	0.0022	12.0039	2.6700	0.5321	0.4679	0.0641	0.0331	76.40
MF	CP	**0.0107**	**0.0116**	0.0091	0.0025	11.9993	2.6900	**0.5318**	**0.4682**	**0.0636**	**0.0331**	**76.45**
WMF	OR	0.0596	0.0615	0.0582	0.0033	7.3859	6.2200	0.6531	0.3469	0.3061	0.1547	0.0
WMF	C	**0.0611**	**0.0641**	**0.0611**	**0.0030**	**7.3903**	**6.3100**	0.5500	0.4500	0.0999	0.0515	66.72
WMF	P	0.0594	0.0602	0.0569	0.0033	7.3177	6.1300	0.4912	0.5088	-0.0176	-0.0072	104.62
WMF	CP	0.0606	0.0636	0.0554	0.0083	7.3223	6.2000	**0.4565**	**0.5440**	**-0.0878**	**-0.0398**	**125.70**
SVD	OR	0.0094	0.0098	0.0065	0.0033	**5.2575**	0.0200	0.8377	0.1623	0.6754	0.3393	0.0
SVD	C	0.0122	0.0143	**0.0117**	**0.0026**	5.2317	0.0300	0.7580	0.2420	0.5161	0.2593	23.58
SVD	P	0.0095	0.0100	0.0073	0.0027	5.1446	0.0200	0.6207	0.3793	0.2414	0.1221	52.93
SVD	CP	**0.0127**	**0.0153**	0.0082	0.0071	5.1388	**0.0300**	**0.5106**	**0.4894**	**0.0213**	**0.0142**	**95.81**
VAECF	OR	0.0163	0.0167	0.0125	0.0042	5.9782	0.4800	0.6365	0.3635	0.2731	0.1386	0.0
VAECF	C	**0.0188**	**0.0212**	**0.0170**	**0.0041**	**5.9807**	**0.4802**	0.6369	0.3631	0.2737	0.1389	-0.21
VAECF	P	0.0161	0.0165	0.0103	0.0062	5.8784	0.4500	0.5183	0.4817	0.0366	0.0214	84.58
VAECF	CP	0.0177	0.0204	0.0131	0.0074	5.8819	0.4600	**0.4986**	**0.5014**	**-0.0028**	**0.0023**	**98.35**

are obtained under fairness constraints on NDCG. On the producer-side, additional beyond-accuracy metrics such as novelty (**Nov.**) and coverage (**Cov.**) metrics are applied to evaluate the recommendations of items (tracks). To measure the exposure parity, we divide items into two categories of Male and Female defined based on their gender attribute. Finally, the best results are highlighted in bold font.

We can observe that the fairness-aware methods (C, P, and CP) in all recommendation models obtained a better performance (*User Relevance*) when we analyze all, male and female users, *All* (CP: 0.0107, C: 0.0611, CP: 0.0127, C: 0.0188 in MF, WMF, SVD, and VAECF, respectively), *Male* (CP: 0.0117, C: 0.0641, CP: 0.0153, C: 0.0212 in MF, WMF, SVD, and VAECF, respectively), and *Female* (C: 0.0107, C: 0.0611, C: 0.0117, C: 0.0170 in MF, WMF, SVD, and VAECF, respectively). Regarding to tracks (items) exposure, fairness-aware methods (C, P, and CP) obtained better results than type OR methods when considering the Novelty (Nov. in the table) (C: 7.3903, C: 5.9807 in WMF and VAECF) and Coverage (Cov.) (C: 2.7100, C: 6.3100, CP: 0.0300, C: 0.4800 for MF, WMF, SVD, and VAECF) metrics. In addition, the metric of exposure of tracks (providers) concerning gender (i.e., male and female) is more equity in the results obtained with the fairness-aware (CP) methods. For example, in MF algorithms, the exposure of male artists is reduced from 0.6400 to 0.5318 in benefit to the exposure of the group of female artists, where the exposure is increased from 0.3600 to 0.4682. The trend is similar to the rest of the algorithms, where the best exposure results are for the fairness-aware methods (CP) in the group of tracks of male artists and the group of tracks of female artists. It is important to highlight that the C and P fairness-aware methods also improve the equity in the exposure of female and male provider groups.

Furthermore, if we only evaluate the DCF metric (i.e., consumer-side) (C: 0.0025, C: 0.0030, C: 0.0026, C: 0.0041) or DPF (i.e., producer-side) (CP: 0.0636, CP: -0.0878, CP:

0.0213, CP: -0.0028), all fairness-aware methods outperformed unfairness-aware method (OR) for MF, WMF, SVD, and VAECF, respectively. Moreover, if we consider both the consumer-side (DCF) and the producer-side (DPF) in $mCPF$ metrics, which evaluates the overall performance of the model taking into account the fairness for consumers and providers, all fairness-aware methods (C, P, and CP) obtained the best results (e.g., in CP, 0.0331, -0.0398, 0.0142, 0.0023). Moreover, the Δ metric demonstrate an improvement of the performance using fairness-aware methods in the range of 76.45% to 125.70%, being CP approach the best one.

Overall, the results of these experiments demonstrate that AMBAR is a suitable dataset for evaluating fairness algorithms for consumers, providers, and consumer-providers with binary attributes.

8.4.2.3 Assessment of a Multi-class Perspective

Table 8.6 shows the behavior of AMBAR dataset with both traditional RSs (i.e., MF, WMF, SVD; and VAECF) and the PFair re-ranking algorithm, which seeks to ensure fairness for providers. In all of the original models, there is some disparity ($\Delta\mathcal{V}_c$ (a)) between the proportion of recommendations given to providers from different continents and those expected to be received according to the distribution of ratings in the training set. We can also observe that these visibility disparities ($\Delta\mathcal{V}_c$ (b)) are almost completely mitigated after applying the PFair model. For example, in the MF algorithm, $\Delta Total$ is reduced from 0.1721 to 0.0000. While it is true that some disparity remains at the exposure level, it is not significant, and all cases are better than in the original models (e.g., $\Delta\mathcal{E}_c$ (b)), in the VAECF algorithm, is reduced from 0.0537 to 0.0002). Finally, one aspect to note is that the impact on the quality of recommendations is almost imperceptible (e.g., the NDCG of VAECF without fairness is 0.1417 while in VAECF using the fairness approach is 0.1416), so it is possible to offer greater visibility and exposure to providers while ensuring the efficiency of those recommendations for users.

Table 8.6 Results of traditional Recommender Systems (fairness unaware) (a) and with PFair (provider fairness) (b). The first row describes the algorithms. The second row contains the NDCG and rows 4-9 report for each demographic group: (i) the Disparate Visibility of continents ($\Delta\mathcal{V}_c$) and (ii) Disparate Exposure of continents ($\Delta\mathcal{E}_c$). The last row represents the sum of the disparities $\Delta Total$. Acronyms stand for AF: Africa, AS: Asia, EU: Europe, NA: North America, OC: Oceania, SA: South America.

	MF				**WMF**				**SVD**				**VAECF**			
	NDCG (a) 0.1381		**NDCG (b)** 0.1380		**NDCG (a)** 0.0349		**NDCG (b)** 0.0348		**NDCG (a)** 0.0211		**NDCG (b)** 0.0209		**NDCG (a)** 0.1417		**NDCG (b)** 0.1416	
	$\Delta\mathcal{V}_c$ **(a)**	$\Delta\mathcal{V}_c$ **(b)**	$\Delta\mathcal{E}_c$ **(a)**	$\Delta\mathcal{E}_c$ **(b)**	$\Delta\mathcal{V}_c$ **(a)**	$\Delta\mathcal{V}_c$ **(b)**	$\Delta\mathcal{E}_c$ **(a)**	$\Delta\mathcal{E}_c$ **(b)**	$\Delta\mathcal{V}_c$ **(a)**	$\Delta\mathcal{V}_c$ **(b)**	$\Delta\mathcal{E}_c$ **(a)**	$\Delta\mathcal{E}_c$ **(b)**	$\Delta\mathcal{V}_c$ **(a)**	$\Delta\mathcal{V}_c$ **(b)**	$\Delta\mathcal{E}_c$ **(a)**	$\Delta\mathcal{E}_c$ **(b)**
AF	-0.0014	0.0000	-0.0004	-0.0003	-0.0005	0.0000	-0.0002	0.0000	0.0016	0.0000	0.0008	0.0000	0.0008	0.0000	0.0005	0.0001
AS	0.0158	0.0000	0.0087	0.0000	-0.0143	0.0000	0.0023	0.0000	0.0421	0.0000	0.0227	0.0030	-0.0044	0.0000	0.0004	0.0000
EU	0.0441	0.0000	0.0177	0.0000	0.0259	0.0000	0.0096	0.0000	0.2036	0.0000	0.0833	0.0000	0.0510	0.0000	0.0225	0.0000
NA	-0.0846	0.0000	-0.0416	-0.0015	-0.0069	0.0000	-0.0064	0.0000	-0.1960	0.0000	-0.1288	-0.0089	-0.0511	0.0000	-0.0268	-0.0001
OC	0.0113	0.0000	0.0059	0.0000	0.0019	0.0000	0.0008	0.0000	-0.0755	0.0000	0.0075	0.0042	-0.0005	0.0000	0.0018	0.0000
SA	0.0149	0.0000	0.0097	0.0018	-0.0061	0.0000	-0.0015	0.0000	0.0243	0.0000	0.0145	0.0017	0.0043	0.0000	0.0017	0.0000
$\Delta Total$	**0.1721**	**0.0000**	**0.0840**	**0.0036**	**0.0556**	**0.0000**	**0.0208**	**0.0000**	**0.5431**	**0.0000**	**0.2576**	**0.0178**	**0.1121**	**0.0000**	**0.0537**	**0.0002**

Our experiments demonstrate that AMBAR enables to evaluate provider fairness in a multi-class setting, with graded sensitive attributes.

8.4.2.4 Assessment of a Multi-Objective Recommendation at the Global and Individual Levels

In Table 8.7 we can observe the behavior of the dataset when calculating the visibility disparity and the miscalibration both in the recommendation lists of the original models and in the lists of re-ranked recommendations obtained after performing the disparity mitigation with the MOReGIn approach. In Table 8.7, we present the sum of disparities, instead of the details per continent of providers and per genre preferences. Mainly, because the purpose of this paper is to show that AMBAR enables cross-analysis of properties at the user and item levels and not to show the effectiveness of MOReGIn at each level. For example, in the SVD algorithm, it can be observed that the disparity of the original model is 0.1869, while in MOReGIn it is 0.0032. Overall, in MOReGIn we notice that there is a significant reduction in disparities, having a greater impact on the group or global approach ($\Delta\mathcal{V}_c$), although to a lesser extent also at the level of individual users, $\Delta\mathcal{M}_{ug}$.

Table 8.7 Results of traditional Recommender Systems without fairness, and with MOReGIn on AMBAR. It shows the $NDCG$, the Disparate visibility of continents ($\Delta\mathcal{V}_c$), and the Miscalibration of genres ($\Delta\mathcal{M}_{ug}$). Each value represents the sum of disparities. Acronyms stand for OR: Original and MG: MOReGIn algorithms.

	NDCG		$\Delta\mathcal{V}_c$		$\Delta\mathcal{M}_{ug}$	
	OR	**MG**	**OR**	**MG**	**OR**	**MG**
MF	**0.1381**	0.1009	0.1548	**0.0017**	0.3645	**0.0522**
WMF	**0.0349**	0.0273	0.0326	**0.0000**	0.0463	**0.0314**
SVD	0.0211	**0.0299**	0.1869	**0.0032**	0.6950	**0.0679**
VAECF	**0.1417**	0.1103	0.1015	**0.0000**	0.1023	**0.0288**

Our experiments demonstrate that AMBAR enables to evaluate multi-objective algorithms, specifically, we have tested in MOReGIn which focuses on the continent of providers while calibrating at the individual level.

8.5 Conclusions

Proper datasets are essential for every research community. In this paper, we present the AMBAR dataset, which is a new resource for the community. AMBAR is a new dataset in the music domain, collected from a well-known platform and anonymized for preserving privacy issues, that offers user-item interactions as well as additional information on both users and items. AMBAR includes both sensitive and non-sensitive attributes of the users, items, and subjects. The proposed dataset's characteristics make it very useful for many approaches that consider beyond-accuracy properties, from multi-objective, fair to calibrated recommendations. We have provided some benchmark results on these three different tasks; it has been used in state-of-the-art Recommender Systems, in a consumer-provider fairness algorithm based on a binary perspective, in a provider fairness algorithm focused on a multi-class perspective, and in a Multi-objective recommendation algorithm. The benchmark results denote the usability of the proposal in all the analyzed tasks.

In summary, the proposed dataset is expected to benefit future research by allowing the study of systems that provide recommendations that take into account both sensitive and non-sensitive attributes of users, items, and subjects. For example, since the dataset contains more than one sensitive attribute, it offers researchers the possibility to analyze the interplay between them in future research. Apart from that, it makes it possible to conduct new analyses, such as the cross-analysis of properties at the user and item level. In addition, it also offers the possibility to analyze consumer-provider fairness approaches with a multi-class perspective, since the dataset also includes graded sensitive attributes.

Part IV

EPILOGUE

CHAPTER 9

Conclusions and Future Lines

The thesis concludes with this chapter. It provides an overview of our contributions according to the set of objectives, discusses open research challenges, and suggest futures researches.

9.1 Contribution Summary

This thesis has systematically explored the issue of algorithmic bias in Recommender Systems, particularly focusing on how data imbalances can exacerbate unfair outcomes for users and content providers. Through a thorough investigation of different types of biases—such as representation bias, disparate impact, and disparate treatment—the research has identified critical areas where traditional Recommender Systems fall short in delivering equitable results.

One of the significant findings of this research is the impact of geographic imbalance on the visibility and exposure of items in Recommender Systems. The study revealed that items from certain regions, especially those from underrepresented countries/continents, are often disadvantaged, leading to reduced visibility and unfair treatment in recommendation outputs. To address this, the thesis proposed several innovative post-processing re-ranking algorithms that adjust recommendation lists to ensure a more balanced and fair distribution of content. These approaches have been shown to effectively mitigate biases while maintaining the overall accuracy and effectiveness of the Recommender Systems.

The contributions of this thesis include the development of new datasets enriched with geographic and demographic information, the proposal of four novel fairness-aware algorithms, and the successful demonstration of these algorithms across multiple domains, including movies, books, music, and educational content.

The four proposed approaches consisted of: (i) addressing and mitigating geographic disparity bias related to the country of origin of providers groups using a binary classification (majority versus rest), (ii) the binary classification was expanded to include multiple providers groups, thereby addressing geographic disparity at the continent level through a multi-class approach. (iii) development of a methodology to ensure fairness in Recommender Systems by balancing the visibility of coarse-grained and fine-grained demographic groups, using countries within continents as a case study, (iv) creating a methodology to handle multiple objectives, aiming to achieve both global and individual balance. At the group level, we considered the continents of origin of item providers, while at the individual level, we ensured fairness for users by tailoring recommendations to their gender preferences regarding the items.

The research outcomes highlight the importance of integrating fairness considerations into the design and evaluation of Recommender Systems to ensure that they serve diverse user groups equitably.

9.2 Future Work

Building on the findings of this thesis, several directions for future research are proposed, such as, extend the methodologies developed in this thesis to other domains, such as e-commerce, news, or social media platforms and exploration of additional fairness dimensions, while this research focused on geographic and demographic fairness, future work could explore other dimensions, such as socioeconomic status, linguistic or cultural diversity, or fairness in specific content types (e.g., promoting sustainable products). This would provide a more comprehensive understanding of fairness in Recommender Systems.

Another line of research could be to develop recommendation methods that take into account multiple dimensions of discrimination (e.g., gender, ethnicity, and economic status) simultaneously, inspired by intersectionality theory. This would allow for improved fairness in recommendations for groups that are at the intersection of several marginalized categories. On the other hand, from a consumer perspective, an interesting line of research would be to investigate how bias in recommendation systems affects the psychological, social, and economic well-being of users. For example, studying how unequal exposure to certain types of content (such as financial or health products) can impact different demographic groups. Additionally, the possibility of developing new methodologies to generate synthetic data sets that are inherently fairer could also be studied, and using them to train recommendation systems, thus avoiding the biases inherent in real-world data.

Bibliography

[1] Himan Abdollahpouri and Masoud Mansoury. Multi-sided exposure bias in recommendation. *CoRR*, abs/2006.15772, 2020. URL `https://arxiv.org/abs/2006.15772`.

[2] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The unfairness of popularity bias in recommendation. *CoRR*, abs/1907.13286, 2019. URL `http://arxiv.org/abs/1907.13286`.

[3] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Augusto Pizzato. Multistakeholder recommendation: Survey and research directions. *User Model. User Adapt. Interact.*, 30(1):127–158, 2020. doi: 10.1007/s11257-019-09256-1. URL `https://doi.org/10.1007/s11257-019-09256-1`.

[4] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The connection between popularity bias, calibration, and fairness in recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, page 726–731, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375832. doi: 10.1145/3383313.3418487. URL `https://doi.org/10.1145/3383313.3418487`.

[5] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. User-centered evaluation of popularity bias in recommender systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '21, page 119–129, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383660. doi: 10.1145/3450613.3456821. URL `https://doi.org/10.1145/3450613.3456821`.

[6] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005. doi: 10.1109/TKDE.2005.99.

[7] Vito Walter Anelli, Alejandro Bellogin, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. Elliot: A comprehensive and rigorous framework for reproducible recommender systems

evaluation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2405–2414, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3463245. URL https://doi.org/10.1145/3404835.3463245.

[8] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104: 671, 2016. doi: http://dx.doi.org/10.2139/ssrn.2477899.

[9] Silvio Barra, Mirko Marras, and Gianni Fenu. Continuous authentication on smartphone by means of periocular and virtual keystroke. In Man Ho Au, Siu-Ming Yiu, Jin Li, Xiapu Luo, Cong Wang, Aniello Castiglione, and Kamil Kluczniak, editors, *Network and System Security - 12th International Conference, NSS 2018, Hong Kong, China, August 27-29, 2018, Proceedings*, volume 11058 of *Lecture Notes in Computer Science*, pages 212–220. Springer, 2018. doi: 10.1007/978-3-030-02744-5_16. URL https://doi.org/10.1007/978-3-030-02744-5_16.

[10] Christine Bauer and Markus Schedl. Global and country-specific mainstreaminess measures: Definitions, analysis, and usage for improving personalized music recommendation systems. *PLOS ONE*, 14(6):1–36, 06 2019. doi: 10.1371/journal.pone.0217389.

[11] Christine Bauer and Eva Zangerle. Leveraging multi-method evaluation for multistakeholder settings. *CoRR*, abs/2001.04348, 2020. URL https://arxiv.org/abs/2001.04348.

[12] Alejandro Bellogín, Pablo Castells, and Iván Cantador. Statistical biases in information retrieval metrics for recommender systems. *Inf. Retr. Journal*, 20(6):606–634, 2017. doi: https://doi.org/10.1007/s10791-017-9312-z.

[13] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In Anssi Klapuri and Colby Leider, editors, *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24-28, 2011*, pages 591–596. University of Miami, 2011. URL http://ismir2011.ismir.net/papers/OS6-1.pdf.

[14] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2019.*, pages 2212–2220. ACM, 2019. doi: 10.1145/3292500.3330745.

[15] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on*

Research and Development in Information Retrieval, SIGIR 2018, pages 405–414. ACM, 2018. doi: 10.1145/3209978.3210063.

[16] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46:109–132, 2013. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2013.03.012. URL `https://www.sciencedirect.com/science/article/pii/S0950705113001044`.

[17] Ludovico Boratto and Mirko Marras. Advances in bias-aware recommendation on the web. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 1147–1149, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450382977. doi: 10.1145/3437963.3441665. URL `https://doi.org/10.1145/3437963.3441665`.

[18] Ludovico Boratto, Salvatore Carta, Walid Iguider, Fabrizio Mulas, and Paolo Pilloni. Predicting workout quality to help coaches support sportspeople. In David Elsweiler, Bernd Ludwig, Alan Said, Hanna Schäfer, Helma Torkamaan, and Christoph Trattner, editors, *Proceedings of the 3rd International Workshop on Health Recommender Systems, HealthRecSys 2018, co-located with the 12th ACM Conference on Recommender Systems (ACM RecSys 2018), Vancouver, BC, Canada, October 6, 2018*, volume 2216 of *CEUR Workshop Proceedings*, pages 8–12. CEUR-WS.org, 2018. URL `http://ceur-ws.org/Vol-2216/healthRecSys18_paper_2.pdf`.

[19] Ludovico Boratto, Gianni Fenu, and Mirko Marras. The effect of algorithmic bias on recommender systems for massive open online courses. In *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part I*, volume 11437 of *Lecture Notes in Computer Science*, pages 457–472. Springer, 2019. doi: 10.1007/978-3-030-15712-8_30.

[20] Ludovico Boratto, Gianni Fenu, and Mirko Marras. Interplay between upsampling and regularization for provider fairness in recommender systems. *CoRR*, abs/2006.04279, 2020. URL `https://arxiv.org/abs/2006.04279`.

[21] Ludovico Boratto, Gianni Fenu, and Mirko Marras. Connecting user and item perspectives in popularity debiasing for collaborative recommendation. *Inf. Process. Manag.*, 58 (1):102387, 2021. doi: 10.1016/j.ipm.2020.102387. URL `https://doi.org/10.1016/j.ipm.2020.102387`.

[22] Ludovico Boratto, Gianni Fenu, and Mirko Marras. Connecting user and item perspectives in popularity debiasing for collaborative recommendation. *Information Processing and Management*, 58(1):102387, Jan 2021. ISSN 0306-4573. doi: 10.1016/j.ipm.2020.102387.

[23] Ludovico Boratto, Gianni Fenu, Mirko Marras, and Giacomo Medda. Consumer fairness in recommender systems: Contextualizing definitions and mitigations. *CoRR*, abs/2201.08614, 2022. doi: https://doi.org/10.1007/978-3-030-99736-6_37.

[24] Rodrigo Borges and Kostas Stefanidis. Enhancing long term fairness in recommendations with variational autoencoders. In *Proceedings of the 11th International Conference on Management of Digital EcoSystems*, MEDES '19, page 95–102, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450362382. doi: 10.1145/3297662.3365798. URL `https://doi.org/10.1145/3297662.3365798`.

[25] Meredith Broussard. *Artificial unintelligence: How computers misunderstand the world*. mit Press, 2018. doi: https://doi.org/10.5860/jifp.v3i2-3.6776.

[26] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018. URL `https://proceedings.mlr.press/v81/buolamwini18a.html`.

[27] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. Balanced neighborhoods for multi-sided fairness in recommendation. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 202–214. PMLR, 23–24 Feb 2018. URL `https://proceedings.mlr.press/v81/burke18a.html`.

[28] Robin Burke, Nicholas Mattei, Vladislav Grozin, Amy Voida, and Nasim Sonboli. Multi-agent social choice for dynamic fairness-aware recommendation. In *UMAP '22: 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 234–244, New York, NY, USA, 2022. ACM. doi: 10.1145/3511047.3538032. URL `https://doi.org/10.1145/3511047.3538032`.

[29] Rocío Cañamares and Pablo Castells. A probabilistic reformulation of memory-based collaborative filtering: Implications on popularity biases. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 215–224. ACM, 2017. doi: 10.1145/3077136.3080836.

[30] Jaime G. Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336. ACM, 1998. doi: 10.1145/290941.291025.

[31] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. Ranking with fairness constraints. In *45th International Colloquium on Automata, Languages, and Programming,*

ICALP 2018, volume 107 of *LIPIcs*, pages 28:1–28:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018. doi: 10.4230/LIPIcs.ICALP.2018.28.

[32] O. Celma. *Music Recommendation and Discovery in the Long Tail.* Springer, 2010. doi: https://doi.org/10.1007/978-3-642-13287-2.

[33] Òscar Celma and Perfecto Herrera. A new approach to evaluating novel recommendations. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, page 179–186, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580937. doi: 10.1145/1454008.1454038. URL `https://doi.org/10.1145/1454008.1454038`.

[34] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ACM Trans. Inf. Syst.*, 41(3), feb 2023. ISSN 1046-8188. doi: 10.1145/3564284. URL `https://doi.org/10.1145/3564284`.

[35] Jiayi Chen, Wen Wu, Liye Shi, Wei Zheng, and Liang He. Long-tail session-based recommendation from calibration. *Appl. Intell.*, 53(4):4685–4702, 2023. doi: 10.1007/s10489-022-03718-7. URL `https://doi.org/10.1007/s10489-022-03718-7`.

[36] Zhaorui Chen and Carrie Demmans Epp. Csclrec: Personalized recommendation of forum posts to support socio-collaborative learning. In Anna N. Rafferty, Jacob Whitehill, Cristóbal Romero, and Violetta Cavalli-Sforza, editors, *Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020, Fully virtual conference, July 10-13, 2020.* International Educational Data Mining Society, 2020. URL `https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_64.pdf`.

[37] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023, 2018. URL `http://arxiv.org/abs/1808.00023`.

[38] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, page 39–46, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589060. doi: 10.1145/1864708.1864721. URL `https://doi.org/10.1145/1864708.1864721`.

[39] Yashar Deldjoo, Vito Walter Anelli, Hamed Zamani, Alejandro Bellogín, and Tommaso Di Noia. Recommender systems fairness evaluation via generalized cross entropy. *CoRR*, abs/1908.06708, 2019. URL `http://arxiv.org/abs/1908.06708`.

[40] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogín, Alessandro Difonzo, and Dario Zanzonelli. A survey of research on fair recommender systems. *CoRR*, abs/2205.11127, 2022. doi: 10.48550/arXiv.2205.11127. URL `https://doi.org/10.48550/arXiv.2205.11127`.

[41] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. Fairness in recommender systems: research landscape and future directions. *User Modeling and User-Adapted Interaction*, 2023. doi: 10.1007/s11257-023-09364-z. URL `https://doi.org/10.1007/s11257-023-09364-z`.

[42] Danilo Dessì, Gianni Fenu, Mirko Marras, and Diego Reforgiato Recupero. Leveraging cognitive computing for multi-class classification of e-learning videos. In *The Semantic Web: ESWC 2017 Satellite Events, Revised Selected Papers*, volume 10577 of *Lecture Notes in Computer Science*, pages 21–25. Springer, 2017. doi: 10.1007/978-3-319-70407-4_5.

[43] Danilo Dessì, Gianni Fenu, Mirko Marras, and Diego Reforgiato Recupero. COCO: semantic-enriched collection of online courses at scale with experimental use cases. In *Trends and Advances in Information Systems and Technologies - Volume 2 [WorldCIST'18]*, volume 746 of *Advances in Intelligent Systems and Computing*, pages 1386–1396. Springer, 2018. doi: 10.1007/978-3-319-77712-2_133.

[44] Danilo Dessì, Mauro Dragoni, Gianni Fenu, Mirko Marras, and Diego Reforgiato Recupero. Evaluating neural word embeddings created from online course reviews for sentiment analysis. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, SAC 2019*, pages 2124–2127. ACM, 2019. doi: 10.1145/3297280.3297620.

[45] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, CIKM '20, page 275–284, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3411962. URL `https://doi.org/10.1145/3340531.3411962`.

[46] Patrik Dokoupil, Ladislav Peska, and Ludovico Boratto. Rows or columns? minimizing presentation bias when comparing multiple recommender systems. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete, editors, *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2354–2358. ACM, 2023. doi: 10.1145/3539618.3592056. URL `https://doi.org/10.1145/3539618.3592056`.

[47] Patrik Dokoupil, Ladislav Peska, and Ludovico Boratto. Looks can be deceiving: Linking user-item interactions and user's propensity towards multi-objective recommendations. *CoRR*, abs/2307.00654, 2023. doi: 10.48550/arXiv.2307.00654. URL https://doi.org/10.48550/arXiv.2307.00654.

[48] Shayan Doroudi and Emma Brunskill. Fairer but not fair enough on the equitability of knowledge tracing. In *Proceedings of the 9th International Conference on Learning Analytics and Knowledge, LAK 2019, Tempe, AZ, USA, March 4-8, 2019*, pages 335–339. ACM, 2019. doi: 10.1145/3303772.3303838. URL https://doi.org/10.1145/3303772.3303838.

[49] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

[50] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 172–186. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/ekstrand18b.html.

[51] Michael D. Ekstrand, Mucun Tian, Mohammed R. Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018*, pages 242–250. ACM, 2018. doi: 10.1145/3240323.3240373.

[52] Michael D Ekstrand, Robin Burke, and Fernando Diaz. Fairness and discrimination in recommendation and retrieval. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, page 576–577, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362436. doi: 10.1145/3298689.3346964. URL https://doi.org/10.1145/3298689.3346964.

[53] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. Fairness in recommender systems. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 679–707. Springer US, 2022. doi: 10.1007/978-1-0716-2197-4_18. URL https://doi.org/10.1007/978-1-0716-2197-4_18.

[54] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. *Fairness in Recommender Systems*, pages 679–707. Springer US, New York, NY, 2022. ISBN 978-1-0716-2197-4. doi: 10.1007/978-1-0716-2197-4_18.

[55] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. Fairness and discrimination in information access systems. *CoRR*, abs/2105.05779, 2022. URL https://arxiv.org/abs/2105.05779.

[56] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. Fairness in information access systems. *Foundations and Trends® in Information Retrieval*, 16(1-2): 1–177, 2022. ISSN 1554-0669. doi: 10.1561/1500000079. URL http://dx.doi.org/10.1561/1500000079.

[57] Mehdi Elahi, Mouzhi Ge, Francesco Ricci, David Massimo, and Shlomo Berkovsky. Interactive food recommendation for groups. In *RecSys Posters*, 2014. URL https://api.semanticscholar.org/CorpusID:14680670.

[58] Asmaa Elbadrawy, Agoritsa Polyzou, Zhiyun Ren, Mackenzie Sweeney, George Karypis, and Huzefa Rangwala. Predicting student performance using personalized analytics. *Computer*, 49(4):61–69, 2016. doi: 10.1109/MC.2016.119. URL https://doi.org/10.1109/MC.2016.119.

[59] Francesco Fabbri, Francesco Bonchi, Ludovico Boratto, and Carlos Castillo. The effect of homophily on disparate visibility of minorities in people recommender systems. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020*, pages 165–175. AAAI Press, 2020. doi: 10.1609/icwsm.v14i1.7288.

[60] Fakhri Fauzan, Dade Nurjanah, and Rita Rismala. Apriori association rule for course recommender system. *Indonesia Journal on Computing (Indo-JC)*, 5(2):1–16, Oct. 2020. doi: 10.34818/INDOJC.2020.5.2.434.

[61] Wenzheng Feng, Jie Tang, and Tracy Xiao Liu. Understanding dropouts in moocs. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 517–524. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.3301517. URL https://doi.org/10.1609/aaai.v33i01.3301517.

[62] Gianni Fenu, Mirko Marras, and Massimiliano Meles. A learning analytics tool for usability assessment in moodle environments. *Journal of e-Learning and Knowledge Society*, 13(3), September 2017. ISSN 1826-6223. URL https://www.learntechlib.org/p/180986.

[63] Gianni Fenu, Hicham Lafhouli, and Mirko Marras. Exploring algorithmic fairness in deep speaker verification. In Osvaldo Gervasi, Beniamino Murgante, Sanjay Misra, Chiara Garau, Ivan Blecic, David Taniar, Bernady O. Apduhan, Ana Maria A. C. Rocha, Eufemia Tarantino, Carmelo Maria Torre, and Yeliz Karaca, editors, *Computational Science and Its Applications - ICCSA 2020 - 20th International Conference, Cagliari, Italy, July 1-4, 2020, Proceedings, Part IV*, volume 12252 of *Lecture Notes in Computer Science*, pages 77–93. Springer, 2020. doi: 10.1007/978-3-030-58811-3_6.

[64] Andres Ferraro, Xavier Serra, and Christine Bauer. What is fair? exploring the artists' perspective on the fairness of music streaming platforms. In *Human-Computer Interaction - INTERACT 2021 - 18th IFIP TC 13 International Conference, Proceedings, Part II*, volume 12933 of *Lecture Notes in Computer Science*, pages 562–584. Springer, 2021. doi: 10.1007/978-3-030-85616-8_33. URL https://doi.org/10.1007/978-3-030-85616-8_33.

[65] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3):330–347, jul 1996. ISSN 1046-8188. doi: 10.1145/230538.230561. URL https://doi.org/10.1145/230538.230561.

[66] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, and Gerard de Melo. Fairness-aware explainable recommendation over knowledge graphs. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 69–78, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401051. URL https://doi.org/10.1145/3397271.3401051.

[67] Ruoyuan Gao and Chirag Shah. How fair can we go: Detecting the boundaries of fairness optimization in information retrieval. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 229–236, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368810. doi: 10.1145/3341981.3344215. URL https://doi.org/10.1145/3341981.3344215.

[68] Vishal Garg and Ritu Tiwari. Hybrid massive open online course (mooc) recommendation system using machine learning. In *International Conference on Recent Trends in Engineering, Science Technology - (ICRTEST 2016)*, pages 1–5, 2016. doi: 10.1049/cp.2016.1479.

[69] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, page 257–260, New York,

NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589060. doi: 10.1145/1864708.1864761. URL https://doi-org.sire.ub.edu/10.1145/1864708.1864761.

[70] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, page 257–260, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589060. doi: 10.1145/1864708.1864761. URL https://doi.org/10.1145/1864708.1864761.

[71] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, and Yongfeng Zhang. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 445–453, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450382977. doi: 10.1145/3437963.3441824. URL https://doi.org/10.1145/3437963.3441824.

[72] Yingqiang Ge, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. Toward pareto efficient fairness-utility trade-off in recommendation through reinforcement learning. In K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang, editors, *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 316–324. ACM, 2022. doi: 10.1145/3488560.3498487. URL https://doi.org/10.1145/3488560.3498487.

[73] Yingqiang Ge, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. Toward pareto efficient fairness-utility trade-off in recommendation through reinforcement learning. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 316–324, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391320. doi: 10.1145/3488560.3498487. URL https://doi.org/10.1145/3488560.3498487.

[74] Salvatore Genovese. Artificial intelligence: A guide for thinking humans. *ORDO*, 71(1): 444–449, 2020. doi: doi:10.1515/ordo-2021-0028. URL https://doi.org/10.1515/ordo-2021-0028.

[75] Alireza Gharahighehi, Celine Vens, and Konstantinos Pliakos. Fair multi-stakeholder news recommender system with hypergraph ranking. *Information Processing and Management*, 58(5):102663, 2021. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.

102663. URL https://www.sciencedirect.com/science/article/pii/S0306457321001515.

[76] Elizabeth Gómez, Ludovico Boratto, and Maria Salamó. Disparate impact in item recommendation: A case of geographic imbalance. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors, *Advances in Information Retrieval*, pages 190–206, Cham, 2021. Springer International Publishing. ISBN 978-3-030-72113-8. doi: 10.1007/978-3-030-72113-8_13.

[77] Elizabeth Gómez, Ludovico Boratto, and Maria Salamó. Disparate impact in item recommendation: A case of geographic imbalance. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I*, volume 12656 of *Lecture Notes in Computer Science*, pages 190–206. Springer, 2021. doi: 10.1007/978-3-030-72113-8_13. URL https://doi.org/10.1007/978-3-030-72113-8_13.

[78] Elizabeth Gómez, Ludovico Boratto, and Maria Salamó. Disparate impact in item recommendation: A case of geographic imbalance. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors, *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I*, volume 12656 of *Lecture Notes in Computer Science*, pages 190–206. Springer, 2021. doi: 10.1007/978-3-030-72113-8_13. URL https://doi.org/10.1007/978-3-030-72113-8_13.

[79] Elizabeth Gómez, Carlos Shui Zhang, Ludovico Boratto, Maria Salamó, and Mirko Marras. The winner takes it all: Geographic imbalance and provider (un)fairness in educational recommender systems. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1808–1812. ACM, 2021. doi: 10.1145/3404835.3463235.

[80] Elizabeth Gómez, Ludovico Boratto, and Maria Salamó. Provider fairness across continents in collaborative recommender systems. *Inf. Process. Manag.*, 59(1):102719, 2022. doi: 10.1016/j.ipm.2021.102719.

[81] Elizabeth Gómez, Carlos Shui Zhang, Ludovico Boratto, Maria Salamó, and Guilherme Ramos. Enabling cross-continent provider fairness in educational recommender systems. *Future Gener. Comput. Syst.*, 127:435–447, 2022. doi: 10.1016/j.future.2021.08.025.

[82] Elizabeth Gómez, David Contreras, Ludovico Boratto, and Maria Salamó. Moregin: Multi-objective recommendation at the global and individual levels. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald,

and Iadh Ounis, editors, *Advances in Information Retrieval*, pages 21–38, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-56027-9. doi: https://doi.org/10.1007/978-3-031-56027-9_2.

[83] Jibing Gong, Shen Wang, Jinlong Wang, Wenzheng Feng, Hao Peng, Jie Tang, and Philip S. Yu. Attentional graph convolutional networks for knowledge concept recommendation in moocs in a heterogeneous view. In Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 79–88. ACM, 2020. doi: 10.1145/3397271.3401057. URL `https://doi.org/10.1145/3397271.3401057`.

[84] Jyotirmoy Gope and Sanjay Kumar Jain. A learning styles based recommender system prototype for edx courses. In *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, pages 414–419, 2017. doi: 10.1109/SmartTechCon.2017.8358407.

[85] Ben Green and Lily Hu. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *International Conference on Machine Learning*, 2018. URL `https://api.semanticscholar.org/CorpusID:49544563`.

[86] Sara Hajian, Francesco Bonchi, and Carlos Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2125–2126. ACM, 2016. doi: 10.1145/2939672.2945386.

[87] Jungkyu HAN and Hayato YAMANA. A survey on recommendation methods beyond accuracy. *IEICE Transactions on Information and Systems*, E100.D(12):2931–2944, 2017. doi: 10.1587/transinf.2017EDR0003.

[88] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015. doi: https://doi.org/10.1145/2827872.

[89] David Hauger, Markus Schedl, Andrej Kosir, and Marko Tkalcic. The million musical tweet dataset - what we can learn from microblogs. In Alceu de Souza Britto Jr., Fabien Gouyon, and Simon Dixon, editors, *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR 2013, Curitiba, Brazil, November 4-8, 2013*, pages 189–194, 2013. URL `http://www.ppgia.pucpr.br/ismir2013/wp-content/uploads/2013/09/85_Paper.pdf`.

[90] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 173–182, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349130. doi: 10.1145/3038912.3052569. URL `https://doi.org/10.1145/3038912.3052569`.

[91] Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 230–237, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581130961. doi: 10.1145/312624.312682. URL `https://doi.org/10.1145/312624.312682`.

[92] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. *Inf. Retr.*, 5(4): 287–310, 2002. doi: 10.1023/A:1020443909834.

[93] Kenneth Holstein and Shayan Doroudi. Fairness and equity in learning analytics systems (fairlak). In *Companion Proceedings of the Ninth International Learning Analytics and Knowledge Conference (LAK 2019)*, 2019.

[94] Qian Hu and Huzefa Rangwala. Towards fair educational data mining: A case study on detecting at-risk students. In Anna N. Rafferty, Jacob Whitehill, Cristóbal Romero, and Violetta Cavalli-Sforza, editors, *Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020, Fully virtual conference, July 10-13, 2020*. International Educational Data Mining Society, 2020. URL `https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_157.pdf`.

[95] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, pages 263–272, 2008. doi: 10.1109/ICDM.2008.22.

[96] Rashidul Islam, Kamrun Naher Keya, Ziqian Zeng, Shimei Pan, and James Foulds. Debiasing career recommendations with neural fair collaborative filtering. In *Proceedings of the Web Conference 2021*, WWW '21, page 3779–3790, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3449904. URL `https://doi.org/10.1145/3442381.3449904`.

[97] Dietmar Jannach. Multi-objective recommender systems: Survey and challenges, 2022. URL `https://arxiv.org/abs/2210.10309`.

[98] Dietmar Jannach. Multi-objective recommendation: Overview and challenges. In Himan Abdollahpouri, Shaghayegh Sahebi, Mehdi Elahi, Masoud Mansoury, Babak Loni, Zahra Nazari, and Maria Dimakopoulou, editors, *Proceedings of the 2nd Workshop on Multi-Objective Recommender Systems co-located with 16th ACM Conference on Recommender Systems (RecSys 2022), Seattle, WA, USA, 18th-23rd September 2022*, volume 3268 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022. URL `https://ceur-ws.org/Vol-3268/paper1.pdf`.

[99] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002. doi: 10.1145/582415.582418.

[100] Di Jin, Luzhi Wang, He Zhang, Yizhen Zheng, Weiping Ding, Feng Xia, and Shirui Pan. A survey on fairness-aware recommender systems. *Information Fusion*, 100:101906, 2023. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.101906. URL `https://www.sciencedirect.com/science/article/pii/S1566253523002221`.

[101] Michael Jugovac, Dietmar Jannach, and Lukas Lerche. Efficient optimization of multiple recommendation quality factors according to individual user tendencies. *Expert Systems with Applications*, 81:321–331, 2017. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2017.03.055. URL `https://www.sciencedirect.com/science/article/pii/S0957417417302075`.

[102] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Recommendation independence. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 187–201. PMLR, 23–24 Feb 2018. URL `https://proceedings.mlr.press/v81/kamishima18a.html`.

[103] Yeqin Kang. An analysis on spoc: Post-mooc era of online education. *Tsinghua Journal of Education*, 35(1):85–93, 2014.

[104] Evangelos Karakolis, Panagiotis Kokkinakos, and Dimitrios Askounis. Provider fairness for diversity and coverage in multi-stakeholder recommender systems. *Applied Sciences*, 12(10), 2022. ISSN 2076-3417. doi: 10.3390/app12104984. URL `https://www.mdpi.com/2076-3417/12/10/4984`.

[105] Baris Kirdemir, Joseph Kready, Esther Mead, Muhammad Nihal Hussain, and Nitin Agarwal. Examining video recommendation bias on youtube. In *Advances in Bias and Fairness in Information Retrieval*, pages 106–116, Cham, 2021. Springer International Publishing. doi: https://doi.org/10.1007/978-3-030-78818-6_10.

[106] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 426–434, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934. doi: 10.1145/1401890.1401944.

[107] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 426–434. ACM, 2008. doi: 10.1145/1401890.1401944.

[108] Yehuda Koren, Robert M. Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009. doi: 10.1109/MC.2009.263.

[109] Dominik Kowald, Markus Schedl, and Elisabeth Lex. The unfairness of popularity bias in music recommendation: A reproducibility study. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II*, volume 12036 of *Lecture Notes in Computer Science*, pages 35–42. Springer, 2020. doi: 10.1007/978-3-030-45442-5_5. URL https://doi.org/10.1007/978-3-030-45442-5_5.

[110] Hugues Labarthe, François Bouchet, Rémi Bachelet, and Kalina Yacef. Does a peer recommender foster students' engagement in moocs? In Tiffany Barnes, Min Chi, and Mingyu Feng, editors, *Proceedings of the 9th International Conference on Educational Data Mining, EDM 2016, Raleigh, North Carolina, USA, June 29 - July 2, 2016*, pages 418–423. International Educational Data Mining Society (IEDMS), 2016. URL http://www.educationaldatamining.org/EDM2016/proceedings/paper_171.pdf.

[111] Jie Li, Yongli Ren, and Ke Deng. Fairgan: Gans-based fairness-aware learning for recommendations with implicit feedback. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 297–307, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965. doi: 10.1145/3485447.3511958. URL https://doi.org/10.1145/3485447.3511958.

[112] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. Towards deep conversational recommendations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Asso-

ciates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/800de15c79c8d840f4e78d3af937d4d4-Paper.pdf.

[113] Yang Li, Kangbo Liu, Ranjan Satapathy, Suhang Wang, and Erik Cambria. Recent developments in recommender systems: A survey [review article]. *IEEE Computational Intelligence Magazine*, 19(2):78–95, 2024. doi: 10.1109/MCI.2024.3363984.

[114] Yangkun Li, Mohamed-Laid Hedia, Weizhi Ma, Hongyu Lu, Min Zhang, Yiqun Liu, and Shaoping Ma. Contextualized fairness for recommender systems in premium scenarios. *Big Data Research*, 27:100300, 2022. ISSN 2214-5796. doi: https://doi.org/10.1016/j.bdr.2021.100300. URL https://www.sciencedirect.com/science/article/pii/S2214579621001179.

[115] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. User-oriented fairness in recommendation. In Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia, editors, *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 624–632. ACM / IW3C2, 2021. doi: 10.1145/3442381.3449866. URL https://doi.org/10.1145/3442381.3449866.

[116] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. Towards personalized fairness based on causal notion. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1054–1063, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462966. URL https://doi.org/10.1145/3404835.3462966.

[117] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 689–698, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356398. doi: 10.1145/3178876.3186150. URL https://doi.org/10.1145/3178876.3186150.

[118] Kun Lin, Nasim Sonboli, Bamshad Mobasher, and Robin Burke. Calibration in collaborative filtering recommender systems: a user-centered analysis. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, HT '20, page 197–206, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370981. doi: 10.1145/3372923.3404793. URL https://doi.org/10.1145/3372923.3404793.

[119] Xiao Lin, Hongjie Chen, Changhua Pei, Fei Sun, Xuanji Xiao, Hanxiao Sun, Yongfeng Zhang, Wenwu Ou, and Peng Jiang. A pareto-efficient algorithm for multiple objec-

tive optimization in e-commerce recommendation. In Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk, editors, *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, pages 20–28. ACM, 2019. doi: 10.1145/3298689.3346998. URL `https://doi.org/10.1145/3298689.3346998`.

[120] Weiwen Liu and Robin Burke. Personalizing fairness-aware re-ranking. *CoRR*, abs/1809.02921, 2018. URL `http://arxiv.org/abs/1809.02921`.

[121] Weiwen Liu, Feng Liu, Ruiming Tang, Ben Liao, Guangyong Chen, and Pheng Ann Heng. Balancing between accuracy and fairness for interactive recommendation with reinforcement learning. In Hady W. Lauw, Raymond Chi-Wing Wong, Alexandros Ntoulas, Ee-Peng Lim, See-Kiong Ng, and Sinno Jialin Pan, editors, *Advances in Knowledge Discovery and Data Mining*, pages 155–167, Cham, 2020. Springer International Publishing. doi: https://doi.org/10.1007/978-3-030-47426-3_13.

[122] Mykola Makhortykh, Aleksandra Urman, and Roberto Ulloa. Detecting race and gender bias in visual representation of ai on web search engines. In *Advances in Bias and Fairness in Information Retrieval*, pages 36–50, Cham, 2021. Springer International Publishing. ISBN 978-3-030-78818-6. doi: https://doi.org/10.1007/978-3-030-78818-6_5.

[123] Masoud Mansoury, Bamshad Mobasher, Robin Burke, and Mykola Pechenizkiy. Bias disparity in collaborative recommendation: Algorithmic evaluation and comparison. *CoRR*, abs/1908.00831, 2019. URL `http://arxiv.org/abs/1908.00831`.

[124] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. Fairmatch: A graph-based approach for improving aggregate diversity in recommender systems. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, page 154–162, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368612. doi: 10.1145/3340631.3394860. URL `https://doi.org/10.1145/3340631.3394860`.

[125] Benjamin Marlin. *Collaborative filtering: A machine learning perspective*. University of Toronto Toronto, 2004.

[126] Mirko Marras, Pawel Korus, Nasir D. Memon, and Gianni Fenu. Adversarial optimization for dictionary attacks on speaker verification. In Gernot Kubin and Zdravko Kacic, editors, *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 2913–2917. ISCA, 2019. doi: 10.21437/Interspeech.2019-2430.

[127] Mirko Marras, Ludovico Boratto, Guilherme Ramos, and Gianni Fenu. Equality of

learning opportunity via individual fairness in personalized recommendations. *CoRR*, abs/2006.04282, 2020. URL https://arxiv.org/abs/2006.04282.

[128] Mirko Marras, Ludovico Boratto, Guilherme Ramos, and Gianni Fenu. Equality of learning opportunity via individual fairness in personalized recommendations. *International Journal of Artificial Intelligence in Education*, 2021. doi: 10.1007/s40593-021-00271-1. URL https://doi.org/10.1007/s40593-021-00271-1.

[129] Sean M. McNee, John Riedl, and Joseph A. Konstan. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, page 1097–1101, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595932984. doi: 10.1145/1125451.1125659. URL https://doi.org/10.1145/1125451.1125659.

[130] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness and satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018*, pages 2243–2251. ACM, 2018. doi: 10.1145/3269206.3272027.

[131] Alessandro B. Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing and Management*, 58(5):102666, 2021. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.102666. URL https://www.sciencedirect.com/science/article/pii/S0306457321001540.

[132] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data and Society*, 3(2): 2053951716679679, 2016. doi: https://doi.org/10.1177/2053951716679679.

[133] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 429–438, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401100. URL https://doi.org/10.1145/3397271.3401100.

[134] Mohammadmehdi Naghiaei, Hossein A. Rahmani, and Yashar Deldjoo. Cpfair: Personalized consumer and producer fairness re-ranking for recommender systems. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 770–779. ACM, 2022. doi: 10.1145/3477495.3531959. URL https://doi.org/10.1145/3477495.3531959.

[135] Mohammadmehdi Naghiaei, Hossein A. Rahmani, and Yashar Deldjoo. Cpfair: Personalized consumer and producer fairness re-ranking for recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 770–779, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531959. URL `https://doi.org/10.1145/3477495.3531959`.

[136] Xia Ning, Christian Desrosiers, and George Karypis. A comprehensive survey of neighborhood-based recommendation methods. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 37–76. Springer, 2015. doi: 10.1007/978-1-4899-7637-6_2. URL `https://doi.org/10.1007/978-1-4899-7637-6_2`.

[137] Jeremy Nixon, Mike Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. *CoRR*, abs/1904.01685, 2019. URL `http://arxiv.org/abs/1904.01685`.

[138] Yanxia Pang, Na Wang, Ying Zhang, Yuanyuan Jin, Wendi Ji, and Wenan Tan. Prerequisite-related mooc recommendation on learning path locating. *Computational Social Networks*, 6(1):1–16, 2019. doi: https://doi.org/10.1186/s40649-019-0065-2.

[139] Gourab K. Patro, Arpita Biswas, Niloy Ganguly, Krishna P. Gummadi, and Abhijnan Chakraborty. Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In *WWW '20: The Web Conference 2020*, pages 1194–1204. ACM / IW3C2, 2020. doi: 10.1145/3366423.3380196.

[140] Michael J Pazzani and Daniel Billsus. Content-based recommendation systems. In *The adaptive web: methods and strategies of web personalization*, pages 325–341. Springer, 2007. doi: https://doi.org/10.1007/978-3-540-72079-9_10.

[141] Ladislav Peska and Patrik Dokoupil. Towards results-level proportionality for multi-objective recommender systems. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 1963–1968. ACM, 2022. doi: 10.1145/3477495.3531787. URL `https://doi.org/10.1145/3477495.3531787`.

[142] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. Fairness in rankings and recommendations: an overview. *The VLDB Journal*, 31(3):431–458, 2022. doi: 10.1007/s00778-021-00697-y. URL `https://doi.org/10.1007/s00778-021-00697-y`.

[143] Boyd A. Potts, Hassan Khosravi, Carl Reidsema, Aneesha Bakharia, Mark Belonogoff, and Melanie Fleming. Reciprocal peer recommendation for learning purposes. In Abelardo Pardo, Kathryn Bartimote-Aufflick, Grace Lynch, Simon Buckingham Shum, Rebecca Ferguson, Agathe Merceron, and Xavier Ochoa, editors, *Proceedings of the 8th International Conference on Learning Analytics and Knowledge, LAK 2018, Sydney, NSW, Australia, March 07-09, 2018*, pages 226–235. ACM, 2018. doi: 10.1145/3170358.3170400. URL `https://doi.org/10.1145/3170358.3170400`.

[144] Jiezhong Qiu, Jie Tang, Tracy Xiao Liu, Jie Gong, Chenhui Zhang, Qian Zhang, and Yufei Xue. Modeling and predicting learning behavior in moocs. In Paul N. Bennett, Vanja Josifovski, Jennifer Neville, and Filip Radlinski, editors, *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016*, pages 93–102. ACM, 2016. doi: 10.1145/2835776.2835842. URL `https://doi.org/10.1145/2835776.2835842`.

[145] Amifa Raj and Michael D. Ekstrand. Measuring fairness in ranked results: An analytical and empirical comparison. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 726–736, New York, NY, USA, 2022. ACM. doi: 10.1145/3477495.3532018. URL `https://doi.org/10.1145/3477495.3532018`.

[146] Guilherme Ramos and Ludovico Boratto. Reputation (in)dependence in ranking systems: Demographics influence over output disparities. In Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 2061–2064. ACM, 2020. doi: 10.1145/3397271.3401278.

[147] Guilherme Ramos and Carlos Caleiro. A novel similarity measure for group recommender systems with optimal time complexity. In Ludovico Boratto, Stefano Faralli, Mirko Marras, and Giovanni Stilo, editors, *Bias and Social Aspects in Search and Recommendation - First International Workshop, BIAS 2020, Lisbon, Portugal, April 14, 2020, Proceedings*, volume 1245 of *Communications in Computer and Information Science*, pages 95–109. Springer, 2020. doi: 10.1007/978-3-030-52485-2_10.

[148] Guilherme Ramos, Ludovico Boratto, and Carlos Caleiro. On the negative impact of social influence in recommender systems: A study of bribery in collaborative hybrid algorithms. *Inf. Process. Manag.*, 57(2):102058, 2020. doi: 10.1016/j.ipm.2019.102058.

[149] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: bayesian personalized ranking from implicit feedback. *CoRR*, abs/1205.2618, 2012. URL `http://arxiv.org/abs/1205.2618`.

[150] Peter J Rentfrow and Samuel D Gosling. The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology*, 84(6):1236, 2003. doi: https://doi.org/10.1037/0022-3514.84.6.1236.

[151] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW '94, page 175–186, New York, NY, USA, 1994. Association for Computing Machinery. ISBN 0897916891. doi: 10.1145/192844.192905. URL `https://doi.org/10.1145/192844.192905`.

[152] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems handbook. *Recommender Systems Handbook*, 1-35:1–35, 10 2010. doi: 10.1007/978-0-387-85820-3_1.

[153] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: Introduction and challenges. In *Recommender Systems Handbook*, pages 1–34. Springer, 2015. doi: 10.1007/978-1-4899-7637-6_1.

[154] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: Techniques, applications, and challenges. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 1–35. Springer US, 2022. doi: 10.1007/978-1-0716-2197-4_1. URL `https://doi.org/10.1007/978-1-0716-2197-4_1`.

[155] Carlos Rojas, David Contreras, and Maria Salamó. Analysis of biases in calibrated recommendations. In *Advances in Bias and Fairness in Information Retrieval*, pages 91–103, Cham, 2022. Springer International Publishing. ISBN 978-3-031-09316-6. doi: https://doi.org/10.1007/978-3-031-09316-6_9.

[156] Aghiles Salah, Quoc-Tuan Truong, and Hady W. Lauw. Cornac: A comparative framework for multimodal recommender systems. *Journal of Machine Learning Research*, 21 (95):1–5, 2020. URL `http://jmlr.org/papers/v21/19-805.html`.

[157] Pablo Sánchez and Alejandro Bellogín. Applying reranking strategies to route recommendation using sequence-aware evaluation. *User Model. User Adapt. Interact.*, 30(4): 659–725, 2020. doi: 10.1007/s11257-020-09258-4. URL `https://doi.org/10.1007/s11257-020-09258-4`.

[158] Piotr Sapiezynski, Wesley Zeng, Ronald E. Robertson, Alan Mislove, and Christo Wilson. Quantifying the impact of user attentionon fair group representation in ranked lists. In *Companion of The 2019 World Wide Web Conference, WWW 2019*, pages 553–562. ACM, 2019. doi: 10.1145/3308560.3317595.

[159] Badrul Munir Sarwar, George Karypis, Joseph A. Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the Tenth International World Wide Web Conference, WWW 10*, pages 285–295. ACM, 2001. doi: 10.1145/371920.372071.

[160] Ryoma Sato. Enumerating fair packages for group recommendations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 870–878, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391320. doi: 10.1145/3488560.3498432. URL `https://doi.org/10.1145/3488560.3498432`.

[161] João Saúde, Guilherme Ramos, Carlos Caleiro, and Soummya Kar. Reputation-based ranking systems and their resistance to bribery. In Vijay Raghavan, Srinivas Aluru, George Karypis, Lucio Miele, and Xindong Wu, editors, *2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017*, pages 1063–1068. IEEE Computer Society, 2017. doi: 10.1109/ICDM.2017.139.

[162] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web: methods and strategies of web personalization*, pages 291–324. Springer, 2007. doi: https://doi.org/10.1007/978-3-540-72079-9_9.

[163] Markus Schedl. The lfm-1b dataset for music retrieval and recommendation. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, page 103–110, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450343596. doi: 10.1145/2911996.2912004. URL `https://doi.org/10.1145/2911996.2912004`.

[164] Dimitris Serbos, Shuyao Qi, Nikos Mamoulis, Evaggelia Pitoura, and Panayiotis Tsaparas. Fairness in package-to-group recommendations. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 371–379, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349130. doi: 10.1145/3038912.3052612. URL `https://doi.org/10.1145/3038912.3052612`.

[165] Sinan Seymen, Himan Abdollahpouri, and Edward C. Malthouse. A constrained optimization approach for calibrated recommendations. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 607–612, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384582. doi: 10.1145/3460231.3478857. URL `https://doi.org/10.1145/3460231.3478857`.

[166] Dhawal Shah. By the Numbers: MOOCs During the Pandemic. `https://www.classcentral.com/report/mooc-stats-pandemic/`, 2020. [Online; accessed 02-March-2021].

[167] Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2018*, pages 2219–2228. ACM, 2018. doi: 10.1145/3219819.3220088. URL https://doi.org/10.1145/3219819.3220088.

[168] Nasim Sonboli and Robin Burke. Localized fairness in recommender systems. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, UMAP 2019*, pages 295–300. ACM, 2019. doi: 10.1145/3314183.3323845. URL https://doi.org/10.1145/3314183.3323845.

[169] Dusan Stamenkovic, Alexandros Karatzoglou, Ioannis Arapakis, Xin Xin, and Kleomenis Katevas. Choosing the best of both worlds: Diverse and novel recommendations through multi-objective reinforcement learning. In K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang, editors, *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 957–965. ACM, 2022. doi: 10.1145/3488560.3498471. URL https://doi.org/10.1145/3488560.3498471.

[170] Harald Steck. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, page 154–162, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359016. doi: 10.1145/3240323.3240372.

[171] Maria Stratigi, Jyrki Nummenmaa, Evaggelia Pitoura, and Kostas Stefanidis. Fair sequential group recommendations. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, SAC '20, page 1443–1452, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368667. doi: 10.1145/3341105.3375766. URL https://doi.org/10.1145/3341105.3375766.

[172] Saedeh Tahery, Seyyede Zahra Aftabi, and Saeed Farzi. A ga-based algorithm meets the fair ranking problem. *Information Processing and Management*, 58 (6):102711, 2021. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.102711. URL https://www.sciencedirect.com/science/article/pii/S0306457321001953.

[173] Sergio Torrijos, Alejandro Bellogín, and Pablo Sánchez. Discovering related users in location-based social networks. In Tsvi Kuflik, Ilaria Torre, Robin Burke, and Cristina Gena, editors, *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2020, Genoa, Italy, July 12-18, 2020*, pages 353–357. ACM, 2020. doi: 10.1145/3340631.3394882. URL https://doi.org/10.1145/3340631.3394882.

[174] Thi Ngoc Trang Tran, Alexander Felfernig, Christoph Trattner, and Andreas Holzinger. Recommender systems in the healthcare domain: state-of-the-art and research issues. *Journal of Intelligent Information Systems*, pages 1–31, 2020. doi: 10.1007/s10844-020-00633-6.

[175] Gabriel Vigliensoni and Ichiro Fujinaga. Automatic music recommendation systems: Do demographic, profiling, and contextual features improve their performance? In *International Society for Music Information Retrieval Conference*, 2016. URL `https://api.semanticscholar.org/CorpusID:17941472`.

[176] Elaine Walster, Ellen Berscheid, and G William Walster. New directions in equity research. *Journal of personality and social psychology*, 25(2):151, 1973. doi: https://doi.org/10.1037/h0033967.

[177] Mengting Wan, Jianmo Ni, Rishabh Misra, and Julian McAuley. Addressing marketing bias in product recommendations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 618–626, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368223. doi: 10.1145/3336191.3371855. URL `https://doi.org/10.1145/3336191.3371855`.

[178] Zhaoyuan Wang, Chuishi Meng, Shenggong Ji, Tianrui Li, and Yu Zheng. Food package suggestion system based on multi-objective optimization: A case study on a real-world restaurant. *Applied Soft Computing*, 93:106369, 2020. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2020.106369. URL `https://www.sciencedirect.com/science/article/pii/S1568494620303094`.

[179] Haolun Wu, Chen Ma, Bhaskar Mitra, Fernando Diaz, and Xue Liu. A multi-objective optimization framework for multi-stakeholder fairness-aware recommendation. *ACM Trans. Inf. Syst.*, 41(2), dec 2022. ISSN 1046-8188. doi: 10.1145/3564285. URL `https://doi.org/10.1145/3564285`.

[180] Haolun Wu, Bhaskar Mitra, Chen Ma, Fernando Diaz, and Xue Liu. Joint multisided exposure fairness for recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 703–714, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3532007. URL `https://doi.org/10.1145/3477495.3532007`.

[181] Le Wu, Lei Chen, Pengyang Shao, Richang Hong, Xiting Wang, and Meng Wang. Learning fair representations for recommendation: A graph-based perspective. In *Proceedings of the Web Conference 2021*, WWW '21, page 2198–2208, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450015. URL `https://doi.org/10.1145/3442381.3450015`.

[182] Yao Wu, Jian Cao, Guandong Xu, and Yudong Tan. Tfrom: A two-sided fairness-aware recommendation model for both customers and providers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1013–1022, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462882. URL `https://doi.org/10.1145/3404835.3462882`.

[183] Chen Xu, Sirui Chen, Jun Xu, Weiran Shen, Xiao Zhang, Gang Wang, and Zhenhua Dong. P-mmf: Provider max-min fairness re-ranking in recommender system. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 3701–3711, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394161. doi: 10.1145/3543507.3583296. URL `https://doi.org/10.1145/3543507.3583296`.

[184] Diyi Yang, David Adamson, and Carolyn Penstein Rosé. Question recommendation with constraints for massive open online courses. In Alfred Kobsa, Michelle X. Zhou, Martin Ester, and Yehuda Koren, editors, *Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA - October 06 - 10, 2014*, pages 49–56. ACM, 2014. doi: 10.1145/2645710.2645748. URL `https://doi.org/10.1145/2645710.2645748`.

[185] Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, pages 22:1–22:6. ACM, 2017. doi: 10.1145/3085504.3085526.

[186] Ding Yanhui, Wang Dequan, Zhang Yongxin, and Li Lin. A group recommender system for online course study. In *Proceedings of the 2015 7th International Conference on Information Technology in Medicine and Education (ITME)*, ITME '15, page 318–320, USA, 2015. IEEE Computer Society. ISBN 9781467383028. doi: 10.1109/ITME.2015.99. URL `https://doi.org/10.1109/ITME.2015.99`.

[187] Sirui Yao and Bert Huang. New fairness metrics for recommendation that embrace differences. *CoRR*, abs/1706.09838, 2017. URL `http://arxiv.org/abs/1706.09838`.

[188] Renzhe Yu, Qiujie Li, Christian Fischer, Shayan Doroudi, and Di Xu. Towards accurate and fair prediction of college success: Evaluating different sources of student data. In Anna N. Rafferty, Jacob Whitehill, Cristóbal Romero, and Violetta Cavalli-Sforza, editors, *Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020, Fully virtual conference, July 10-13, 2020*. International Educational Data Mining Society, 2020. URL `https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_194.pdf`.

[189] Guanghu Yuan, Fajie Yuan, Yudong Li, Beibei Kong, Shujie Li, Lei Chen, Min Yang, Chenyun YU, Bo Hu, Zang Li, Yu Xu, and Xiaohu Qie. Tenrec: A large-scale multipurpose benchmark dataset for recommender systems. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 11480–11493. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/4ad4fc1528374422dd7a69dea9e72948-Paper-Datasets_and_Benchmarks.pdf.

[190] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 609–616, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.

[191] Meike Zehlike and Carlos Castillo. Reducing disparate exposure in ranking: A learning to rank approach. In *WWW '20: The Web Conference 2020*, pages 2849–2855. ACM / IW3C2, 2020. doi: 10.1145/3366424.3380048.

[192] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017*, pages 1569–1578. ACM, 2017. doi: 10.1145/3132847.3132938.

[193] Hao Zhang, Heng Yang, Tao Huang, and Gaoqiang Zhan. Dbncf: Personalized courses recommendation system based on dbn in mooc environment. In *2017 International Symposium on Educational Technology (ISET)*, pages 106–108, 2017. doi: 10.1109/ISET.2017.33.

[194] Hao Zhang, Tao Huang, Zhihan Lv, Sanya Liu, and Zhili Zhou. MCRS: A course recommendation system for moocs. *Multim. Tools Appl.*, 77(6):7051–7069, 2018. doi: 10.1007/s11042-017-4620-2. URL https://doi.org/10.1007/s11042-017-4620-2.

[195] Jing Zhang, Bowen Hao, Bo Chen, Cuiping Li, Hong Chen, and Jimeng Sun. Hierarchical reinforcement learning for course recommendation in moocs. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 435–442. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.3301435. URL https://doi.org/10.1609/aaai.v33i01.3301435.

[196] Yong Zheng and David (Xuejun) Wang. A survey of recommender systems with multi-objective optimization. *Neurocomputing*, 474:141–153, 2022. doi: 10.1016/j.neucom.2021.11.041. URL `https://doi.org/10.1016/j.neucom.2021.11.041`.

[197] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010. doi: 10.1073/pnas.1000488107. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1000488107`.

[198] Ziwei Zhu, Xia Hu, and James Caverlee. Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018*, pages 1153–1162. ACM, 2018. doi: 10.1145/3269206.3271795.

[199] Ziwei Zhu, Jianling Wang, and James Caverlee. Measuring and mitigating item under-recommendation bias in personalized ranking systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 449–458, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401177. URL `https://doi.org/10.1145/3397271.3401177`.

[200] Ziwei Zhu, Jingu Kim, Trung Nguyen, Aish Fenton, and James Caverlee. Fairness among new items in cold start recommender systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 767–776, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462948. URL `https://doi.org/10.1145/3404835.3462948`.

[201] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, page 22–32, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595930469. doi: 10.1145/1060745.1060754. URL `https://doi.org/10.1145/1060745.1060754`.

Appendices

Appendix A

Results tables related to Chapter 4

This appendix contains Tables that will help the reader to reproduce the results obtained in our experiments.

Table A.1 Disparate impact in the Movies dataset. Disparate impact metrics returned by the different models for each continent (AF: Africa, AS: Asia, EU: Europe, NA: North America, OC: Oceania, SA: South America) considering the Movies data. For each algorithm, we report the results obtained by the original state-of-the-art algorithm and the binary mitigation proposed in [78], in terms of disparate visibility and exposure when considering the rating-based representation as a reference ($\Delta\mathcal{V}_R$ and $\Delta\mathcal{E}_R$ lines) and with the item-based representation ($\Delta\mathcal{V}_I$ and $\Delta\mathcal{E}_I$ lines). Under each metric, we report the gain or loss we obtained when moving from the original model to the binary mitigation.

		Movies											
		AF		**AS**		**EU**		**NA**		**OC**		**SA**	
		original	**binary**	**original**	**binary**	**original**	**binary**	**original**	**binary**	**original**	**binary**	**original**	**binary**
MostPop	$\Delta\mathcal{V}_R$	0.0031	0.0060	-0.0233	-0.0228	-0.0893	-0.0062	0.1151	0.0237	-0.0053	-0.0003	-0.0003	-0.0003
	(gain/loss)		0.0028		0.0005		0.0831		-0.0914		0.0050		0.0000
	$\Delta\mathcal{E}_R$	0.0018	0.0051	-0.0233	-0.0229	-0.1068	-0.0042	0.1357	0.0233	-0.0070	-0.0009	-0.0003	-0.0003
	(gain/loss)		0.0033		0.0005		0.1026		-0.1124		0.0060		0.0000
	$\Delta\mathcal{V}_I$	0.0022	0.0182	-0.0391	-0.0378	-0.1416	-0.0047	0.1874	0.0133	-0.0064	0.0136	-0.0025	-0.0025
	(gain/loss)		0.0161		0.0012		0.1369		-0.1741		0.0200		0.0000
	$\Delta\mathcal{E}_I$	0.0008	0.0188	-0.0391	-0.0380	-0.1591	-0.0007	0.2080	0.0100	-0.0080	0.0124	-0.0025	-0.0025
	(gain/loss)		0.0180		0.0011		0.1584		-0.1979		0.0205		0.0000
RandomG	$\Delta\mathcal{V}_R$	0.0007	-0.0002	0.0135	0.0036	0.0360	-0.0209	-0.0566	0.0168	0.0033	-0.0012	0.0031	0.0020
	(gain/loss)		-0.0009		-0.0100		-0.0569		0.0734		-0.0045		-0.0011
	$\Delta\mathcal{E}_R$	0.0006	-0.0003	0.0135	0.0036	0.0366	-0.0211	-0.0570	0.0170	0.0033	-0.0011	0.0030	0.0020
	(gain/loss)		-0.0009		-0.0099		-0.0577		0.0740		-0.0044		-0.0011
	$\Delta\mathcal{V}_I$	-0.0003	-0.0003	-0.0022	-0.0027	-0.0163	-0.0191	0.0157	0.0193	0.0022	0.0020	0.0009	0.0009
	(gain/loss)		0.0000		-0.0005		-0.0029		0.0036		-0.0002		0.0000
	$\Delta\mathcal{E}_I$	-0.0004	-0.0004	-0.0023	-0.0029	-0.0157	-0.0191	0.0153	0.0196	0.0022	0.0020	0.0008	0.0008
	(gain/loss)		0.0000		-0.0006		-0.0034		0.0042		-0.0002		-0.0001
UserKNN	$\Delta\mathcal{V}_R$	0.0031	0.0056	-0.0228	-0.0220	-0.0719	-0.0022	0.1003	0.0231	-0.0083	-0.0042	-0.0003	-0.0003
	(gain/loss)		0.0025		0.0008		0.0697		-0.0771		0.0041		0.0000
	$\Delta\mathcal{E}_R$	0.0024	0.0052	-0.0230	-0.0222	-0.0811	-0.0009	0.1113	0.0235	-0.0093	-0.0053	-0.0003	-0.0003
	(gain/loss)		0.0028		0.0008		0.0802		-0.0878		0.0040		0.0000
	$\Delta\mathcal{V}_I$	0.0021	0.0069	-0.0386	-0.0365	-0.1241	0.0049	0.1726	0.0278	-0.0094	-0.0006	-0.0025	-0.0025
	(gain/loss)		0.0048		0.0021		0.1290		-0.1447		0.0087		0.0000
	$\Delta\mathcal{E}_I$	0.0015	0.0065	-0.0388	-0.0368	-0.1333	0.0064	0.1836	0.0286	-0.0104	-0.0021	-0.0025	-0.0025
	(gain/loss)		0.0050		0.0020		0.1397		-0.1550		0.0083		0.0000
ItemKNN	$\Delta\mathcal{V}_R$	0.0006	0.0059	-0.0234	-0.0230	-0.0765	0.0083	0.1117	0.0180	-0.0121	-0.0088	-0.0003	-0.0003
	(gain/loss)		0.0053		0.0004		0.0847		-0.0937		0.0033		0.0000
	$\Delta\mathcal{E}_R$	-0.0003	0.0055	-0.0234	-0.0231	-0.0924	0.0097	0.1288	0.0173	-0.0123	-0.0091	-0.0003	-0.0003
	(gain/loss)		0.0058		0.0003		0.1021		-0.1115		0.0032		0.0000
	$\Delta\mathcal{V}_I$	-0.0004	0.0085	-0.0391	-0.0382	-0.1288	0.0090	0.1840	0.0281	-0.0132	-0.0049	-0.0025	-0.0025
	(gain/loss)		0.0089		0.0010		0.1378		-0.1559		0.0084		0.0000
	$\Delta\mathcal{E}_I$	-0.0013	0.0078	-0.0392	-0.0383	-0.1447	0.0123	0.2011	0.0264	-0.0134	-0.0057	-0.0025	-0.0025
	(gain/loss)		0.0091		0.0009		0.1570		-0.1747		0.0077		0.0000
BPR	$\Delta\mathcal{V}_R$	0.0022	0.0028	-0.0140	-0.0117	-0.0326	-0.0047	0.0436	0.0106	0.0008	0.0029	0.0000	0.0000
	(gain/loss)		0.0006		0.0024		0.0280		-0.0330		0.0020		0.0000
	$\Delta\mathcal{E}_R$	0.0026	0.0033	-0.0152	-0.0129	-0.0357	-0.0050	0.0472	0.0109	0.0011	0.0035	0.0000	0.0001
	(gain/loss)		0.0007		0.0023		0.0308		-0.0363		0.0024		0.0000
	$\Delta\mathcal{V}_I$	0.0012	0.0033	-0.0298	-0.0229	-0.0849	0.0016	0.1159	0.0137	-0.0003	0.0063	-0.0022	-0.0021
	(gain/loss)		0.0021		0.0069		0.0865		-0.1022		0.0066		0.0001
	$\Delta\mathcal{E}_I$	0.0016	0.0038	-0.0310	-0.0243	-0.0880	0.0015	0.1195	0.0142	0.0000	0.0068	-0.0021	-0.0021
	(gain/loss)		0.0022		0.0067		0.0895		-0.1053		0.0069		0.0001
BiasedMF	$\Delta\mathcal{V}_R$	0.0077	0.0026	0.1076	0.0468	0.0674	-0.0322	-0.1728	-0.0060	-0.0108	-0.0113	0.0009	0.0000
	(gain/loss)		-0.0051		-0.0608		-0.0996		0.1669		-0.0005		-0.0009
	$\Delta\mathcal{E}_R$	0.0074	0.0026	0.1245	0.0501	0.0549	-0.0358	-0.1767	-0.0056	-0.0108	-0.0112	0.0006	0.0000
	(gain/loss)		-0.0048		-0.0744		-0.0907		0.1711		-0.0005		-0.0007
	$\Delta\mathcal{V}_I$	0.0067	0.0040	0.0918	0.0594	0.0151	-0.0432	-0.1005	-0.0060	-0.0119	-0.0122	-0.0012	-0.0020
	(gain/loss)		-0.0027		-0.0324		-0.0583		0.0945		-0.0003		-0.0007
	$\Delta\mathcal{E}_I$	0.0064	0.0038	0.1087	0.0685	0.0026	-0.0532	-0.1044	-0.0049	-0.0118	-0.0122	-0.0016	-0.0021
	(gain/loss)		-0.0026		-0.0402		-0.0558		0.0994		-0.0003		-0.0005
SVD++	$\Delta\mathcal{V}_R$	0.0039	0.0012	0.0800	0.0381	0.0686	-0.0343	-0.1433	0.0070	-0.0113	-0.0116	0.0021	-0.0003
	(gain/loss)		-0.0027		-0.0419		-0.1029		0.1503		-0.0004		-0.0024
	$\Delta\mathcal{E}_R$	0.0029	0.0009	0.0954	0.0427	0.0569	-0.0374	-0.1452	0.0058	-0.0114	-0.0117	0.0014	-0.0003
	(gain/loss)		-0.0020		-0.0526		-0.0943		0.1510		-0.0003		-0.0017
	$\Delta\mathcal{V}_I$	0.0029	0.0011	0.0642	0.0440	0.0164	-0.0384	-0.0710	0.0076	-0.0124	-0.0125	-0.0001	-0.0018
	(gain/loss)		-0.0018		-0.0202		-0.0547		0.0786		-0.0001		-0.0017
	$\Delta\mathcal{E}_I$	0.0019	0.0005	0.0796	0.0547	0.0047	-0.0474	-0.0729	0.0069	-0.0125	-0.0127	-0.0008	-0.0021
	(gain/loss)		-0.0014		-0.0249		-0.0521		0.0798		-0.0001		-0.0013

Table A.2 Disparate impact in the Books dataset. Disparate impact metrics returned by the different models for each continent (EU: Europe, NA: North America, OC: Oceania, SA: South America) considering the Books data. For each algorithm, we report the results obtained by the original state-of-the-art algorithm and the binary mitigation proposed in [78], in terms of disparate visibility and exposure when considering the rating-based representation as a reference ($\Delta\mathcal{V}_R$ and $\Delta\mathcal{E}_R$ lines) and with the item-based representation ($\Delta\mathcal{V}_I$ and $\Delta\mathcal{E}_I$ lines). Under each metric, we report the gain or loss we obtained when moving from the original model to the binary mitigation.

		Books							
		Europe		**NA**		**OCE**		**SA**	
		original	**binary**	**original**	**binary**	**original**	**binary**	**original**	**binary**
MostPop	$\Delta\mathcal{V}_R$	-0.0697	0.0102	0.0700	-0.0099	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)		0.0800		-0.0800		0.0000		0.0000
	$\Delta\mathcal{E}_R$	-0.0697	0.0102	0.0700	-0.0099	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)		0.0800		-0.0800		0.0000		0.0000
	$\Delta\mathcal{V}_I$	-0.1042	0.0157	0.1049	-0.0151	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)		0.1199		-0.1199		0.0000		0.0000
	$\Delta\mathcal{E}_I$	-0.1042	0.0157	0.1049	-0.0151	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)		0.1200		-0.1200		0.0000		0.0000
RandomG	$\Delta\mathcal{V}_R$	0.0357	-0.0026	-0.0360	0.0025	0.0003	0.0001	0.0001	0.0000
	(gain/loss)		-0.0383		0.0385		-0.0002		0.0000
	$\Delta\mathcal{E}_R$	0.0356	-0.0028	-0.0359	0.0027	0.0003	0.0001	0.0001	0.0000
	(gain/loss)		-0.0384		0.0386		-0.0002		0.0000
	$\Delta\mathcal{V}_I$	0.0012	-0.0033	-0.0012	0.0033	0.0000	0.0000	0.0000	0.0000
	(gain/loss)		-0.0045		0.0045		0.0000		0.0000
	$\Delta\mathcal{E}_I$	0.0011	-0.0035	-0.0011	0.0035	0.0000	-0.0001	0.0000	0.0000
	(gain/loss)		-0.0046		0.0046		0.0000		0.0000
UserKNN	$\Delta\mathcal{V}_R$	0.0059	0.0057	-0.0057	-0.0055	-0.0002	-0.0001	-0.0001	-0.0001
	(gain/loss)		-0.0002		0.0002		0.0000		0.0000
	$\Delta\mathcal{E}_R$	0.0126	0.0062	-0.0124	-0.0060	-0.0002	-0.0001	-0.0001	-0.0001
	(gain/loss)		-0.0064		0.0064		0.0000		0.0000
	$\Delta\mathcal{V}_I$	-0.0286	0.0088	0.0292	-0.0082	-0.0005	-0.0004	-0.0001	-0.0001
	(gain/loss)		0.0373		-0.0374		0.0000		0.0000
	$\Delta\mathcal{E}_I$	-0.0219	0.0095	0.0225	-0.0089	-0.0005	-0.0004	-0.0001	-0.0001
	(gain/loss)		0.0313		-0.0314		0.0000		0.0000
ItemKNN	$\Delta\mathcal{V}_R$	-0.0273	0.0062	0.0276	-0.0060	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)		0.0335		-0.0336		0.0000		0.0000
	$\Delta\mathcal{E}_R$	-0.0263	0.0065	0.0265	-0.0063	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)		0.0328		-0.0328		0.0000		0.0000
	$\Delta\mathcal{V}_I$	-0.0618	0.0081	0.0624	-0.0075	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)		0.0699		-0.0699		0.0000		0.0000
	$\Delta\mathcal{E}_I$	-0.0607	0.0085	0.0614	-0.0079	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)		0.0692		-0.0693		0.0000		0.0000
BPR	$\Delta\mathcal{V}_R$	0.0252	0.0080	-0.0251	-0.0079	0.0000	-0.0001	0.0000	0.0000
	(gain/loss)		-0.0172		0.0172		0.0000		0.0000
	$\Delta\mathcal{E}_R$	0.0252	0.0081	-0.0251	-0.0080	-0.0001	-0.0001	0.0000	0.0000
	(gain/loss)		-0.0171		0.0171		0.0000		0.0000
	$\Delta\mathcal{V}_I$	-0.0093	0.0114	0.0097	-0.0110	-0.0003	-0.0003	-0.0001	-0.0001
	(gain/loss)		0.0207		-0.0208		0.0000		0.0000
	$\Delta\mathcal{E}_I$	-0.0093	0.0116	0.0097	-0.0112	-0.0004	-0.0003	-0.0001	-0.0001
	(gain/loss)		0.0209		-0.0209		0.0000		0.0000
BiasedMF	$\Delta\mathcal{V}_R$	-0.0698	0.0102	0.0701	-0.0099	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)		0.0800		-0.0800		0.0000		0.0000
	$\Delta\mathcal{E}_R$	-0.0698	0.0102	0.0701	-0.0099	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)		0.0800		-0.0800		0.0000		0.0000
	$\Delta\mathcal{V}_I$	-0.1043	0.0157	0.1049	-0.0151	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)		0.1200		-0.1200		0.0000		0.0000
	$\Delta\mathcal{E}_I$	-0.1043	0.0157	0.1049	-0.0151	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)		0.1200		-0.1200		0.0000		0.0000
SVD++	$\Delta\mathcal{V}_R$	-0.0698	0.0094	0.0701	-0.0091	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)		0.0792		-0.0792		0.0000		0.0000
	$\Delta\mathcal{E}_R$	-0.0698	0.0094	0.0701	-0.0091	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)		0.0792		-0.0792		0.0000		0.0000
	$\Delta\mathcal{V}_I$	-0.1043	0.0141	0.1049	-0.0134	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)		0.1183		-0.1183		0.0000		0.0000
	$\Delta\mathcal{E}_I$	-0.1043	0.0140	0.1049	-0.0134	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)		0.1183		-0.1183		0.0000		0.0000

Table A.3 Disparate impact after mitigation in the Movies dataset. Disparate impact metrics returned by the different models for each continent (AF: Africa, AS: Asia, EU: Europe, NA: North America, OC: Oceania, SA: South America) considering the Movies data. For each algorithm, we report the results obtained by the original algorithm and by our multi-group mitigation, in terms of disparate visibility and exposure when considering the rating-based representation as a reference ($\Delta\mathcal{V}_R$ and $\Delta\mathcal{E}_R$ lines) and with the item-based representation ($\Delta\mathcal{V}_I$ and $\Delta\mathcal{E}_I$ lines). Under each metric, we report the gain or loss we obtained when moving from the original model to our multi-group mitigation.

		Movies											
		AF		**AS**		**EU**		**NA**		**OC**		**SA**	
		original	**multi**	**original**	**multi**	**original**	**multi**	**original**	**multi**	**original**	**multi**	**original**	**multi**
MostPop	$\Delta\mathcal{V}_R$	0.0031	0.0007	-0.0233	-0.0230	-0.0893	0.0001	0.1151	0.0226	-0.0053	0.0000	-0.0003	-0.0003
	(gain/loss)		-0.0025		0.0003		0.0894		-0.0925		0.0053		0.0000
	$\Delta\mathcal{E}_R$	0.0018	-0.0005	-0.0233	-0.0231	-0.1068	-0.0392	0.1357	0.0655	-0.0070	-0.0024	-0.0003	-0.0003
	(gain/loss)		-0.0023		0.0002		0.0676		-0.0702		0.0046		0.0000
	$\Delta\mathcal{V}_I$	0.0022	-0.0005	-0.0391	-0.0389	-0.1416	-0.0053	0.1874	0.0468	-0.0064	0.0003	-0.0025	-0.0025
	(gain/loss)		-0.0026		0.0002		0.1363		-0.1406		0.0067		0.0000
	$\Delta\mathcal{E}_I$	0.0008	-0.0015	-0.0391	-0.0389	-0.1591	-0.0543	0.2080	0.0995	-0.0080	-0.0022	-0.0025	-0.0025
	(gain/loss)		-0.0023		0.0002		0.1047		-0.1085		0.0058		0.0000
RandomG	$\Delta\mathcal{V}_R$	0.0007	0.0000	0.0135	0.0000	0.0360	0.0000	-0.0566	-0.0011	0.0033	0.0000	0.0031	0.0011
	(gain/loss)		-0.0007		-0.0136		-0.0360		0.0555		-0.0033		-0.0019
	$\Delta\mathcal{E}_R$	0.0006	0.0000	0.0135	0.0000	0.0366	-0.0001	-0.0570	-0.0009	0.0033	0.0000	0.0030	0.0010
	(gain/loss)		-0.0006		-0.0135		-0.0367		0.0561		-0.0033		-0.0020
	$\Delta\mathcal{V}_I$	-0.0003	0.0000	-0.0022	0.0000	-0.0163	0.0000	0.0157	0.0000	0.0022	0.0000	0.0009	0.0000
	(gain/loss)		0.0003		0.0022		0.0163		-0.0157		-0.0022		-0.0009
	$\Delta\mathcal{E}_I$	-0.0004	0.0000	-0.0023	0.0000	-0.0157	0.0000	0.0153	0.0000	0.0022	0.0000	0.0008	0.0000
	(gain/loss)		0.0004		0.0023		0.0157		-0.0153		-0.0022		-0.0008
UserKNN	$\Delta\mathcal{V}_R$	0.0031	0.0023	-0.0228	-0.0199	-0.0719	0.0000	0.1003	0.0179	-0.0083	0.0000	-0.0003	-0.0003
	(gain/loss)		-0.0007		0.0029		0.0719		-0.0824		0.0083		0.0000
	$\Delta\mathcal{E}_R$	0.0024	0.0019	-0.0230	-0.0208	-0.0811	-0.0277	0.1113	0.0499	-0.0093	-0.0029	-0.0003	-0.0003
	(gain/loss)		-0.0006		0.0022		0.0534		-0.0614		0.0064		0.0000
	$\Delta\mathcal{V}_I$	0.0021	0.0009	-0.0386	-0.0359	-0.1241	-0.0002	0.1726	0.0377	-0.0094	0.0000	-0.0025	-0.0025
	(gain/loss)		-0.0012		0.0027		0.1239		-0.1349		0.0094		0.0000
	$\Delta\mathcal{E}_I$	0.0015	0.0004	-0.0388	-0.0364	-0.1333	-0.0360	0.1836	0.0768	-0.0104	-0.0023	-0.0025	-0.0025
	(gain/loss)		-0.0011		0.0024		0.0974		-0.1068		0.0081		0.0000
ItemKNN	$\Delta\mathcal{V}_R$	0.0006	0.0000	-0.0234	-0.0184	-0.0765	0.0002	0.1117	0.0185	-0.0121	0.0000	-0.0003	-0.0003
	(gain/loss)		-0.0006		0.0050		0.0767		-0.0932		0.0121		0.0000
	$\Delta\mathcal{E}_R$	-0.0003	-0.0008	-0.0234	-0.0192	-0.0924	-0.0294	0.1288	0.0520	-0.0123	-0.0023	-0.0003	-0.0003
	(gain/loss)		-0.0005		0.0042		0.0630		-0.0768		0.0101		0.0000
	$\Delta\mathcal{V}_I$	-0.0004	-0.0015	-0.0391	-0.0341	-0.1288	0.0000	0.1840	0.0381	-0.0132	0.0000	-0.0025	-0.0025
	(gain/loss)		-0.0011		0.0050		0.1288		-0.1459		0.0132		0.0000
	$\Delta\mathcal{E}_I$	-0.0013	-0.0022	-0.0392	-0.0345	-0.1447	-0.0354	0.2011	0.0764	-0.0134	-0.0018	-0.0025	-0.0025
	(gain/loss)		-0.0009		0.0047		0.1092		-0.1247		0.0117		0.0000
BPR	$\Delta\mathcal{V}_R$	0.0022	0.0000	-0.0140	-0.0001	-0.0326	0.0000	0.0436	0.0000	0.0008	0.0000	0.0000	0.0001
	(gain/loss)		-0.0022		0.0140		0.0326		-0.0436		-0.0008		0.0001
	$\Delta\mathcal{E}_R$	0.0026	0.0009	-0.0152	-0.0041	-0.0357	-0.0086	0.0472	0.0113	0.0011	0.0003	0.0000	0.0002
	(gain/loss)		-0.0017		0.0110		0.0271		-0.0359		-0.0007		0.0001
	$\Delta\mathcal{V}_I$	0.0012	0.0001	-0.0298	0.0000	-0.0849	0.0000	0.1159	0.0000	-0.0003	0.0002	-0.0022	-0.0003
	(gain/loss)		-0.0011		0.0298		0.0849		-0.1159		0.0005		0.0018
	$\Delta\mathcal{E}_I$	0.0016	0.0005	-0.0310	-0.0065	-0.0880	-0.0117	0.1195	0.0183	0.0000	0.0001	-0.0021	-0.0007
	(gain/loss)		-0.0012		0.0244		0.0763		-0.1012		0.0001		0.0015
BiasedMF	$\Delta\mathcal{V}_R$	0.0077	0.0000	0.1076	0.0000	0.0674	0.0000	-0.1728	0.0000	-0.0108	0.0000	0.0009	0.0000
	(gain/loss)		-0.0077		-0.1076		-0.0674		0.1729		0.0108		-0.0010
	$\Delta\mathcal{E}_R$	0.0074	0.0001	0.1245	0.0009	0.0549	0.0000	-0.1767	0.0000	-0.0108	-0.0009	0.0006	-0.0001
	(gain/loss)		-0.0073		-0.1236		-0.0549		0.1767		0.0098		-0.0007
	$\Delta\mathcal{V}_I$	0.0067	0.0003	0.0918	0.0005	0.0151	0.0000	-0.1005	0.0000	-0.0119	0.0000	-0.0012	-0.0007
	(gain/loss)		-0.0065		-0.0913		-0.0151		0.1005		0.0119		0.0005
	$\Delta\mathcal{E}_I$	0.0064	0.0003	0.1087	0.0015	0.0026	0.0000	-0.1044	0.0000	-0.0118	-0.0011	-0.0016	-0.0007
	(gain/loss)		-0.0061		-0.1072		-0.0026		0.1044		0.0107		0.0009
SVD++	$\Delta\mathcal{V}_R$	0.0039	0.0004	0.0800	0.0000	0.0686	0.0000	-0.1433	0.0000	-0.0113	-0.0004	0.0021	0.0000
	(gain/loss)		-0.0034		-0.0800		-0.0686		0.1433		0.0109		-0.0021
	$\Delta\mathcal{E}_R$	0.0029	0.0005	0.0954	0.0000	0.0569	0.0000	-0.1452	0.0000	-0.0114	-0.0004	0.0014	-0.0001
	(gain/loss)		-0.0024		-0.0954		-0.0569		0.1452		0.0111		-0.0015
	$\Delta\mathcal{V}_I$	0.0029	0.0000	0.0642	0.0000	0.0164	0.0000	-0.0710	0.0000	-0.0124	0.0000	-0.0001	0.0000
	(gain/loss)		-0.0029		-0.0642		-0.0164		0.0710		0.0124		0.0001
	$\Delta\mathcal{E}_I$	0.0019	0.0000	0.0796	0.0019	0.0047	0.0001	-0.0729	0.0000	-0.0125	-0.0019	-0.0008	-0.0001
	(gain/loss)		-0.0019		-0.0777		-0.0046		0.0729		0.0106		0.0007

Table A.4 Disparate impact after mitigation in the Books dataset. Disparate impact metrics returned by the different models for each continent (EU: Europe, NA: North America, OC: Oceania, SA: South America) considering the Books data. For each algorithm, we report the results obtained by the original algorithm and by our multi-group mitigation, in terms of disparate visibility and exposure when considering the rating-based representation as a reference ($\Delta\mathcal{V}_R$ and $\Delta\mathcal{E}_R$ lines) and with the item-based representation ($\Delta\mathcal{V}_I$ and $\Delta\mathcal{E}_I$ lines). Under each metric, we report the gain or loss we obtained when moving from the original model to our multi-group mitigation.

		Books							
		EU		**NA**		**OC**		**SA**	
		original	**multi**	**original**	**multi**	**original**	**multi**	**original**	**multi**
MostPop	$\Delta\mathcal{V}_R$	-0.0697	0.0000	0.0700	0.0003	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)		0.0697		-0.0697		0.0000		0.0000
	$\Delta\mathcal{E}_R$	-0.0697	-0.0227	0.0700	0.0230	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)		0.0471		-0.0471		0.0000		0.0000
	$\Delta\mathcal{V}_I$	-0.1042	0.0000	0.1049	0.0006	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)		0.1042		-0.1042		0.0000		0.0000
	$\Delta\mathcal{E}_I$	-0.1042	-0.0322	0.1049	0.0328	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)		0.0720		-0.0720		0.0000		0.0000
RandomG	$\Delta\mathcal{V}_R$	0.0357	0.0000	-0.0360	0.0000	0.0003	0.0000	0.0001	0.0000
	(gain/loss)		-0.0357		0.0360		-0.0003		0.0000
	$\Delta\mathcal{E}_R$	0.0356	0.0000	-0.0359	0.0000	0.0003	0.0000	0.0001	0.0000
	(gain/loss)		-0.0356		0.0359		-0.0003		-0.0001
	$\Delta\mathcal{V}_I$	0.0012	0.0000	-0.0012	0.0000	0.0000	0.0000	0.0000	0.0000
	(gain/loss)		-0.0012		0.0012		0.0000		0.0000
	$\Delta\mathcal{E}_I$	0.0011	0.0000	-0.0011	0.0000	0.0000	0.0000	0.0000	0.0000
	(gain/loss)		-0.0011		0.0011		0.0000		0.0000
UserKNN	$\Delta\mathcal{V}_R$	0.0059	0.0001	-0.0057	0.0000	-0.0002	-0.0001	-0.0001	-0.0001
	(gain/loss)		-0.0058		0.0057		0.0001		0.0000
	$\Delta\mathcal{E}_R$	0.0126	0.0001	-0.0124	0.0000	-0.0002	-0.0001	-0.0001	-0.0001
	(gain/loss)		-0.0125		0.0124		0.0001		0.0000
	$\Delta\mathcal{V}_I$	-0.0286	0.0000	0.0292	0.0004	-0.0005	-0.0002	-0.0001	-0.0001
	(gain/loss)		0.0286		-0.0288		0.0002		0.0000
	$\Delta\mathcal{E}_I$	-0.0219	0.0000	0.0225	0.0004	-0.0005	-0.0003	-0.0001	-0.0001
	(gain/loss)		0.0219		-0.0220		0.0002		0.0000
ItemKNN	$\Delta\mathcal{V}_R$	-0.0273	0.0000	0.0276	0.0002	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)		0.0273		-0.0273		0.0000		0.0000
	$\Delta\mathcal{E}_R$	-0.0263	-0.0032	0.0265	0.0034	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)		0.0231		-0.0231		0.0000		0.0000
	$\Delta\mathcal{V}_I$	-0.0618	0.0000	0.0624	0.0006	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)		0.0618		-0.0618		0.0000		0.0000
	$\Delta\mathcal{E}_I$	-0.0607	-0.0121	0.0614	0.0127	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)		0.0486		-0.0486		0.0000		0.0000
BPR	$\Delta\mathcal{V}_R$	0.0252	0.0000	-0.0251	0.0000	0.0000	0.0000	0.0000	0.0000
	(gain/loss)		-0.0252		0.0251		0.0000		0.0000
	$\Delta\mathcal{E}_R$	0.0252	0.0000	-0.0251	0.0000	-0.0001	0.0000	0.0000	0.0000
	(gain/loss)		-0.0252		0.0251		0.0001		0.0000
	$\Delta\mathcal{V}_I$	-0.0093	0.0000	0.0097	0.0000	-0.0003	0.0000	-0.0001	0.0000
	(gain/loss)		0.0093		-0.0098		0.0003		0.0001
	$\Delta\mathcal{E}_I$	-0.0093	0.0000	0.0097	0.0001	-0.0004	-0.0001	-0.0001	0.0000
	(gain/loss)		0.0093		-0.0096		0.0003		0.0001
BiasedMF	$\Delta\mathcal{V}_R$	-0.0698	0.0000	0.0701	0.0003	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)		0.0698		-0.0698		0.0000		0.0000
	$\Delta\mathcal{E}_R$	-0.0698	-0.0215	0.0701	0.0218	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)		0.0483		-0.0483		0.0000		0.0000
	$\Delta\mathcal{V}_I$	-0.1043	0.0000	0.1049	0.0006	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)		0.1043		-0.1043		0.0000		0.0000
	$\Delta\mathcal{E}_I$	-0.1043	-0.0296	0.1049	0.0302	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)		0.0747		-0.0747		0.0000		0.0000
SVD++	$\Delta\mathcal{V}_R$	-0.0698	0.0000	0.0701	0.0003	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)		0.0698		-0.0698		0.0000		0.0000
	$\Delta\mathcal{E}_R$	-0.0698	-0.0213	0.0701	0.0216	-0.0002	-0.0002	-0.0001	-0.0001
	(gain/loss)		0.0485		-0.0485		0.0000		0.0000
	$\Delta\mathcal{V}_I$	-0.1043	0.0000	0.1049	0.0006	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)		0.1043		-0.1043		0.0000		0.0000
	$\Delta\mathcal{E}_I$	-0.1043	-0.0296	0.1049	0.0302	-0.0005	-0.0005	-0.0001	-0.0001
	(gain/loss)		0.0747		-0.0747		0.0000		0.0000