UNIVERSITY OF COPENHAGEN
DEPARTMENT OF COMPUTER SCIENCE

PhD Thesis

Theresia Veronika Rampisela

Offline Evaluation Measures of Fairness in Recommender Systems

Principal supervisor: Christina Lioma
Co-supervisors: Tuukka Ruotsalo and Maria Maistro

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen on 31 May 2025.

Abstract

The evaluation of recommender system fairness has become increasingly important, especially with recent legislation that emphasises the development of fair and responsible artificial intelligence. This increasing importance has led to the emergence of various fairness evaluation measures, which quantify fairness based on different fairness definitions. However, many of such measures are simply proposed and further used without further analysis on their robustness. As a result, there is insufficient understanding and awareness of the measures' limitations. Among other issues, it is not known what kind of model outputs produce the (un)fairest score, how the measure scores are empirically distributed, and whether there are cases where the measures cannot be computed (e.g., due to division by zero). These issues cause difficulty in interpreting the measure scores and confusion on which measure(s) should be used for a specific case.

To address the above issues, this thesis presents a series of papers that assess and overcome various theoretical, empirical, and conceptual limitations of existing recommender system fairness evaluation measures. Altogether, the papers in this thesis investigate a wide range of offline evaluation measures for different fairness notions, which are divided based on the evaluation subjects (users and items) and for different evaluation granularities (groups of subjects and individual subjects). Firstly, we perform theoretical and empirical analysis on a set of fairness measures, exposing flaws that limit their interpretability, expressiveness, or applicability. Secondly, we contribute novel evaluation approaches and measures that overcome these limitations. Finally, considering the measures' limitations, we recommend guidelines for the appropriate measure usage, thereby allowing for more precise selection of fairness evaluation measures in practical scenarios.

Overall, the papers in this thesis collectively contribute to advancing the state-of-the-art offline evaluation of fairness in recommender systems.

Resumé

Evaluering af fairness (retfærdighed) i anbefalingssystemer er blevet stadig vigtigere, især i lyset af ny lovgivning, der understøtter udviklingen af fair og ansvarlig kunstig intelligens. Denne stigende opmærksomhed har ført til fremkomsten af forskellige fairness-evalueringsmetoder, som kvantificerer fairness baseret på forskellige definitioner. Imidlertid er mange af disse metoder blevet foreslået og anvendt uden en dybere analyse af deres robusthed. Dette har resulteret i en utilstrækkelig forståelse og bevidsthed om metoderne og deres begrænsninger. Blandt andet er det uklart, hvilke modeloutput der producerer de mest (un)fair scores, hvordan målingerne empirisk fordeler sig, og om der findes situationer, hvor målingerne ikke kan beregnes (f.eks. ved division med nul). Disse udfordringer vanskeliggør fortolkningen af målingernes resultater og skaber uklarhed om, hvilke metoder der bør anvendes i specifikke tilfælde.

For at adressere disse udfordringer præsenterer denne afhandling en serie af artikler, der vurderer og løser forskellige teoretiske, empiriske og konceptuelle begrænsninger ved eksisterende fairness-evalueringsmetoder i anbefalingssystemer. Artiklerne undersøger et bredt udvalg af offline evalueringsmetoder, opdelt efter evalueringssubjekter (brugere og genstande) samt på forskellige evalueringsniveauer (grupper af subjekter og individuelle subjekter). Først udføres en teoretisk og empirisk analyse af et sæt fairness-målinger, hvor der identificeres begrænsninger, der påvirker deres fortolkningsevne, udtrykskraft og anvendelighed. Derefter introduceres nye evalueringsmetoder, som overvinder disse begrænsninger. Endelig anbefales retningslinjer for korrekt anvendelse af fairness-målinger, hvilket muliggør en mere præcis udvælgelse af passende evalueringsmetoder i praktiske scenarier.

Samlet set bidrager artiklerne i denne afhandling til at fremme den nyeste udvikling inden for offline evaluering af fairness i anbefalingssystemer.

List of Publications

This thesis includes the following papers as chapters, listed in the order of their appearance in the thesis.[1]

1. Theresia Veronika Rampisela, Maria Maistro, Tuukka Ruotsalo, and Christina Lioma. Evaluation measures of individual item fairness for recommender systems: A critical study. *ACM Trans. Recomm. Syst.*, 3(2):1–52, November 2024. doi: 10.1145/3631943. URL `https://doi.org/10.1145/3631943`. 🏆 *Women in RecSys Journal Paper of the Year Award 2024, Junior Category.*

2. Theresia Veronika Rampisela, Tuukka Ruotsalo, Maria Maistro, and Christina Lioma. Can We Trust Recommender System Fairness Evaluation? The Role of Fairness and Relevance. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 271–281, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657832. URL `https://doi.org/10.1145/3626772.3657832`. *Acceptance rate: 20%*

3. Theresia Veronika Rampisela, Maria Maistro, Tuukka Ruotsalo, Falk Scholer, and Christina Lioma. Relevance-aware individual item fairness measures for recommender systems: Limitations and usage guidelines. *ACM Trans. Recomm. Syst.*, September 2025. doi: 10.1145/3765624. URL `https://doi.org/10.1145/3765624`.

4. Theresia Veronika Rampisela, Tuukka Ruotsalo, Maria Maistro, and Christina Lioma. Joint Evaluation of Fairness and Relevance in Recommender Systems with Pareto Frontier. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 1548–1566, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714589. URL

[1] The publication status is updated as of the defence date.

https://doi.org/10.1145/3696410.3714589. *Acceptance rate (oral presentation): 7.5%,* *ACM Artifacts Available v1.1 badge.*

5. Theresia Veronika Rampisela, Maria Maistro, Tuukka Ruotsalo, and Christina Lioma. Measuring Individual User Fairness with User Similarity and Effectiveness Disparity. 2025. 10 pages. *Under review.*

6. Theresia Veronika Rampisela, Maria Maistro, Tuukka Ruotsalo, Falk Scholer, and Christina Lioma. Stairway to fairness: Connecting group and individual fairness. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, RecSys '25, page 677–683, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713644. doi: 10.1145/3705328.3748031. URL https://doi.org/10.1145/3705328.3748031.

Acknowledgements

This PhD thesis and the entirety of my PhD journey would not have been possible without the support of many people. I would like to acknowledge them here.

First and foremost, I am extremely grateful for my PhD supervisors. From the start to the end of my studies, they have provided me with invaluable scientific and professional guidance, critical and detailed feedback, as well as many opportunities to develop my career, all of which I deeply appreciated. I have learned a lot from them. To Christina, thank you for consistently emphasising the importance of conducting high-quality scientific research and for showing how to explain complex concepts so that they can be easily understood. I am also thankful that you helped me understand the "why's" behind certain practices, in a safe environment. To Tuukka, I am grateful for your fresh ideas and perspectives, which have been greatly useful. Thank you for always encouraging me to look at the bigger picture. To Maria, thank you for your technical and mathematical expertise, as well as for your help in translating abstract ideas and thoughts into clarity. Thank you for proactively facilitating my participation in activities that are relevant to my professional growth.

I sincerely thank the Algorithms, Data & Democracy (ADD) project for fully funding my PhD. Being part of the project has allowed me to meet wonderful researchers from various disciplines, providing me with an inter-disciplinary collaboration experience and broadening my perspective on the societal impacts of algorithms.

To the Information Retrieval (IR) Lab at the University of Copenhagen, thank you for the enriching presentations and discussions, as well as for the 3Cs: (paper) clinics, chats, and cakes. A special thanks goes to Vadym, Pietro, and Ervin for being a significant part of my PhD journey. Thank you for being there for me through the ups and downs! Another special thanks goes to Qiuchi and Sara, with whom I had the pleasure to share an office and hold insightful conversations.

During my studies, I had the pleasure of visiting Falk Scholer and the IR research group at RMIT University in Melbourne. This visit has given me new, invaluable insights into the diverse spectrum of IR evaluation research.

Next, I would like to thank my parents for financially supporting my education up to my master's degree and for ensuring that I receive a good education. To my sister, thank you for being my moral support from 10,000 km away.

To July, Kim, and Morgan, thank you for letting me stay at your lovely home during my research visit in Melbourne, and for making me feel welcome in your family. To the Indonesian Student Association in Denmark, thank you for helping me navigate life in a foreign country. I am also thankful for the Mensa Foundation, which has awarded me an additional scholarship in relation to my PhD studies. I would also like to acknowledge Chloé Rouyer, who has provided a convenient LaTeX template that I have used for this thesis.

Finally, I thank my colleagues and friends, whom I cannot possibly name one by one, for their various forms of support and encouragement that I have received throughout this journey.

Table of Contents

Abstract ii
Resumé iii
List of Publications iv
Acknowledgements vi

1 Executive Summary **1**
1.1 Introduction 1
1.2 Theoretical Background 5
1.2.1 Overview of Recommender Systems 5
1.2.2 Fairness in Recommender Systems 8
1.2.3 Offline Evaluation of Recommender System Fairness 12
1.3 Scientific Contributions 18
1.3.1 Paper 1: Evaluation Measures of Individual Item Fairness for Recommender Systems: A Critical Study 19
1.3.2 Paper 2: Can We Trust Recommender System Fairness Evaluation? The Role of Fairness and Relevance 21
1.3.3 Paper 3: Relevance-aware Individual Item Fairness Measures in Recommender Systems: Limitations and Usage Guidelines 21
1.3.4 Paper 4: Joint Evaluation of Fairness and Relevance in Recommender Systems with Pareto Frontier 23
1.3.5 Paper 5: Measuring Individual User Fairness with User Similarity and Effectiveness Disparity 25
1.3.6 Paper 6: Stairway to Fairness: Connecting Group and Individual Fairness 26
1.4 Summary and Future Work 28

2 Evaluation Measures of Individual Item Fairness for Recommender Systems: A Critical Study **32**
2.1 Introduction 33

2.2 Individual Item Fairness Measures 35
2.2.1 Notation and Examination Functions 36
2.2.2 Measures of Individual Item Fairness 36
2.3 Measure Limitations 42
2.3.1 Limitation 1: Non-realisability 42
2.3.2 Limitation 2: Quantity-insensitivity (QF) 45
2.3.3 Limitation 3: Undefinedness (Ent) 45
2.3.4 Limitation 4: Always-fair (↑FSat) 47
2.3.5 Limitation 5: Item-representation-dependence (↓VoCD) 47
2.4 Resolving Limitations 48
2.4.1 Resolving Non-realisability (Limitation 1) and Undefinedness (Limitation 3) 48
2.4.2 Resolving Quantity-insensitivity (Limitation 2) 52
2.4.3 Unresolvable Limitations 52
2.5 Empirical Analysis 53
2.5.1 Experimental Setup 53
2.5.2 Analysis of Relevance and Fairness 54
2.5.3 Correlation between Measures 58
2.5.4 Max/min Achievable Fairness 62
2.5.5 Sliding Window: Relevance and Fairness at Different Rank Positions 65
2.5.6 Measure Strictness and Sensitivity through Artificial Insertion of Items 68
2.6 Related Work 71
2.7 Discussion 72
2.7.1 Summary of Theoretical Corrections 72
2.7.2 Summary of Empirical Findings 73
2.7.3 Guidelines of the Appropriate Use of the Fairness Measures . 74
2.8 Conclusions 75
2.9 Appendix 77
2.9.1 Mathematical Workings for Bounds in Tab. 2.5 77
2.9.2 Extended Results of Experiments 82

3 Can We Trust Recommender System Fairness Evaluation? The Role of Fairness and Relevance 101
3.1 Introduction 102
3.2 Individual Item Fairness & Relevance 103
3.2.1 Notation and Definitions 103

3.2.2 Joint Measures of Fairness and Relevance 104
3.3 Experimental Setup 110
3.4 Empirical Analysis 111
3.4.1 Evaluation Results of All Measures 112
3.4.2 Correlation between Measures (RQ1 & RQ2) 113
3.4.3 Measure Sensitivity at Different Ranks (RQ3) 117
3.4.4 Artificial Insertion of Items (RQ4) 117
3.5 Related Work 119
3.6 Appropriate Usage of Joint Measures 121
3.7 Conclusions and Future Work 123

4 Relevance-aware Individual Item Fairness Measures in Recommender Systems: Limitations and Usage Guidelines 124
4.1 Introduction 125
4.2 Evaluation Measures of Fairness for Individual Items 126
4.2.1 Notation and Examination Functions 126
4.2.2 Exposure-based Fairness Measures 127
4.2.3 Relevance-aware Fairness Measures 128
4.3 Measure Limitations 134
4.3.1 Limitation 1: Non-realisability 135
4.3.2 Limitation 2: Non-localisation (↓IAA, ↓IFD$_{\div}$, ↓HD, ↑IBO/↓IWO, ↓II-F, ↓AI-F) 138
4.3.3 Limitation 3: Undefinedness (IAA, IBO/IWO) 139
4.3.4 Limitation 4: Zero-exposure (↓IAA) 140
4.3.5 Limitation 5: Top-k-insensitivity (↓IFD$_{\div}$) 141
4.4 Resolving Limitations 141
4.4.1 Resolving Non-realisability (Limitation 1) and Top-k-insensitivity (Limitation 5) 142
4.4.2 Resolving Undefinedness (Limitation 2) and Zero-exposure (Limitation 4) 147
4.4.3 Unresolvable Limitations 148
4.5 Experimental Setup 149
4.6 Empirical Analysis 152
4.6.1 Evaluation Results of All Measures 152
4.6.2 Measure Correlations 157
4.6.3 Achievable Fairness Scores (Related to Non-realisability) . . . 160
4.6.4 Sensitivity to Cut-off k (Related to Top-k-insensitivity) 164

4.6.5 Effect of Relevant Items Beyond the Top-k (Related to Non-localisation) . . . 166
4.6.6 Sensitivity to Varying Recommendation Effectiveness and Item Exposure Distribution . . . 170
4.7 Guidelines for Measure Usage . . . 174
4.8 Related Work . . . 180
4.9 Conclusion . . . 181
4.10 Appendix . . . 182
4.10.1 Mathematical workings for ranking strategies to achieve the (un)fairest JOINT scores . . . 182

5 Joint Evaluation of Fairness and Relevance in Recommender Systems with Pareto Frontier 187
5.1 Introduction . . . 188
5.2 Related Work . . . 190
5.3 Distance to Pareto Frontier (DPFR) . . . 190
5.3.1 Definitions . . . 191
5.3.2 Pareto Frontier generation . . . 191
5.3.3 Distance computation . . . 193
5.3.4 Efficient computation of Pareto Frontier . . . 193
5.4 Experimental Setup . . . 194
5.5 Experimental Results . . . 197
5.5.1 Groundwork Runs . . . 197
5.5.2 Measure Compatibility with DPFR . . . 197
5.5.3 The Generated PF . . . 199
5.5.4 Agreement between Measures . . . 200
5.5.5 Efficient DPFR . . . 204
5.6 Discussion and Conclusions . . . 206
5.7 Appendix . . . 208
5.7.1 Extended Dataset Statistics . . . 208
5.7.2 Algorithms for Generating Pareto Frontier . . . 208
5.7.3 Modifications to the GS Algorithm . . . 213
5.7.4 Extended Results . . . 214
5.7.5 Further Discussions . . . 214

6 Measuring Individual User Fairness with User Similarity and Effectiveness Disparity 223
6.1 Introduction . . . 224

6.2 Individual User Fairness . . . 225
6.2.1 Definition of Individual User Fairness . . . 225
6.2.2 Current Individual User Fairness Measures . . . 226
6.3 Pairwise User unFairness (PUF) . . . 228
6.4 Empirical Analysis . . . 230
6.4.1 Experimental Setup . . . 230
6.4.2 Comparison of All Evaluation Measures . . . 231
6.4.3 Measure Agreement . . . 233
6.4.4 Varying Relevance Score Distribution . . . 235
6.4.5 Varying User Similarity Distribution . . . 236
6.4.6 PUF and UF under Extreme Cases . . . 238
6.4.7 Using the Same Similarity Measure for PUF and Recommendation Model . . . 239
6.4.8 Using Item Features for PUF and UF . . . 240
6.5 Related Work . . . 241
6.6 Discussion and Conclusions . . . 242
6.7 Appendix . . . 243
6.7.1 Extended Empirical Analysis . . . 243

7 Stairway to Fairness: Connecting Group and Individual Fairness 249
7.1 Introduction . . . 249
7.2 Methodology . . . 250
7.3 Empirical Analysis . . . 254
7.4 Conclusion . . . 260
7.5 Appendix . . . 260
7.5.1 Details on Experimental Setup . . . 260
7.5.2 Extended Results . . . 265

Errata 275

Bibliography 276

Chapter 1

Executive Summary

1.1 Introduction

Recommender systems are a type of Artificial Intelligence (AI) that help their users discover relevant items in the form of content, products, services, and so forth based on the users' preferences, needs, or interests [4]. These systems are highly useful in both professional and daily activities, with a wide range of applications. Some common applications include recommending job positions to apply for [206], suggesting news articles to read [109], and providing ideas for songs, videos, or movies as entertainment options [46, 94, 197].

Alongside the development of methods for improving recommendation quality, various ways of evaluating recommender systems have been proposed [91, 248]. The purpose of evaluation is not only to assess the systems' effectiveness, but also to capture additional important aspects, such as novelty and serendipity [98]. This thesis focuses on one particular 'beyond effectiveness' evaluation aspect: fairness.

Ensuring fairness in AI applications, such as recommender systems, is crucial and even compulsory according to legislations such as the EU AI Act [69]. While the overall idea of fairness is generally to avoid disadvantaging individuals or specific groups of entities, there is no universal definition of fairness. Even in the specific context of recommender systems, many fairness notions arise from what different stakeholders may perceive as 'fair' [57, 74, 207, 209, 211, 228] (see §1.2.2.1). Recommender system stakeholders include the consumers and the item providers, which are associated with *user*- and *item*-side fairness, respectively [28]. User-side fairness typically concerns disparity in the users' recommendation effectiveness, which has to do with ensuring that all users are equally satisfied with their recommendations [126]. On the other hand, item-side fairness is usually based on the exposure received by the items, which

is about recommending items to users in a way that would benefit the items or their providers [65]. For instance, recommending a product to users may increase their awareness of the product, benefiting its seller in the form of advertisement. Fairness can also be evaluated at different levels of granularity: for *groups* of users/items and for the *individual* subjects [228]. Group fairness usually emphasises that different groups of users/items should receive similar treatment, or have a similar experience [65, 228]. For example, users from different age groups should get recommendations that are equally good [64]. In contrast, individual fairness focuses more on providing equal treatment for similar users or items [227, 235].

The deeply contextual nature of fairness, along with the inherent limitation of not being able to label fairness for a given data instance without considering other data instances or additional information, leads to the difficulty in evaluating fairness. Most research on recommender system fairness rely on offline experiments [54], where one or more evaluation measures[1] are computed as a proxy of how fair a system is. There is a multitude of such measures under numerous fairness concepts, which can be categorised according to their evaluation granularity or the fairness subject (e.g., users or items), or other grouping criteria [228]. The sheer number of recommender system fairness evaluation measures, the various fairness definitions they follow, and other factors such as how different their formulations can be, collectively make it challenging to understand their appropriate usage [144, 188]. These factors motivate a more careful selection of these measures by determining the most suitable measure(s) for specific use cases [221], considering their practical constraints [208].

Generally, the strengths and pitfalls of an evaluation measure can be unveiled through theoretical and empirical analyses [26, 137]. Despite the existence of numerous surveys on recommender system fairness [1, 9, 54, 133, 173, 209, 222, 228, 237, 252, 261], there is a lack of studies that theoretically dissect the fairness evaluation measures or empirically compare them. At the same time, several measures are proposed without clear analysis on their limitations. Moreover, there is no clear standard on how to assess the goodness of recommender system fairness evaluation measures (§1.2.3.2). This leads to several problems:

1. There is insufficient understanding of how recommender system fairness measures quantify fairness. Often, the specific kind of recommendation lists producing the (un)fairest score is unknown, and sometimes, even the theoretical measure range is unclear as it is not explicitly stated (e.g., [103, 191, 242]) and it is not obvious from the mathematical formulation. Suppose that a measure

[1]Though some may refer to evaluation measures as *metrics*, this thesis distinguishes them to avoid confusion with distance metrics, which must satisfy additional properties [159].

is bounded at 0, and that this is the fairest score. If its upper bound (the most unfair score) is unknown, it is not possible to understand whether a score of 50 is closer to the most unfair score or to the fairest. If the upper bound is 100, then 50 can be deemed moderately fair. However, if the upper bound is 1000 instead, 50 would seem highly fair. Without knowing the measure's range or what the 'ideal' output looks like, the score interpretation is highly ambiguous. This necessitates a look into the measures' **theoretical limitations**.

2. Several empirical properties of recommender system fairness measures are poorly understood. One such property is related to how the measure scores are distributed in their range. For example, two measures with a $[0, 1]$-range score the same set of systems conflictingly [126, 264]: one has a compressed range of $[0, 0.04]$, while the other has a wider range of $[0.6, 1]$. Considering that 0 is the fairest score and 1 is the unfairest score, the systems could be seen as extremely fair based on the first measure, but moderately to highly unfair based on the second measure. The difference in the measures' observed empirical range may cause confusion in their score interpretation. Knowing the measures' expressiveness is important, as a measure with a compressed empirical range could give the impression that two systems are comparably fair when they are not. Hence, such **empirical limitations** also need to be investigated.

3. Recommender system fairness measures may not quantify the intended aspects. For instance, individual fairness is generally understood as providing equal treatment to *similar* individuals [63], but most evaluation measures for individual fairness in recommender systems simply account for disparity or variation across all individuals, without considering their similarity. To ensure that the measures account for the intended aspects [15], this type of **conceptual limitations** must be addressed.

Objectives. The main objectives of this thesis are:

1. To assess the theoretical, empirical, and conceptual limitations of existing recommender system fairness evaluation measures;

2. To design new evaluation approaches or measures that resolve the above limitations; and

3. To recommend the appropriate measure usage, considering their limitations.

Overall, this thesis aims to improve state-of-the-art offline evaluation measures of fairness in recommender systems.

Table 1.1: The three fairness types covered by the six papers included as chapters in this thesis (P1–P6). Fairness is categorised based on the combination of evaluation granularity (individuals vs. groups) and evaluation subject (items vs. users). Fairness for groups of items is not included as it has been extensively studied [175, 194, 198].

Fairness type	P1	P2	P3	P4	P5	P6
Individual items	✓	✓	✓	✓		
Individual users					✓	✓
Groups of users						✓

Scope. The scope of the papers that are part of this thesis is as follows. This thesis examines existing offline fairness evaluation measures in recommender systems, starting from individual item fairness (*Papers 1–4* Chapters 2–5) and continues with studying user-side fairness for individuals (*Paper 5*, Chapter 6) and for groups (*Paper 6*, Chapter 7). Individual item fairness can be evaluated by considering only item exposure (e.g., how many times it gets recommended to users) or jointly with item relevance (e.g., whether the item suits the users' needs or preferences). To avoid complexity arising from having item relevance as an additional variable, we first study measures that account only for exposure (*Paper 1*, Chapter 2) separately from those that account for both exposure and relevance (*Papers 2–4*, Chapters 3–5). The fairness types evaluated and investigated in this thesis are summarised in Tab. 1.1.

In summary, all papers that are part of this thesis evaluate fairness in an offline setting. Altogether, the papers cover fairness for both subjects (user and item fairness), as well as for both granularity levels (group and individual fairness). Furthermore, all evaluations are based on the model output (i.e., the recommendation lists) instead of the process, e.g., based on how the users or items are represented by a model's internal mechanism [129, 131, 233]). For the most part, the focus is on evaluation at a fixed point in time, rather than over a time period [19].

While we have covered three common types of fairness within the above-mentioned scopes, several types are excluded. Firstly, item-side group fairness evaluation measures are excluded as they have been extensively studied in the broader field of ranked outputs (e.g., [175, 194, 198]). Additionally, when studying individual fairness measures, we have excluded measures of counterfactual fairness and fair representation learning for recommendation, as they are not computed based on the recommendation list (e.g., [131, 201]). We have also excluded (item) subject fairness, as it is highly similar to item-side fairness [65, 118]. Finally, our studies also exclude case-specific

fairness measures due to their limited applicability to the general recommendation task. For example, some measures are precisely designed only for fairness among items with little historical data [268] or only for group recommendations, where the task is to recommend the same item(s) to a group of users [190, 199, 238].

Chapter Outline. The rest of this chapter is organised as follows. Section 1.2 provides an overview of recommender systems, an introduction to fairness in recommender systems, as well as the theoretical background for offline evaluation measures of recommender system fairness and their limitations. Section 1.3 presents a detailed overview of the contributions made by the papers included as chapters in this thesis. Section 1.4 summarises the contributions of this thesis and discusses promising directions for future research.

1.2 Theoretical Background

This section provides an overview of recommender systems (§1.2.1), an introduction to fairness in recommender systems (§1.2.2), as well as the theoretical background on offline evaluation of recommender system fairness that is necessary for understanding the subsequent chapters of this thesis (§1.2.3). Cross-references to the papers included in this thesis are made when relevant.

1.2.1 Overview of Recommender Systems

The overview of recommender systems is explained in three parts: the general recommendation task (§1.2.1.1), the basic and advanced approaches in recommendation (§1.2.1.2), as well as how recommender systems are evaluated (§1.2.1.3).

1.2.1.1 The Recommendation Task

The main goal of recommender systems is to help users find the most appealing items that would match their preferences, suit their needs, and so on, from a (usually large) set of items. For simplicity, this thesis refers to such items as 'relevant items'. In this thesis, the recommendation task is to first predict the relevance score of each item to the user, then rank the items based on relevance, and finally present the list of most relevant items in decreasing relevance [43, 149].[2] This is done for each user in a

[2]There are also other ways of generating or presenting the recommendations, e.g., displaying them in a grid [176] or carousels [45].

given set of users. Typically, the focus is on recommending items that have not been presented to the user [31], even though re-recommendation can be useful in some cases, such as re-recommending consumables [127] and songs [146].

1.2.1.2 Approaches in Recommendation

To personalise the item recommendation, recommender systems learn from the users' past interactions. These interactions exist in the form of explicit and implicit feedback, also referred to as ratings. Explicit feedback indicates whether a user likes or dislikes an item (e.g., via a like/dislike button), and to what extent (e.g., via a 5-star scale) [4]. For implicit feedback, user preferences are inferred from their activities rather than explicitly expressed by the user [4]. For example, clicking on or buying an item may indirectly indicate a positive preference for that item [4]. Other information, such as user/item attributes, as well as contextual information (e.g., time and the user's location), may also be used to enhance the recommendation [3, 4].

Two of the most common recommendation methods are collaborative filtering methods and content-based methods [4]. Collaborative filtering is based on the idea that similar users would prefer similar items [195]. Content-based methods mainly utilise item attributes/features to select other items that are similar to the ones that the users preferred in the past [169]. This thesis mainly utilises collaborative filtering methods. Some examples of collaborative filtering methods are popularity-based models, which recommend the most popular items [104]; matrix factorisation models, which characterise users and items as vectors of latent factors and predict with inner product the user ratings for the unrated items [119]; and neighbourhood-based models, which average ratings from similar users or items [55, 186]. More advanced collaborative filtering methods are usually deep learning-based [95, 256]. Recent work has also utilised large language models as recommenders [100, 136].

1.2.1.3 Evaluation of Recommender Systems

At a high level, recommender system evaluation protocols can be categorised into three: offline evaluation, user study, and online evaluation [54, 91, 248]. Offline evaluation is done on a pre-collected dataset that contains the users' past interactions (e.g., ratings [94], listening history [196]), with the purpose of collecting quantitative data for predictive performance evaluation [248]. This is contrary to the other two protocols, which require real-time user input [76], enabling the collection of quantitative and qualitative data through eye-tracking experiments, action logs, and/or questionnaires [248]. In user studies, a small number of participants are recruited and tasked to engage with a prototype recommender system, in a laboratory setting

(controlled environment), whereas online evaluation involves collecting data from real-world users' interactions with different versions of recommender systems in production [91], such as through A/B testing [89].

User studies and online evaluations are typically much more expensive than offline evaluation, in terms of planning time, experiment time, and financial cost [248]. In contrast, the relatively low cost of conducting offline evaluations allows researchers to compare model performance on several datasets relatively quickly [98]. As such, offline evaluation can serve as the first evaluation phase, for example, to eliminate worse-performing recommender models from the candidate pool of models, before further testing in user studies or online evaluation [91]. This thesis focuses on offline evaluation; this evaluation protocol is described in more detail below.

Offline evaluation for recommender systems largely follows experimental practice in machine learning and information retrieval [33]. To conduct an offline evaluation with a pre-collected dataset, the dataset needs to be split into at least two parts: training and testing. Others split the dataset into three: training, validation, and testing [151]. The training split would be input to a recommender model [98], such that the model learns and infers user preferences from past interaction data. The validation split may be used to select the final model in the hyperparameter tuning process. The testing split serves as the ground truth for evaluating the predictions or recommendations generated by the model. The datasets can be split in many ways, for example, randomly or temporally [31]. Another factor to consider is whether the data should be split per user or split globally, considering all users. Recent work recommend performing a global temporal split, as other data splitting strategies may cause the recommender model to learn from 'future' interactions that would not be available yet at prediction time [106, 151].

Based on the testing split, model performance can be assessed through *evaluation measures*. Usually, these measures are computed to assess the recommendation effectiveness, but they can also evaluate other aspects such as novelty [34], diversity [31], or fairness [228]; this thesis evaluates recommendation effectiveness and fairness (§1.2.3.1). Generally, these evaluation measures map a ranking of items to a numeric score. Traditional information retrieval measures such as Precision at cut-off k (P@k) and Normalized Discounted Cumulative Gain at cut-off k (NDCG@k [102]) are often used to evaluate the effectiveness of recommender systems [33].[3] The measure P@k quantifies the proportion of relevant items that are within the top-k recommendation list, while NDCG@k compares how close a ranking is to the ideal ranking that follows a decreasing order of relevance. The choice of k is typically

[3]Many more effectiveness measures exist; these are just two examples, and they are not necessarily the optimal ones for all use cases.

small due to the limited size of the interface [45], e.g., between 1 and 20 [6]. This means that only a few specific items will be presented to the users [31, 43].

Pre-collected datasets are often incomplete [33, 91]; there may be additional items that are relevant to a user other than items that have been highly rated by the user, clicked, and so on. As such, the recommender model may struggle to learn from the limited amount of data, which could result in very low effectiveness scores [33]. The incompleteness issue could also affect evaluation in another way: items without ratings are usually deemed as non-relevant, resulting in an underestimation of the model's effectiveness [31]. Thus, the low scores do not necessarily mean that the model is ineffective [33].

1.2.2 Fairness in Recommender Systems

Recommender system fairness is introduced in two parts: the taxonomy of fairness followed by this thesis (§1.2.2.1) and strategies for fair recommendation (§1.2.2.2). Note that fairness evaluation is covered separately in §1.2.3.

1.2.2.1 Taxonomy of Fairness in Recommender Systems

There are many ways of viewing fairness in recommender systems. Two common ways of categorising recommender system fairness across various surveys (e.g., [1, 9, 132, 228]) are by considering the fairness *subject* and the partition *granularity* of the subjects. In terms of subject, recommender system fairness is often classified into *item fairness* and *user fairness* [228].[4] Based on partition granularity, fairness is commonly divided into *individual fairness* and *group fairness* [9].

While it is not possible to provide a general definition of fairness that always holds, even after categorising fairness in terms of subject and/or granularity, the intuition behind each fairness type and its importance is provided below.

Item fairness. Fairness for items is often defined in terms of disparity in item utility. The more equal the utility across (types of) items, the fairer the recommendation. Item utility has typically been associated with the *exposure* provided by a recommender system to various (types of) items. The utility can also be a combination of item exposure and the system's *effectiveness* of recommending items to users for whom the item would be relevant [9].

[4]Some literature also refer to item and user fairness as provider-side and consumer-side fairness, respectively [65].

Firstly, item exposure can be understood as how often an item gets recommended to users and at which rank position. The more often an item is recommended to users, the higher the exposure it receives. Similarly, top-ranked items can be said to receive more exposure than bottom-ranked items, as position bias causes them to be more likely seen and further actioned upon [107]. As such, item fairness is important mainly for the item provider/supplier side, whose goal is for someone to discover the products, services, or content that they have provided. This discovery can lead to many actions and potential benefits [27]. For example, on a music streaming platform, if a user gets recommended a song and listens to it, this may count towards the royalty that the artist earns.[5] Likewise, if a user purchases an item at an e-commerce site, the seller could earn money. Therefore, if a recommender system only recommends the same few items to everyone, many items will never be shown to users, ultimately disadvantaging the item providers.

Secondly, regarding effectiveness-based item fairness, it is important to recommend the items to the right users. In other words, exposure should be allocated proportionally to the item's relevance [19]. For instance, a fair movie recommender system may consider that blockbuster movies would cater to the broader audience's taste and recommend these movies more frequently than movies with a niche genre, which may have fewer fans. Instead, if the system recommends niche movies to the general audience, this would increase the items' exposure and promote item fairness (considering the definition that is purely based on exposure), but at the expense of users' satisfaction. An example of an unfair recommendation for items would be a job recruitment platform that ranks applicants with similar qualifications far apart [54]; here, the applicants are the 'items'. The lower-ranked candidate would experience reduced visibility to recruiters, which decreases their chance of being selected to proceed in the recruitment process, jeopardising their chances of employment and thus their livelihood.

This thesis studies item fairness in *Papers 1–4* (Chapters 2–5).

User fairness. Fairness for users is commonly defined in terms of user utility, e.g., the recommendation effectiveness. The most common definition revolves around the idea of having a more equal recommendation effectiveness across different (types of) users would mean that the recommendation is fairer [64]. It is important for all users to receive recommendations that effectively match their preferences and needs; otherwise, the users may not be satisfied with the recommendations and may abandon the platform [228]. For instance, a job recommendation platform that constantly provides irrelevant suggestions for job positions to certain (groups of)

[5] `https://support.spotify.com/us/artists/article/royalties/`

users could be deemed unfair for these users. This is because the (groups of) users may be disadvantaged in their job searching process, as they have a harder time finding relevant jobs than other users.

Another possible definition for user fairness involves providing different treatment for different groups of users. For example, Deldjoo et al. [52] argue that paying users should receive better recommendations than non-paying users. The idea is that because paying users spend money to use the platform, their recommendation quality should be better than that of the non-paying users. Hence, fairness for users does not always mean that the recommendation effectiveness has to be equal for all users or user groups.

In this thesis, user fairness is covered in *Papers 5–6* (Chapters 6–7).

Individual fairness. Individual fairness generally refers to providing equal treatment to similar individuals [63]. The idea is that users with similar characteristics (e.g., in their past interactions [235]) should receive similar recommendations, or that equal exposure should be given to similar items (e.g., similarity in terms of their popularity in the training data [82, 171] or in their learned representations [227]). As there is no agreement on how the similarity should be measured, and that the similarity measure is often task-specific [63], individual fairness is also often based on the idea that all individuals should be treated equally, regardless of their similarity [126, 167]. Based on this notion of individual fairness, a recommender system would be highly unfair towards individual users, for example, if the recommendation effectiveness varies widely across users, meaning that some users receive high-quality recommendations and others receive low-quality recommendations. Similarly, a system would be highly unfair towards individual items, for instance, if some items get recommended to users much more frequently than others, while some items never get recommended at all.

Individual fairness is investigated in each paper included in this thesis (*Papers 1–6*, Chapters 2–7).

Group fairness. Group fairness is typically associated with ensuring that the experience of different user groups or item groups is comparable [65]. Under this fairness notion, similar users and items are grouped based on their attributes or characteristics. Often, these attributes (e.g., ethnicity, religion, gender) are defined in anti-discrimination laws, which prohibit discrimination based on specific group membership [14]. However, users and items can also be grouped based on other attributes that do not directly relate to socio-demographic background (e.g., user activity level [77, 130], user personality traits [39], or item popularity [77]). The idea

is that regardless of the socio-demographic background, their activity level, or their prior popularity, different groups of users/items should be able to receive similar treatment or similar utility (e.g., similar recommendation effectiveness for different user groups [64] or similar exposure for relevant item from different groups [174]). Even if these attributes are not explicitly part of the input, they may be implicitly present [72, 240]. Hence, ensuring that the recommendations are fair towards different groups requires more effort than omitting the use of group membership information from the input.

Unfairness may also arise from combinations of the attributes, i.e., intersectionality [44, 65, 111]). For example, the generated recommendations can be fair towards users when they are grouped based on only gender or only age, but extremely unfair when the intersectionality between gender and age is considered [50]. Other work considers fairness for different groups of users and items simultaneously [116, 229]. Even though the recommendation effectiveness may be equal across different user groups, different groups of items may be recommended to each user group, which can be seen as unfair, as it may perpetuate certain existing social stereotypes [116].

This thesis investigates group fairness in *Paper 6* (Chapter 7).

Other fairness categories. There are other ways of categorising fairness beyond item/user fairness or individual/group fairness. One such category is *subject fairness*, which concerns how item subjects are treated [65]. Subject fairness refers to having a diverse representation of individuals that the items are about, e.g., people or entities covered in news articles [65, 118]. In this regard, subject fairness could be seen as highly similar to item-side fairness [65]. However, in subject fairness, the idea of exposing/recommending a diverse list of items to each user is even more important than having an overall diverse recommendation across all users, which is the typical goal of item-side fairness [65]. Another type of fairness is *counterfactual fairness*, which ensures that individuals receive the same outcome if their group membership is changed to another group [170, 228]. Counterfactual fairness can also be seen as a specific type of individual fairness [121, 228]. In addition, there is also *within-group fairness*, which is related to the idea of providing an equitable outcome to all individuals within a group [171, 173]. These are just some examples of other fairness categories; many more ways of viewing fairness exist, such as static vs. dynamic fairness or single-sided vs. multi-sided fairness [132].

While this thesis also covers within-group fairness (*Paper 6*, Chapter 7), it does not cover subject fairness or counterfactual fairness, as explained in the scope of the thesis (§1.1).

1.2.2.2 Strategies for Fair Recommendation

Improving recommendation fairness can be done in several ways. However, increasing fairness may decrease recommendation effectiveness [92, 110, 251]. Hence, rather than achieving the maximum possible fairness, fairness-enhancing strategies focus on increasing fairness while maintaining an acceptable level of recommendation effectiveness. Broadly, these strategies can be categorised into data-oriented methods, ranking methods, and re-ranking methods [228].

Data-oriented methods refer to the modification or augmentation of training data, such that the recommender learns from more balanced interaction samples (e.g., in terms of user group proportions [64], or user preferences [183]). The idea is that, with these adjusted samples, the recommender would generate more equitable recommendations towards various users or user groups. Meanwhile, fair ranking strategies are primarily in the form of recommendation models that are optimised for fairness, e.g., by incorporating fairness-related terms in the model's objective function [28, 244]. As such, the model is optimised for both recommendation effectiveness and fairness. Lastly, fair re-ranking methods receive a recommendation list generated by a recommendation model and reorder the items in the list, with the aim of improving fairness (e.g., by pushing underrepresented items closer towards the top of the list [250]). The re-ranking can be done independently per user or simultaneously across all users.

While not the main contribution of this thesis, we explore fair re-ranking methods to improve item fairness in Papers *2–4* (Chapters 3–5).

1.2.3 Offline Evaluation of Recommender System Fairness

This thesis focuses on evaluating recommender system fairness in an offline setting. As such, fairness evaluation is done by computing measures as a proxy of the disparity in the recommendations. The general idea of how fairness is quantified through the measures is explained in §1.2.3.1, while their limitations are covered in §1.2.3.2.

1.2.3.1 Evaluation Measures of Fairness in Recommender Systems

Many survey papers have collected existing measures for quantifying fairness in recommender systems and categorised them in various ways based on their use cases [1, 9, 54, 133, 173, 209, 228, 237, 252, 261]. A common practice across previous work is to cluster these measures according to the evaluation subject for whom fairness is measured (i.e., users vs. items) and the evaluation granularity, which refers to the level at which fairness is assessed (i.e., individuals vs. groups). This clustering is

aligned with the fairness taxonomy used in this thesis (§1.2.2.1). Note that some measures overlap across categories as they can quantify more than one fairness type [9, 228].

To illustrate the workings of existing evaluation measures for fairness in recommender systems, the typical variables in the measure equations and how they are used are summarised here, following the categorisation based on evaluation subject and evaluation granularity. The specific measures and their mathematical equations are introduced in the later chapters of this thesis; the chapters are referenced in each category below.

Item fairness measures. Item fairness is widely defined in terms of how recommender systems allocate exposure to different (types of) items [228]. As such, item fairness measures generally quantify disparity in item exposure, which can be inferred from the frequency of an item's appearance in the recommendation lists across all users. The list can either be the top-k recommended items or all items. Often, item exposure is weighed based on the rank position at which the item appears. Most measures compare exposure between items either in a pairwise manner or altogether across all items. The item exposure is also compared to the number of items in the dataset and to the number of recommendation slots, i.e., the total number of items that are recommended to all users. This is because the total amount of exposure that can be distributed to the items is limited by the number of slots.

A subset of these measures also considers item relevance to the user; this thesis refers to them as 'joint' measures or 'relevance-aware' measures. Some measures compare item relevance to its exposure; the idea is that items with higher relevance should receive more exposure. Alternatively, item relevance and item exposure are combined to indicate user interaction: if a relevant item is exposed to a user, it is assumed that the user will interact with the item [191]. Unfairness is then measured as the disparity of these interactions across all items. The higher the disparity, the more unfair the recommendation.

Paper 1 (Chapter 2) of this thesis presents item fairness measures that are based only on exposure (detached from relevance). *Papers 2–3* (Chapters 3–4) detail measures that quantify relevance-aware item fairness.

User fairness measures. User fairness generally refers to how fair a recommender system is to different (types of) users [228]. Measures of this type commonly score fairness based on the difference in user utility. First, as a proxy of user utility, the recommendation for each user is evaluated for effectiveness with measures such as Precision at k (P@k) or NDCG@k (§1.2.1.3). Afterwards, the effectiveness score per

user is fed into another measure that aggregates the disparity in recommendation effectiveness across individuals or across groups. The aggregated score would be taken as the fairness score. The higher the disparity in recommendation effectiveness, the more unfair it is.

Some measures quantify fairness for users through the difference in the list of recommended items, regardless of the items' relevance to the users [235]. These measures first gauge the dissimilarity between the recommended items in two lists, e.g., by considering the number of overlapping items. Similar to how utility-based user fairness measures operate, the aggregated recommendation similarity across users or groups of users is regarded as the fairness score. More similar recommendations translate to a fairer score.

In this thesis, the formulations of user fairness measures are given in *Papers 5–6* (Chapters 6–7). *Paper 5* (Chapter 6) focuses on both effectiveness-based and non-effectiveness-based user fairness measures, while *Paper 6* (Chapter 7) examines only the former.

Individual fairness measures. Individual fairness is often defined as treating similar individuals similarly [63]. There are two types of individual fairness measures: those that are similarity-based and those that are not. Many measures for individual fairness in recommender systems are not similarity-based; they quantify individual fairness as the disparity or variation of utility (e.g., item exposure or recommendation effectiveness) across all individuals. A handful of measures are similarity-based; they compute similarity between pairs of individuals based on the representation of the individual, e.g., the user/item embeddings [227, 235]. In these measures, only pairs of individuals exceeding a similarity threshold contribute to the fairness score.

The details of individual fairness measures can be found in *Papers 1–3* and *Papers 5–6* (Chapters 2–4 and Chapters 6–7 respectively). *Papers 1* and *5* (Chapters 2 and 6) cover both similarity-based and non-similarity-based individual fairness measures, while *Paper 2, 3* and *6* investigate only the latter.

Group fairness measures. Based on the number of groups, two types of group fairness measures exist: measures that are defined only for two groups and those that can operate with two or more groups. Measures of the first type tend to directly compare utility between the two groups, e.g., taking the difference or the ratio between them. Some measures that can be used for two or more groups average the utility difference between group pairs, while others quantify fairness with respect to variation across groups. The group utility scores can either be weighted or used as they are (unweighted). The weights can be set arbitrarily, but they can also be de-

termined based on the proportion of group utility (to the total utility), or proportion of the group size (to the total population) [9].

This thesis employs group fairness measures that are explicitly defined for two or more (user) groups. These measures are overviewed in *Paper 6* (Chapter 7).

1.2.3.2 Limitations in Existing Fairness Evaluation Measures

Evaluation measures for recommender system fairness may have some limitations. These limitations are either in the measure formulas or arise from the empirical use of the measure. While in most cases the measures can still be computed to evaluate fairness, these limitations may restrict the measure's applicability, interpretability, expressiveness, stability, or efficiency. Hence, it is important to identify, analyse, and resolve these limitations when possible.

Determining the limitations of a fairness measure is a complex problem of its own. Before the first paper in this thesis (Chapter 2, [177]), there were not many existing work that examines the possible limitations of fairness measures in recommender systems. To the best of our knowledge, the only work that touches upon the issue was Raj and Ekstrand [175], which mentions and addresses the issue of edge cases in group fairness measures for ranking, as part of a framework for identifying the conceptual differences between measures. Previous work studies the limitations of other measure types by first defining a list of properties or desiderata. However, for recommender system fairness measures, there is no standard 'checklist' to follow.

Other prior work from a wide range of fields within and outside computer science have constructed a number of lists detailing properties for various types of evaluation measures: effectiveness [153], diversity [5, 8], fairness in computer network [101], and income inequality in the economics domain [7, 24]. Some general numerical properties, such as whether the measure score is bounded, can be readily adopted as ideal characteristics for recommender system fairness evaluation measures. However, not all properties would serve as suitable criteria. Firstly, some of these properties are not specifically tailored for fairness objectives in recommender systems, which means that they may not be particularly meaningful or relevant in the context of recommender system fairness. Secondly, some of them are not directly applicable and require modification to suit the fairness context. Thus, determining the measure limitations demands significant effort in collecting properties from existing work and adapting them for recommender system fairness evaluation measures, or even proposing new, specific criteria that suit the context.

This thesis explores a series of measure limitations and groups them into three classes. The following part explains the three classes of measure limitations in terms

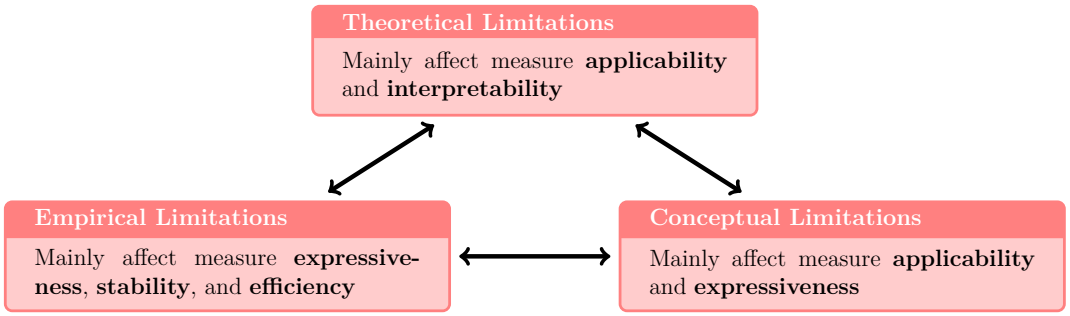


Figure 1.1: Overview of the three types of limitations identified, analysed, or resolved in this thesis for existing evaluation measures of recommender system fairness.

of how they can be identified, their practical implications, and their relationships. Fig. 1.1 overviews the measure limitations.

Theoretical limitations. These limitations arise directly from the measure's mathematical formulation and can be primarily identified by scrutinising the equation or testing it with different values to simulate common and extreme cases. For example, the equation may be ill-defined or fail to handle some scenarios. Consequently, for these cases, the measure cannot be computed due to, e.g., division by zero, and no score is produced. While some prior work handled these cases in their implementation, the issue with the general measure formulation itself remains, decreasing the measure's **applicability**.

In other cases, the precise range of the measure is unknown. Various reasons for this include ambiguous measure equations (e.g., details such as the logarithmic base are not stated) and the measure range is not mentioned in the original work proposing the measure. Even if the measure range is known, there are cases where the (un)fairest possible recommendation is not mapped to the endpoints of the range.[6] As the maximum or minimum achievable score is unknown, it is not possible to assess how close the recommendation is to the (un)fairest scenario. Thus, it is difficult to **interpret** the measure scores.

Paper 1 (Chapter 2) and *Paper 3* (Chapter 4) critically analyse theoretical limitations in existing individual item fairness measures for recommender systems. These papers pinpoint the specific causes and conditions resulting in the limitations. Additionally, they formulate corrections to the measures to counter the limitations or justify why the limitations are not resolvable.

[6]This is similar to the realisability property in effectiveness measures [153].

Empirical limitations. These are properties that emerge during or after the actual computation of the measure. Several measure characteristics can be tested through experiments with real and synthetic datasets, for example, **expressiveness**: Does a measure tend to always score very low or very high? How does it respond to changes in the recommendation list or item relevance? A measure with a consistently compressed empirical range has limited sensitivity, which makes it hard to distinguish the fairness of different recommendations.

On the other hand, there exist conditions that ideally should not impact the measure score or only minimally change it (e.g., ties in item relevance should not change the score considerably). If the measure score is largely affected by such conditions, its **stability** is questioned. Furthermore, there is also the question of **efficiency**: can the measure be computed in a reasonable time? A measure that requires extensive computational time may not be desirable.

All of the above concern limitations within a single measure. When comparing two or more measures, there may be concerns whether a particular measure is redundant: can the same conclusion be achieved by computing another measure instead? Existing literature on ranking effectiveness measures [230] and on machine learning fairness measures [144] highlight the **redundancy** of several measures, as computing another measure leads to similar conclusions. One way of assessing the conclusion similarity is by computing the measure agreement in ranking recommender models from the best to the worst (e.g., from the fairest to the most unfair) [192, 193]. This can be done by, for example, calculating the Kendall's τ correlation coefficient [112] between the two rankings [194].

Each of the six papers included in this thesis covers empirical limitations. Specifically, stability-related issues are analysed in *Paper 3* (Chapter 4); measure efficiency is investigated in *Paper 3* (Chapter 4), *Paper 4* (Chapter 5), and *Paper 5* (Chapter 6); measure expressiveness and redundancy are investigated in each paper.

Conceptual limitations. These limitations refer to certain important aspects pertaining to fairness that cannot be quantified by existing evaluation measures. This is because the measure may not be created for that purpose. An example of such aspects is user/item similarity. This aspect is often left out of individual fairness measures, even though individual fairness is commonly defined in terms of equality across similar individuals [63]. Naturally, a measure that does not account for a particular aspect in its equation can neither be used to evaluate nor expected to be sensitive to that aspect. This results in a need to modify existing measures or to formulate a new evaluation measure that is custom-built for the specific aspect(s), e.g., by including the missing aspect in the measure equation.

In this thesis, conceptual limitations are addressed in *Papers 4–6* (Chapters 5–7). *Paper 4* (Chapter 5) proposes a new evaluation approach to jointly quantify individual item fairness and recommendation effectiveness. *Paper 5* (Chapter 6) contributes a novel individual user fairness measure that considers both disparity in recommendation effectiveness and user similarity. *Paper 6* (Chapter 7) unifies for the first time individual fairness and group fairness evaluation for recommender system users by collecting and computing measures that are compatible with both.

Relationship between limitations. While this thesis broadly groups the limitations into three classes, these classes are not fully separable as they are still related to each other. An empirical limitation, such as the expensive computational time of a measure, may be traced back to the measure equation, which relates more to theoretical limitations. Yet, it is more practical to study how computational complexity translates to the actual computational time. Likewise, conceptual limitations can also be inferred directly from the measure equations. However, unlike theoretical limitations, which mainly concern flaws within the equation itself, conceptual limitations stem from missing variables or components. Lastly, about the relationship between conceptual and empirical limitations, the unique aspects quantified by a measure contribute to its expressiveness in those aspects. All in all, the three types of measure limitations should not be viewed or analysed separately but rather addressed together, as done in this thesis.

1.3 Scientific Contributions

This section provides a detailed summary of the papers included in the thesis. Altogether, the papers examine and resolve three categories of limitations in existing recommender system fairness evaluation measures: theoretical, empirical, and conceptual (see §2.3). Tab. 1.2 overviews the main contributions of each paper in assessing the measure limitations, addressing the limitations by designing a new evaluation approach or correcting the measures, and recommending general usage guidelines across a collection of measures.

Table 1.2: Overview of the main contributions made by the papers in this thesis based on the contribution type: assessment of **limitations** in existing recommender system fairness evaluation measures, resolution of the limitations through a new evaluation measure/approach and corrections to existing measures (**measure**), as well as **guidelines** that cover a set of fairness measures. The type of limitation spans across Theoretical (T), Empirical (E), and Conceptual (C) (Fig. 1.1).

	Limitation			**Measure**	**Guideline**
	T	**E**	**C**		
Paper 1 (Chapter 2)	✓	✓		✓	✓
Paper 2 (Chapter 3)		✓			✓
Paper 3 (Chapter 4)	✓	✓		✓	✓
Paper 4 (Chapter 5)		✓	✓	✓	
Paper 5 (Chapter 6)		✓	✓	✓	
Paper 6 (Chapter 7)		✓	✓		

1.3.1 Paper 1: Evaluation Measures of Individual Item Fairness for Recommender Systems: A Critical Study

In recent years, there has been an increasing amount of work on item-side fairness in recommender systems [228]. Along with this growing research interest, various evaluation measures for this type of fairness emerge. However, the practical differences and similarities among them, as well as their properties are not well-understood yet. Motivated by the need to investigate these aspects, we survey all existing exposure-based evaluation measures for individual item fairness in recommender systems (see §1.2.3.1), collect eight families of such measures, and analyse each measure in depth.

This paper contributes a novel analysis on five **theoretical limitations** of individual item fairness measures (summarised in Tab. 1.3). Out of the five limitations, we are the first to identify three within these measures. For the remaining two that have been previously discovered, we provide a formal explanation for their underlying causes. Furthermore, we address three limitations by correcting the measures or suggesting the use of another measure without the same limitation.

Two types of **measure corrections** were proposed, each addressing a specific limitation. The first method deals with the *undefinedness* limitation, a product of an undefined mathematical operation. To counter this limitation, we redefine the measure equation to account for edge cases, so that the score can now be computed

Table 1.3: Measures of exposure-based individual item fairness and their theoretical limitations. The table is reused from Rampisela et al. [177].

Legend ●: we fully resolve the limitation ○: the limitation is unresolvable ✓: another measure resolves the limitation	Source	Jain [101]	QF [264]	Ent [200]	Gini [85]	Gini-w [60]	FSat [167]	VoCD [227]	IL-D [234]	AL-D [234]
non-realisability: cannot reach max/min score (cause number denoted by C)										
C1. Most unfair score is only given to an impossible scenario	us	●	●	●	●	●	●			
C2. Fewer recommendation slots compared to number of items	us	●	●	●	●	●	●		○	○
C3. Number of recommendation slots is indivisible by number of items	us	●		●	●	○			○	○
C4. Non-realisability due to unknown formulation of max/min score	us					○		○	○	○
quantity-insensitivity: ignores frequency of item recommendation	[264]		✓							
undefinedness: cannot be computed (undefined value)	us			●						
always-fair: gives fairest score regardless of recommendation contents	[167]						○			
item-representation-dependence: depends on how items are represented	us							○		

for those cases. Our second method resolves the issue where measures that range between $[0, 1]$ cannot actually reach 0 or 1 for various reasons (*see non-realisability*, *C1–C4* in Tab. 1.3). Only by looking at the scores, it is not possible to understand how close the output is, to the fairest or unfairest case, which makes the scores hard to interpret. We correct this issue by normalising the measures based on the scores obtained from the (un)fairest possible recommendations. In this way, one end of the measure range is now mapped to the unfairest recommendation and the other to the fairest recommendation, easing the score interpretation. Furthermore, we verify through experiments that the corrected measures indeed resolve the limitations; we provide reasons why these limitations remain unresolved in isolated cases.

Additionally, we uncover the **empirical limitations** of the measures. We discover that some measures have limited expressiveness as they constantly score near-perfect fairness, even if the item exposure distribution is highly imbalanced. We also find that several measures may be redundant, as they tend to produce similar rankings of models.

The comprehensive investigation distinguishes this paper from prior work in two ways: Firstly, prior work studies fairness for item groups [175], while we focus on fairness for individual items. Secondly, existing work explain the high-level underlying idea behind the measures, but do not compute or compare all of them, whereas our study conducts both a theoretical and an empirical investigation.

Finally, considering the unique characteristics and challenges of each measure, we release **guidelines** to help researchers and practitioners select the appropriate measure for evaluating exposure-based individual item fairness in recommender systems.

1.3.2 Paper 2: Can We Trust Recommender System Fairness Evaluation? The Role of Fairness and Relevance

Broadly, there are two types of individual item fairness measures for recommender systems: purely exposure-based measures and 'joint' measures that account for item relevance on top of item exposure, i.e., 'relevance-aware' measures (§1.2.3.1). Our prior work (*Paper 1*, Chapter 2) conducts an extensive study on the exposure-based measures, but the properties and tendencies of the joint measures remain unclear as they have not been investigated. To this end, we investigate the **empirical limitations** of all individual item fairness measures that consider both item exposure and relevance.

Our investigation concentrates on two aspects: measure agreement and measure expressiveness. The purpose of studying measure agreement is to determine whether joint measures yield the same conclusion as single-aspect measures or other joint measures. We analyse measure expressiveness to understand how sensitive the measure scores are to changes in item exposure and item relevance.

This work contributes novel insights on the empirical tendencies of existing individual item fairness measures. First, we expose the fact that some joint measures are rather similar, and thus, one measure can be computed as a reasonable proxy for the other. Moreover, we discover their limited sensitivity towards changes in item relevance and exposure, even though both variables are part of the measures. We highlight several measures that consistently score very close to the fairest, impairing their interpretability, as their scores are always extremely fair.

Considering these alarming findings, we conclude with **guidelines** for the measure usage: we advise against computing similar measures to avoid redundancy; we encourage a more careful interpretation of the measure scores; and we suggest measuring fairness and relevance separately.

1.3.3 Paper 3: Relevance-aware Individual Item Fairness Measures in Recommender Systems: Limitations and Usage Guidelines

This work extends *Paper 2* (Chapter 3) and contributes three things: an in-depth investigation on the **theoretical limitations** of existing relevance-aware individual item fairness measures; **measure corrections** that resolve a subset of the limitations; and detailed **guidelines** that specify factors to consider for measure selection (Tab. 1.4).

Table 1.4: Factors related to the measure Alignment (A), Computability (C), Interpretability (I), Expressiveness (X), Stability (S), and Efficiency (E) to consider when selecting relevance-aware item fairness measures. Statements related to (A) are neutral; the rest are desirable.

Legend ✓: the statement applies to the measure in general ⚠: the statement applies to the measure, for binary relevance and single-round recommendation ✗: the statement does not apply to the measure	IAA_{ori}	IAA_{our}	$IFD_{\div\text{-}ori}$	$IFD_{\div\text{-}our}$	$IFD_{\times\text{-}ori}$	$IFD_{\times\text{-}our}$	HD_{ori}	MMF_{ori}	IBO/IWO_{ori}	IBO/IWO_{our}	$IL\text{-}F_{ori}$	$IL\text{-}F_{our}$	$AI\text{-}F_{ori}$
A1. The measure tends to agree with effectiveness measures	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗
A2. The measure tends to agree with fairness measures or disagree with effectiveness measures	✗	✗	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✓
C1. The measure can be computed even if the full ranking of items (not just the top-k) is unavailable	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
C2. The measure can be computed even if item relevance beyond top-k is unknown	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✓	✓	✓
C3. The measure can be computed even if the total # of relevant items per user is unknown	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗	✗
C4. The measure can be computed even if there is an item that is not relevant to any user	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓
C5. The measure can be computed for any k	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
I1. The measure maps the unfairest and the fairest recommendation to the endpoints of its theoretical range	✗	⚠	✗	⚠	✗	⚠	✗	✗	✗	✗	✗	⚠	✗
I2. The measure distinguishes between an item exposed at k and an unexposed item at $k+1$	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
X1. The measure scores are not compressed in a small interval	✗	✓	✗	✓	✗	✓	✓	✗	✓	✓	✗	✓	✗
X2. The measure responds to changes in k	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
S1. The measure is unaffected by item relevance beyond the top-k	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗	✗
S2. The measure is stable or becomes fairer if there are more relevant items beyond the top-k	✗	✗	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗
S3. The measure is unaffected by ties in item relevance	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓
S4. The measure is unaffected by the number of users with only one relevant item	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
E1. The measure can be computed separately per user	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓	✓	✗
E2. The measure can be computed separately per item	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓
E3. The measure is efficient to compute (<60 s) for large datasets (e.g., > 1K users and > 10K items)	✓	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✓	✓

We have previously conducted an empirical investigation on all existing relevance-aware individual item fairness measures and linked several findings on the measures' **empirical limitations** to the measure equations. Similar to *Paper 1* (Chapter 2), we critically analyse the measures' theoretical limitations by looking closely at their formulations. The analysis focuses on cases that would restrict the measure's appli-

cability or interpretability. To the best of our knowledge, this is the first work that identifies the theoretical limitations of the measures, and we find five such limitations.

Most limitations are resolvable; we justify why the rest are not. To overcome the resolvable limitations, we formulate extensions of the measures in two forms. The first form is a normalised version of the measure, which correctly maps the fairest possible recommendation to 0 and the unfairest to 1. This normalisation significantly improves the score interpretability, as previously the scores were always near 0, giving an illusion of all-time maximally fair output, even for the most unfair case. The second form of measure extension is a redefinition of the measure, which accounts for cases that cause the measure to be incomputable (e.g., division by 0). We kept these modifications to a minimum, so as not to radically change the original measure design, yet the measures can now be used in a wider range of cases.

In addition, through a series of experiments with real and synthetic datasets, this paper shows the extent of the measures' theoretical limitations in practical cases. At the same time, we confirm that our measure extensions have resolved the issues. The paper concludes with specific guidelines that connect the measure limitations and their practical implications on a wide range of evaluation scenarios. Our guidelines specify which measure should (not) be used in different cases and conditions. We hope that these guidelines will ease the selection of relevance-aware individual item fairness measures based on both their properties and their possible use cases.

1.3.4 Paper 4: Joint Evaluation of Fairness and Relevance in Recommender Systems with Pareto Frontier

Previous work, including ours (*Papers 1–3*, Chapters 2–4), has measured recommendation fairness and its relevance either separately or together. When the two aspects are quantified separately, it is hard to select the recommender with the best balance between the two. This is because the fairest recommender may not provide the most relevant recommendations or vice versa. Yet, evaluation measures that consider both fairness for individual items and their relevance to users also do not reliably quantify the balance between both aspects. These joint measures either align exclusively with one of the two aspects, or have inconsistent alignments (*Papers 2–3*, Chapters 3–4). In other words, there is no existing approach that can reliably quantify both relevance and individual item fairness via a single score.

To address this **conceptual limitation**, we present a **novel evaluation approach**, *Distance to Pareto Frontier* (DPFR), to find the most balanced model in terms of both fairness and relevance (see example in Fig. 1.2). We contribute a new algorithm that generates the Pareto Frontier based on the dataset and a selected cut-

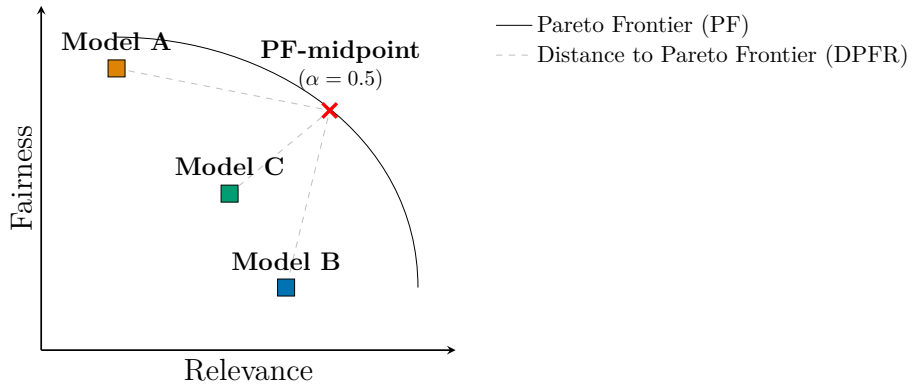


Figure 1.2: In this example, Model A is best for fairness, Model B is best for relevance, and Model C is the closest to the Pareto Frontier (PF) midpoint, when relevance and fairness are equally weighted ($\alpha = 0.5$). The figure has been reused from Rampisela et al. [181] with some modifications for clarity.

off k. Given these inputs, the algorithm first maximises recommendation relevance using the test set and iteratively improves fairness, until the maximally fair recommendation is formed. The final generated frontier contains a set of Pareto-optimal solutions, representing the best fairness-relevance trade-offs that are achievable based on the dataset.

The advantages of DPFR are threefold. First, it is modular, which means that it is model-agnostic and can be computed with various pairs of existing single-aspect measures. Secondly, DPFR allows evaluation against controlled, dataset-based trade-offs of fairness and relevance that fulfill Pareto-optimality; this is important as improving recommendation fairness often comes at the cost of sacrificing its overall effectiveness [110, 251]. Thirdly, our approach is intuitive and is easily interpretable as it directly measures how far the model score is from a desired fairness-relevance balance. No existing joint evaluation approaches for recommender system fairness and relevance have all three advantages.

In our experiments with six datasets of varying characteristics, we test the compatibility of existing single-aspect measures with DPFR. We find 12 pairs of relevance and fairness measures that are compatible with DPFR. Then, we compare DPFR to existing single-aspect measures and joint measures. Our results show that neither measure type can be a reliable proxy for DPFR, as they rank models from the best to worst differently from DPFR; these are part of the measures' **empirical limitations**. Finally, we show that DPFR can be computed efficiently without compromising its evaluation effectiveness. Thus, as an alternative to existing joint

Table 1.5: Overview of individual user fairness measures in recommender systems.

Measure	Reference	Effectiveness	User Similarity
Standard deviation	[20, 128, 167, 183, 236]	✓	-
Gini index	[77, 126]	✓	-
Envy-based	[20, 61, 167]	✓	-
UF	[235]	-	✓
PUF	ours	✓	✓

evaluation approaches, we recommend computing DPFR to select the empirical best possible recommendation based on both relevance and fairness.

1.3.5 Paper 5: Measuring Individual User Fairness with User Similarity and Effectiveness Disparity

Fairness in recommender systems can be evaluated not only for items–as we have investigated in *Papers 1–4* (Chapters 2–5)–but also for users. In this paper, we shift our focus to fairness evaluation measures for individual users.

Individual fairness is commonly defined on the basis of having equal treatment for *similar* individuals [63]. Yet, upon collecting all existing evaluation measures for individual user fairness (overviewed in Tab. 1.5), we discover that only one such measure [235] actually follows this definition by accounting for user similarity. This measure quantifies fairness based on the representation of items in two similar users' recommendation lists. A couple of **conceptual limitations** arise from quantifying fairness in this way. Firstly, having similar recommended items may not directly translate to having similar recommendation effectiveness: if two highly similar users get recommended identical items, they may not equally like their recommendations, as their exact preferences could differ. Secondly, even if similar users are equally satisfied with their recommendations, the measure would deem this unfair if the recommendations consist of entirely different sets of items with dissimilar representations. Thus, it is important to evaluate fairness beyond differences in item representation.

In this paper, we introduce *Pairwise User unFairness* (PUF), a **novel evaluation measure** for individual user fairness that is both aligned with user utility as a key objective in recommender systems and with the individual fairness definition. PUF evaluates unfairness as the average pairwise difference of user utility, weighted by the user pair similarity. The user utility is the recommendation effectiveness score (e.g., NDCG). The idea is that if a user pair has either low similarity or a small difference in

their effectiveness scores, the pair would only contribute a small increase in the PUF score, which translates to less unfairness. Meanwhile, highly similar user pairs with a large effectiveness score difference would have a higher PUF score, which means unfairer recommendations. With this measure, fairness for individual users can now be evaluated based on effectiveness disparity while also considering user similarity.

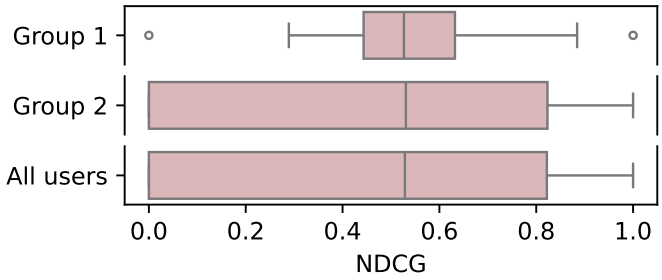
Our experiments demonstrate that PUF can reliably capture both user similarity and recommendation effectiveness, while existing measures fail to do so; these are the **empirical limitations** that we uncovered from the measures. Moreover, as no existing measures consistently yield similar conclusions to PUF, none can be used as a reliable estimate of our measure. By having a measure that is aligned with the individual fairness definition and user utility, we hope that our measure can serve as a tool for practitioners to meaningfully estimate performance disparity across users.

1.3.6 Paper 6: Stairway to Fairness: Connecting Group and Individual Fairness

Most work on recommender system fairness focuses exclusively on group fairness, rarely evaluating for both group and individual fairness [54]. Furthermore, in prior studies on both fairness types, group fairness and individual fairness are evaluated with different measure families, which makes it difficult to compare them, as different measures may not have equal sensitivity or empirical ranges [177, 178, 198]. Other literature instead evaluates fairness for different subjects or objectives, e.g., equal recommendation effectiveness across individual users and equal exposure between item groups [236], which are also not directly comparable. Consequently, the interplay between group and individual fairness is not scientifically well-understood. Analysing the relationship between the two fairness types is key to understanding how improving one fairness type can impact the other.

We contribute the first empirical study on the relationship between user-side group and individual fairness evaluation measures. To allow a proper comparison of group fairness and individual fairness, we evaluate both with the same measure family. We gather all existing user-side group fairness measures and identify a subset of them that have also been used as a user-side individual fairness measure. In the given example (Fig. 1.3), both fairness types are measured with Gini [85]. This example shows in practice how a recommender can be very fair towards user groups and, at the same time, much more unfair towards individual users.

In this paper, we also address a **conceptual limitation** of user-side group fairness measures. No existing user-side group fairness measure satisfies both the following criteria: (1) it ranges between 0 and 1, which enables a proper comparison of



	#user	↑NDCG	↓Gini
Group 1	11	0.546	0.037
Group 2	901	0.471	
Individual	912	0.472	0.446

Figure 1.3: Left: NDCG score distribution of two handpicked, non-overlapping user groups from our experiment and for all individual users of the two groups. Circles denote outliers. Right: Fairness score (Gini) for groups and individuals. The lower the Gini score, the fairer. Fairness is higher between groups than across all individuals as both groups have similar average NDCG, but within-group variance is high, which means that recommendation quality varies widely across users (see the boxplots).

group and individual fairness; and (2) it accounts for within-group variance, which relates to individual fairness. To this end, we identify an extra measure that fulfills them both and compute it to evaluate user-side group and individual fairness.

Our empirical analyses with 8 recommenders and 3 datasets reveal that group fairness scores tend to mask individual-level unfairness, regardless of how the users are grouped or the group size. We find that this is due to high within-group unfairness, where effectiveness scores vary widely in each user group, resulting in high utility variation across all users. Even when the group fairness measure formulation accounts for the within-group variance, the masking effect remains. More importantly, no group fairness measures consistently reach the same conclusion as any individual fairness measure, which is an **empirical limitation** of the measures. This necessitates the evaluation of both group and individual fairness, as the fairest model for user groups may not be the fairest for individual users.

All in all, we encourage a more comprehensive fairness evaluation by assessing both group and individual fairness. We especially highlight the importance of evaluating individual fairness on top of group fairness, as disparity across individuals goes undetected by group-level scores. Considering that most fairness-aware efforts have been centred on mitigating disparity between groups, our work motivates a push towards initiatives for improving fairness across individuals, such that all recommender system users/items can obtain a similar level of benefit.

1.4 Summary and Future Work

The papers included in this thesis collectively contribute to the advancement of state-of-the-art offline evaluation approaches for recommender system fairness. In particular, this thesis contributes: (1) a study and guidelines on exposure-based individual item fairness measures and our proposed extensions of the measures; (2) an empirical study on relevance-aware individual item fairness measures and their usage guidelines; (3) a theoretical study, empirical analyses, and guidelines on relevance-aware fairness measures, as well as our proposed measure extensions; (4) a new joint evaluation approach for both recommendation fairness and effectiveness; (5) a new evaluation measure for individual user fairness; and (6) a work that connects group fairness and individual fairness. More than 29 unique fairness measures for individual items, for individual users, or for user groups, were studied across the six papers. This number consists of both existing and new measures.

Through theoretical and empirical investigations, we expose serious flaws of existing recommender system evaluation measures in the form of theoretical, empirical, and conceptual limitations. Altogether, these flaws may severely limit the measures' interpretability, expressiveness, and applicability. On one hand, our study reveals redundancy among existing measures, meaning that not all of them must be computed to have a good overview of fairness, yet some of them conflict in their conclusions in ranking models from the fairest to the unfairest (*Paper 2*, Chapter 3). On the other hand, we show that evaluating for only one type of fairness is insufficient (*Paper 6*, Chapter 7). In addition, we find that even fairness measures that operate on the same idea (e.g., parity), for the same subject (users/items), and are evaluated for the same granularity (individual/group) often differ in their theoretical and empirical properties. Practically, this means that some measures are more suitable for relative comparison among models than for a standalone evaluation of a single system. This also means that the measure scores need to be carefully interpreted as some are stricter than others. Furthermore, fairness-optimisation methods guided by these measures may require different approaches due to the measures' various alignment to effectiveness measures. Finally, we have condensed the insights from our investigations into practical guidelines for the appropriate measure usage.

While this thesis aims to provide a comprehensive coverage of offline evaluation of fairness in recommender systems and has studied a wide range of fairness measures, it is by no means exhaustive. Many questions remain unanswered, leaving gaps for further investigation. Therefore, three key potential research areas for future work on fairness in recommender systems are identified and discussed below:

Online and user-centred fairness evaluation. This thesis has concentrated on a specific evaluation paradigm: offline evaluation. While this paradigm is widely employed in recommender system fairness research, it is not the sole way of evaluating fairness. There are not many studies on online evaluation for recommender system fairness (e.g., [18]), but they are worth further investigating, as they directly reflect recommender systems' performance in the real world. Similarly, user studies on what stakeholders perceive as fair, such as [57, 67, 74, 207, 209–211, 217], are still rather limited in coverage or participants' background diversity. This opens up avenues for large-scale online studies, as well as user studies that target a more diverse set of participants and a wider range of fairness notions.

Even though offline evaluation is relatively faster and simpler than conducting online evaluation or user studies, their results may conflict [16, 33, 80]. This is especially crucial for fairness evaluation, as prior user studies have pointed out that some desirable fairness notions based on user perception align more with the concepts of privacy, inclusion, and transparency, rather than conventional fairness concepts in offline evaluation of recommender system fairness [67, 74, 210]. Although these aspects, together with fairness, are pillars of responsible artificial intelligence (AI), they are not commonly accounted for in existing fairness measures [210]. Hence, to have a more complete picture of recommender system fairness and its connection to broader societal issues, there is a need to investigate other evaluation paradigms, especially in their alignment with existing offline approaches and in relation to other pillars of responsible AI. By leveraging user studies, future work could design offline evaluation measures that cater to actual stakeholder needs and perceptions of fairness.

Multi-stakeholder, multi-round, and multi-attribute fairness. The papers included in this thesis have conducted an in-depth investigation into single-stakeholder fairness at a fixed point in time. However, in reality, recommender systems operate across multiple rounds of recommendations and involve multiple stakeholders, each of whom may be associated with multiple attributes. This calls for multi-stakeholder, multi-round, and multi-attribute fairness evaluation.

Firstly, about multi-stakeholder fairness, recommender systems are two-sided platforms that require a balance between users, items, and the possible third parties involved [65]. As such, fairness evaluation should also consider the multi-stakeholder nature of the ecosystem to minimise unfairness towards all parties. Recent work has proposed several measures to jointly evaluate fairness across stakeholders (e.g., [116, 229, 234]). Yet, the relationship between multi- and single-stakeholder fairness measures is currently unknown. For example, it is not known if optimising for a multi-stakeholder fairness measure would automatically improve single-stakeholder

fairness scores (or vice versa). More importantly, as there are no existing studies that 'stress-test' these measures for a wide range of scenarios, it is not well-understood if these measures would work properly in all cases.

Secondly, regarding multi-round fairness, real-world recommender systems users typically engage with the platform over a period of time. Over time and over multiple rounds of engagements, the system learns the user preferences better and iteratively provides updated recommendations that are learned from the users' previous interactions. Consequently, recommendation effectiveness may change over time. As user-side fairness is typically defined based on disparity in recommendation effectiveness, it is important to study how this type of fairness changes over time. Multi-round evaluation is also relevant for item-side fairness, which is often defined in terms of the exposure received by items. The total amount of exposure is limited directly by the number of recommendation slots. As there are more slots in multi-round recommendations, there are also more opportunities for items to appear in a user's recommendation list. Therefore, it would be interesting to study fairness considering all rounds of recommendations.

Fairness in multiple rounds has been evaluated either separately per round [73, 75, 245] or jointly considering all rounds [23, 234]. In *Paper 1* and *Paper 3* (Chapter 2 and Chapter 4), we analyse measures that quantify multi-round fairness for items. While we have briefly discussed their theoretical limitations under the multi-round evaluation setup, their in-depth and empirical investigations are left for future work.

Thirdly, about multi-attribute fairness, group fairness evaluation is closely tied to user/item sensitive attributes (e.g., gender, race). Our work (*Paper 6*, Chapter 7) studied fairness for user groups and found that fairness worsens when the groups are formed with more sensitive attributes. As such, we highlight the importance of multi-attribute fairness assessment. While our investigation is limited to treating all attributes as nominal variables, future work could explore multi-attribute fairness evaluation measures that accommodate ordinal and interval values (e.g., user age and years of work experience) in conjunction with nominal values [152, 194].

Recommender system fairness and Large Language Models (LLMs). The rise of LLMs in recent years has opened up multiple new areas of research that build upon LLMs' world knowledge [172, 258], as well as their 'reasoning' [49, 162] and 'instruction-following' ability [163, 254, 263]. Beyond using LLMs as zero-shot recommenders, e.g., [100] and *Paper 6* (Chapter 7), several ideas on how recommender system fairness research can benefit from LLMs are outlined below.

Recommender system fairness research can potentially exploit LLMs in at least three ways: for data bias detection, for fairness-aware approaches, and for fairness

evaluation. Existing work has leveraged LLMs to detect bias in tabular datasets [246]; similar tools can be developed for identifying potential biases in recommender system datasets, so that unfair biases arising from data can be managed prior to model training. LLMs have also been used as a fairness-aware reranker for recommendations [78]. As this reranking approach is still limited to item-side fairness and for a specific fairness measure, future work could extend the approach for other fairness notions and measures. Additionally, the use of LLM-as-a-judge has gained increasing popularity; LLMs have been used to evaluate personalisation [62], document relevance [13], and recommender system explanation [257] with varying levels of success. Hence, one possible research direction is to experiment with LLMs to assess fairness. By doing this, LLM alignment to various fairness concepts can also be examined. Finally, another area that is worth exploring concerns a newly proposed fairness evaluation framework for LLM-based recommenders [50, 51, 255]. In this framework, fairness evaluation is done by comparing two recommendation lists, each generated by including and excluding a user's sensitive attributes in the prompt. Understanding how this framework aligns with other existing evaluation approaches can be crucial to assessing its validity and robustness.

Overall, the topic of fairness in recommender systems remains relevant not only in the specific field of responsible AI research, but also in the broader society. With the development of new technologies, such as LLMs, there is both a legal and practical necessity to uncover their hidden bias before further leveraging them to assess and improve fairness in other applications. Evaluating the potential harms and risks arising from these biases is the first step in mitigating and reducing unfairness.

Chapter 2

Evaluation Measures of Individual Item Fairness for Recommender Systems: A Critical Study

Abstract

Fairness is an emerging and challenging topic in recommender systems. In recent years, various ways of evaluating and therefore improving fairness have emerged. In this study, we examine existing evaluation measures of fairness in recommender systems. Specifically, we focus solely on exposure-based fairness measures of individual items that aim to quantify the disparity in how individual items are recommended to users, separate from item relevance to users. We gather all such measures and we critically analyse their theoretical properties. We identify a series of limitations in each of them, which collectively may render the affected measures hard or impossible to interpret, to compute, or to use for comparing recommendations. We resolve these limitations by redefining or correcting the affected measures, or we argue why certain limitations cannot be resolved. We further perform a comprehensive empirical analysis of both the original and our corrected versions of these fairness measures, using real-world and synthetic datasets. Our analysis provides novel insights into the relationship between measures based on different fairness concepts, and different levels of measure sensitivity and strictness. We conclude with practical suggestions of which fairness measures should be used and when. Our code is publicly available. To our knowledge, this is the first critical comparison of individual item fairness measures in recommender systems.

2.1 Introduction

The concept of fairness in Recommender Systems (RSs) is commonly understood as treating users or items that are alike, in a similar way. It can be studied either for individual items or users (individual fairness), or for groups of items or users (group fairness). Individual fairness typically refers to similar individuals being treated similarly [19], where sometimes the similarity of individuals is measured through a certain metric, e.g. distance of the individual representation [133]. In this work, we focus solely on fairness for individual items, and specifically on evaluation measures designed to quantify individual item fairness in RSs. While there exist comprehensive surveys on group fairness measures [175, 252], to the best of our knowledge, no critical analysis of evaluation measures of individual item fairness for RSs has been presented.

Individual item fairness is an important type of fairness that occurs in many RS scenarios. For example, it is helpful for new item discovery and for ensuring that recommended items come from different providers or creators. Individual item unfairness may occur due to popularity bias, which causes some items to be recommended more often than others [267]. Some items may not even be recommended at all. For instance, interesting content from emerging content creators could be recommended less frequently than less or equally interesting content from popular creators.

In a traditional recommendation scenario, a model produces a top-k list of recommendations across all users. This output is typically evaluated by measuring the recommendation relevance of the top-k recommendations to each user. On top of that, one can measure the individual item fairness of the top-k recommendations. On a high level, we distinguish between two definitions of individual item fairness. In the **first definition**, individual item fairness is understood as *all items*[1] *having equal exposure*. Exposure (also known as attention [19] or coverage [227]) refers to an item appearing in the top k recommendations for a user. The definition of fairness above does not consider how relevant the recommended items are to the users; it only considers how uniform their exposure is. Given that relevance is a key aim in RSs, fairness has also been given a **second definition** as *all items having an equal opportunity for exposure, where the opportunity is based on item relevance to users or similar other criteria [56, 234].*

Several evaluation measures have been used to empirically evaluate fairness for individual item fairness in RSs [228]. All of the measures evaluate fairness based on the recommendation list; the input is the recommendation list, and in most cases,

[1] All items in the dataset or all items in the recommendations given to all users. Both definitions are used in the literature and also in our work in §2.2.2.

the measure compares the exposure of items in the recommendation list and the total number of items in the dataset. Some of the measures follow the first definition of fairness, which is purely exposure-based, while other measures evaluate fairness jointly with relevance. Here, we focus solely on fairness measures of the first type, that is measures that evaluate only fairness without considering relevance. Even if these fairness-only measures have been used to evaluate fairness in prior work, they have not been extensively analysed for their limitations and possible consequences arising thereof. They have also not been analysed in relation to one another, as most research in individual item fairness only uses a single fairness measure. As a result, it is currently unclear how to interpret these measures and select which measures should be used under various circumstances. For instance, we find cases where a measure is theoretically bound between $[0, 1]$ but empirically cannot reach either of the endpoints, which makes the interpretation of the measure extremely difficult. It is also unknown if there are theoretical limitations and cases where the measures would fail. For instance, we find that a measure will give the highest scores no matter the recommendation, given a certain number of users, items, or cut-off. We also find that a measure cannot be computed as the formulation is not well-defined for a commonly-encountered case.

The goals of this work are to address the above research gaps, by examining existing measures of individual item fairness in RSs, presenting analytical limitations and solving them (or justifying why they cannot be solved), and examining how the scores of the measures change in relation to each other in different scenarios. As such, we contribute the following:

- We review individual item fairness measures in RSs (§2.2).
- We identify and analyse 5 theoretical limitations of those measures, where three of the limitations are novel and two limitations have already been identified but without a formal/theoretical explanation, which we provide (§2.3).
- We propose theoretical corrections of the measures to resolve their limitations or explain why some limitations are unresolvable (§2.4). Note that none of the measures is without limitations (Tab. 2.4).
- We present an empirical analysis of the existing measures to study the correlation among measures, investigate the relations between fairness and relevance, and compare the corrected measures against the original measures on real-world and synthetic datasets (§2.5).
- We provide insights to guide the correct use of different types of individual item fairness measures (§2.7).

Table 2.1: Summary of the notation.

Notation	Explanation
$U = \{u_1, u_2, \ldots, u_m\}$	The set of users
$I = \{i_1, i_2, \ldots, i_n\}$	The set of items
$\lvert U\rvert = m$	The unique number of users in dataset
$\lvert I\rvert = n$	The unique number of items in dataset
k	The cut-off threshold
km	The total number of recommendation slots
$r_{u,i} \in \{0, 1\}$	The relevance of item i to user u
$z(u, i) \in \{1, 2, \ldots, n\}$	The rank position of item i for user u
$z(u, i, w) \in \{1, 2, \ldots, n\}$	The rank position of item i for user u in round w
R_u^k	The top k recommendations for user u
$R_{u,w}^k$	The top k recommendations for user u in round w
R	The set of top k unique items recommended to all users
$1_{\mathcal{A}}(x) = 1$ if $x \in \mathcal{A}$, else 0	Indicator function

Table 2.2: Examination functions used in fairness measures.

	Equation	Measure	Reference
uniform	$e_{\text{unif}}(u, i) \equiv 1, \forall(u, i)$	Jain, QF, Ent, Gini, FSat, VoCD	[85, 101, 167, 200, 227, 264]
DCG	$e_{\text{DCG}}(u, i, w) = 1/\log_2(z(u, i, w) + 1)$	Gini-w	[60]
RBP	$e_{\text{RBP}}(u, i, w) = \gamma^{z(u,i,w)-1}$	II-D, AI-D	[234]

2.2 Individual Item Fairness Measures

We present our notation (§2.2.1) and the eight exposure-based evaluation measures of individual item fairness that we study in this work (§2.2.2). To the best of our knowledge, these eight measures are all the measures of individual item fairness that exist in RSs.[2] Several of these measures are taken from [228].

2.2.1 Notation and Examination Functions

Given a set of users $U = \{u_1, u_2, \ldots, u_m\}$, with $|U| = m$, and a set of items $I = \{i_1, i_2, \ldots, i_n\}$, with $|I| = n$, for each user $u \in U$ we rank all n items to produce the full recommendation list. The list of the top k recommended items to user u is R_u^k, and the set of all top k recommended items to all users is $R = \bigcup_{u \in U} R_u^k$. If u finds the recommended item i relevant, we write $r_{u,i} = 1$, otherwise $r_{u,i} = 0$. The rank position of item i in the recommendation list for user u is $z(u, i)$. As there exist fairness measures that consider multiple rounds of recommendation, we also use the following notations for such measures: the rank position of item i for user u in round w is $z(u, i, w)$ and $R_{u,w}^k$ is the list of the top k recommended items to user u in round w. Tab. 2.1 summarizes our notation.

Fairness for individual items is closely linked to exposure, which is identified as the appearance of an item in the top k recommendations for a user. Exposure can be quantified in several ways using different examination functions $e(\cdot)$ in various binary or graded ways. Examination functions are functions modelling the probability of a user seeing an item that is exposed to the user. All examination functions in this paper assume that this probability only depends on $z(u, i)$, the rank position of an item i for user u. Tab. 2.2 presents the examination functions used by the individual item fairness measures that are included in this paper. These examination functions apply either no discount, logarithmic discount, or exponential-like discounting.

The simplest examination function is uniform, e_{unif}, and assumes a constant weight of 1 for each rank position [85, 101, 167, 200, 227, 264]. The two other examination functions are e_{DCG} [60] and e_{RBP} [234], which use the discount function based on Discounted Cumulative Gain (DCG) [102] and Rank-Biased Precision (RBP) [154] respectively. In e_{RBP}, the parameter γ is the user's patience, i.e., the probability of the user examining the next ranked item. For example, $\gamma = 0.8$ in [234] and $\gamma = 0.5$ in [56].

2.2.2 Measures of Individual Item Fairness

We present the eight measures that so far have been used to quantify fairness for individual items (without considering item relevance), as well as the context in which they are used in the original work. We use the subscript $\cdot_{\text{ori}}$ to denote the original formulation of a measure as opposed to our corrected version that we present later in §2.4. Note that these measures are typically used with a fixed cut-off k, and therefore the number of recommendation slots km is also fixed.

[2]Based on publications up to August 2022.

2.2.2.1 Jain's Index (Jain) [101]

Jain, which was originally defined for fairness in computer networks, has been used in RSs [264] to measure how consistent item exposure is in relation to the number of times an item is recommended. The original work uses this measure to evaluate "fairness of the recommendation opportunity". Jain is the ratio between the square of the number of recommendation slots and the sum of squares of the number of times each item is recommended, where the ratio is divided by the number of items in the dataset. It is calculated as follows:

$$\text{Jain}_{\text{ori}} = \frac{\left[\sum_{i \in I} \sum_{u \in U} 1_{R_u^k}(i)\right]^2}{n \sum_{i \in I} \left[\sum_{u \in U} 1_{R_u^k}(i)\right]^2} = \frac{(km)^2}{n \sum_{i \in I} \left[\sum_{u \in U} 1_{R_u^k}(i)\right]^2} \tag{2.1}$$

where $1_{R_u^k}(i) = 1$ if item i is in the top k recommendations for user u, and 0 otherwise. $\sum_{u \in U} 1_{R_u^k}(i)$ counts how many times item i is recommended in the top k across all users. The range of Eq. (2.1) is $[0, 1]$.[3] The range of Jain matters as [264] analysed the absolute values of Jain, in addition to observing the difference of the scores. The higher the Jain score, the fairer the recommendation with respect to individual items (i.e., items are exposed consistently with respect to other items in the dataset). E.g., if 60% of the items in the dataset are exposed equally to all users, for instance, $R^3_{u_1} = [i_1, i_2, i_3]$, $R^3_{u_2} = [i_4, i_5, i_6]$, $n = 10$, then Jain $= 0.6$. However, this interpretation does not hold and becomes less intuitive when items are not exposed equally, which is often the case. For instance, $R^3_{u_1} = [i_1, i_2, i_3]$, $R^3_{u_2} = [i_1, i_2, i_4]$, $R^3_{u_3} = [i_1, i_5, i_6]$, $n = 10$. In this case, 60% of the items are exposed but Jain $= 0.476$. In real-life, it is unlikely that items are exposed equally, which means that Jain's interpretation suffers from this limitation, more often than not.

2.2.2.2 Qualification Fairness (QF) [264]

QF is a modification of Jain that measures how many items are in the set of top k recommended items R, divided by n, the total number of items in the dataset. The authors of the original measure explained in [264] that the measure only considers whether an item in the dataset is recommended, as opposed to how many times it is recommended.

[3] This range is based on the original paper, [101]

$$\mathrm{QF}_{\mathrm{ori}} = \frac{\left[\sum_{i \in I} 1_R(i)\right]^2}{n \sum_{i \in I} \left[1_R(i)\right]^2} = \frac{\left[\sum_{i \in I} 1_R(i)\right]^2}{n \sum_{i \in I} 1_R(i)} = \frac{\sum_{i \in I} 1_R(i)}{n} = \frac{|R|}{n} \tag{2.2}$$

The QF range is $[0, 1]$. The higher the score, the fairer the recommendation. Along with the relative comparison of the QF scores, the absolute values of QF scores are also taken into account in [147, 264], where a score of 1 means that all items in the dataset are in the top k at least once. Formally, QF (Eq. 2.2) is equivalent to Coverage [98], a measure of diversity, which has been used to evaluate fairness too [147].

2.2.2.3 Entropy (Ent) [200]

Ent measures how uniform the exposure of the recommended items is [147, 148, 167]. In [167], Lorenz curves were used to detect massive differences in individual item exposures and therefore, Entropy-like measure was proposed to quantify the inequality of item exposure:

$$\mathrm{Ent}_{\mathrm{ori}} = -\sum_{i \in I} p(i) \log p(i) \qquad \text{and} \qquad p(i) = \frac{\sum_{u \in U} 1_{R_u^k}(i)}{km} \tag{2.3}$$

where $p(i)$ is the recommendation frequency of i, i.e., how often item i is recommended in the top k to any user in the dataset, divided by the available recommendation slots km. In [167], log is the log base-n. It is unclear what log base is used in [147, 148]. When the log base is b, Ent ranges between $[0, \log_b n]$ while for log base-n, the range is $[0, 1]$.

In the past, this measure has been used by [147, 148, 167] to compare recommender models using absolute values, where a higher Ent is interpreted as having a more uniform distribution of the recommended items, and thus fairer.

2.2.2.4 Gini Index (Gini) [85]

Gini is a measure of variability i.e., the mean difference from all observed quantities [37]. It is most commonly used to measure inequality in the distribution of economic income, where the intuition is that a Gini score of 1 means that one entity receives all the income. Similarly in RSs, it is used to measure how much the distribution of item exposure deviates from an equal/uniform distribution [59, 60, 147]. Formally, Gini is defined as follows:

$$\text{Gini}_{\text{ori}} = \frac{\sum_{j=1}^{n}(2j-n-1)Ex_j}{n\sum_{j=1}^{n}Ex_j} \quad \text{and} \quad Ex_j = \sum_{u\in U}\sum_{w=1}^{W} 1_{R^k_{u,w}}(x_j)\cdot e_{(\cdot)}(u,x_j,w) \tag{2.4}$$

where x_j is the item with the j-th least amount of Ex_j, the total exposure received by that item across W rounds of recommendations.[4] W is the number of rounds, and $1_{R^k_{u,w}}(i) = 1$ when item i is in user u's top k recommendation list in round w. The examination function $e_{(\cdot)}(u, x_j, w)$ used in [147] is uniform, $e_{\text{unif}}(u, x_j, w) \equiv 1$ and in [59, 60] the examination function e_{DCG} (see Tab. 2.2) is used. We refer to the latter case as Gini-w. Gini and Gini-w's range is $[0, 1]$, where 0 means that there is an equal distribution of item exposure for all items in the dataset (fairest case). The range is important for the interpretability of the measure. Having an interpretable range is more important for how Gini and Gini-w quantify fairness when looking at the absolute values of the measure, like in [60, 147], than when looking at the difference in model rankings, like in [59].

2.2.2.5 Fraction of Satisfied Items (FSat) [167]

FSat is defined in the context of maximin-shared fairness, where fairness means that each item is recommended at least $\frac{km}{n}$ times, as there are only km slots that should ideally be distributed equally between n items. However, this distribution is impossible if the number of slots is not divisible by the number of items in the dataset, $n \nmid km$. The requirement is relaxed to recommending each item at least $\left\lfloor \frac{km}{n} \right\rfloor$ times (the maximin share) for it to be a fair recommendation. An item is satisfied iff its exposure is more than or equal to the maximin share. FSat measures the number of satisfied items divided by the total number of items:

$$\text{FSat}_{\text{ori}} = \frac{1}{n}\sum_{i\in I}\delta\left(\sum_{u\in U} 1_{R^k_u}(i) \geq \left\lfloor \frac{km}{n} \right\rfloor\right) \tag{2.5}$$

where $\delta(\cdot) = 1$ when the expression $\cdot$ is True and 0 otherwise. FSat has a range of $[0, 1]$, and the higher, the fairer. The range of values matters, for example [167] has used both absolute values and difference in values to interpret FSat.

[4] This is the total exposure received by each item, including items that are not recommended to any users. The scores are sorted as $Ex_1, Ex_2, \ldots Ex_n$. Ties between Ex_j's do not affect the final score.

2.2.2.6 Violation of Coverage Disparity (VoCD) [227]

VoCD is a fairness constraint. In [227], VoCD is used to optimise the recommendations for fairness during the training process, but not used for evaluating the final recommendation output. We include VoCD in this work to provide insights into what transpires when VoCD is used as an evaluation measure in its current formulation. VoCD is also the only measure operating on the Lipschitz condition [63], which requires similar individuals to be treated similarly. The idea behind VoCD is that any two α-similar recommended items should receive similar coverage. Two distinct items $i, i' \in R$ are α-similar if $d(i, i') = 1 - sim(i, i') \leq \alpha$, where $d(i, i')$ is the cosine distance and $sim(i, i')$ is the cosine similarity between the embeddings[5] of item i and i' respectively, and α is a parameter. Similar coverage means that the Coverage Disparity (CD) of those items, which is proportional to their exposure difference, must not exceed a threshold β. VoCD thus measures the average violation of the maximum allowed coverage disparity β in all pairs of α-similar recommended items:

$$\begin{aligned} \text{VoCD}_{\text{ori}} &= \frac{1}{|A|} \sum_{\forall (i,i') \in A} \max\left(CD(i, i') - \beta, 0\right) \\ CD(i, i') &= \left| \frac{\sum_{u \in U} 1_{R_u^k}(i) - \sum_{u \in U} 1_{R_u^k}(i')}{\max\left(\sum_{u \in U} 1_{R_u^k}(i), \sum_{u \in U} 1_{R_u^k}(i')\right)} \right| \end{aligned} \tag{2.6}$$

where A is the set of all pairs of α-similar items and $CD(i, i')$ is the coverage disparity between item i and i'. VoCD $= 0$ means that there is no violation of coverage disparity between any pairs (i.e., fair). In [227], the absolute value of VoCD affects the parameter that is used to control fairness during the training process. VoCD is customisable w.r.t. fairness and similarity: a lower[6] α means a stricter similarity requirement and a lower β means a stricter fairness requirement.

2.2.2.7 Individual-user-to-individual-item disparity (II-D) [234]

II-D was first defined by [56] to quantify the mean squared difference between system exposure and random exposure in individual queries and individual items, where there is a distribution of rankings (stochastic rankings). II-D is a resulting component

[5]The original paper specifically defines that they use embeddings, but in principle, any representation could work.

[6]This is *lower* instead of *higher* as the original authors [227] defined the term α-similar items based on the cosine distance between the two items being not more than α.

of decomposing another measure in the original work, which quantifies item fairness proportional to item relevance to users. It was redefined by [234] for W rounds of recommendations in RSs as:

$$\text{II-D}_{\text{ori}} = \frac{1}{m}\frac{1}{n}\sum_{u\in U}\sum_{i\in I}\left(E_{u,i} - E^{\sim}_{u,i}\right)^2 \tag{2.7}$$

$$E_{u,i} = \frac{1}{W}\sum_{w=1}^{W} 1_{R^k_{u,w}}(i)\cdot e_{\text{RBP}}(u,i,w) \qquad \text{and} \qquad E^{\sim}_{u,i} = \frac{1-\gamma^k}{n(1-\gamma)} \tag{2.8}$$

where $E_{u,i}$ is the expected exposure of i to u as per a stochastic ranking policy, $E^{\sim}_{u,i}$ is the expected exposure of i to u based on a uniformly random distribution over all permutations of items, and γ is an arbitrarily-set parameter for user patience. The examination function based on RBP (see Tab. 2.2) is used in $E_{u,i}$ and the equation of $E^{\sim}_{u,i}$ is derived based on the same examination function [234]. The range of II-D is not well-known,[7] but a lower value means a fairer recommendation. In [234], min-max normalisation is performed on II-D post-computation, such that the range is $[0, 1]$. This range of values matters; for example [234] has used the absolute values of II-D to analyse fairness and relevance trade-off, in addition to looking at the difference in model rankings based on II-D scores.

2.2.2.8 All-users-to-individual-item disparity (AI-D) [234]

AI-D computes the mean of the squared difference between system exposure and random exposure in each item. AI-D is similar to II-D in the sense that it is originally used for multiple rounds of recommendations and also a component resulting from the decomposition of another measure proposed by [234] that considers fairness w.r.t. relevance. However, unlike II-D, AI-D is sensitive to whether an item is recommended to multiple users due to the aggregation of the difference in exposure being done per item.

$$\text{AI-D}_{\text{ori}} = \frac{1}{n}\sum_{i\in I}\left(\frac{1}{m}\sum_{u\in U}E_{u,i} - \frac{1}{m}\sum_{u\in U}E^{\sim}_{u,i}\right)^2 \tag{2.9}$$

where $E_{u,i}$, $E^{\sim}_{u,i}$ are as per Eq. (2.8). The range of AI-D is not well-known, but a lower value means fairer recommendation. Like on II-D, a post-computation min-max normalisation is also performed by [234] on AI-D, resulting in a $[0, 1]$-range. The range of values matters; for instance [234] has used the absolute values of AI-D to

[7] The authors of the original measure did not state the range of II-D.

analyse fairness and relevance trade-off, on top of looking at the difference in model rankings based on AI-D scores.

2.3 Measure Limitations

We identify 5 theoretical limitations in the measures presented in §2.2 (summarised in Tab. 2.4). We use the term 'limitation' in the sense that regardless of the reason, a measure fails to quantify or fulfill properties that are important for evaluating fairness. Some of these limitations rarely occur, e.g. related to edge cases (§2.3.4), yet some are more likely to occur in practical scenarios (§2.3.1 & §2.3.3). In the headings, we put the name of the affected measures in brackets.

In practice, even if the limitation transpires by design, the design of the measure still restricts its usage under the conditions that we explain below. The identified limitations are independent of the recommender algorithm, as long as the recommender is a top-k recommender, which is the most common recommendation scenario in practice. We accompany each measure name by ↑ or ↓, denoting that the higher (↑) or the lower (↓) the score of the measure, the fairer the recommendation.

2.3.1 Limitation 1: Non-realisability

This is a novel limitation identified by us and affects *all measures*. We define non-realisability as the limitation whereby the max/min score of the evaluation measure cannot be reached at the top-k. As argued in [153], a desirable property of effectiveness measures is their realisability. While the realisability property in [153] is related to the number of relevant items, non-realisability for fairness measures is related to the number of recommendation slots (km) and the number of items that are in the dataset (n); we explain this relationship below as part of the causes of this limitation. In practice, the non-realisability limitation makes fairness scores hard to interpret because when the worst or best possible fairness score varies based on the dataset (m, n) and experimental choice of threshold k, it is unknown whether the fairness score obtained for a model is closer to the max or min score. For instance, if a higher-is-fairer measure ranges in $[0, 1]$ and a model achieves a score of 0.2, one might think that the model is not very fair. However, if the maximum achievable score for that case (e.g., if all top k items are fair) is 0.22, then the model might actually be fair, but this cannot be known from the score of the evaluation measure. We identify four different causes of non-realisability.

2.3.1.1 Cause 1 (↑Jain, ↑QF, ↑Ent, ↑FSat, ↓Gini, ↓Gini-w)

Non-realisability can occur if the most unfair score is only given to an unrealistic recommendation scenario, which we explain next as it differs per measure. Specifically, the score can never be 0 for ↑Jain, unless the number of slots is 0, i.e., $k = 0$ or $m = 0$, which does not make sense because $k = 0$ means that no recommendation at all is outputted, and $m = 0$ means that there are zero users to recommend to. The score of ↑Ent can only be 0 when there are no items in the dataset ($n = 0$), which also does not make sense because it means that there is nothing to recommend. For ↑QF/FSat, the score can only be 0 if there are *no* recommended items to *any* users ($|R| = 0$), which does not make sense either. On the other hand, the score can only be 1 for ↓Gini/Gini-w when a single item is recommended at all k slots for each user, which is a highly unlikely artificial outcome; to our knowledge, no reasonably performing recommender model can produce such an output. All of the above conditions are unrealistic. A consequence of the above is that fairness is overestimated by these measures. Instead of the unrealistic situations above, it is the realistically unfairest recommendation for ↑Jain/QF/Ent/FSat that should be mapped to 0 and for ↓Gini/Gini-w to 1.

2.3.1.2 Cause 2 (↑Jain, ↑QF, ↑Ent, ↓Gini, ↓Gini-w, ↑FSat, ↓II-D, ↓AI-D)

The second cause of non-realisability that we identify is that when the number of recommendation slots km is less than the number of items n, then the score cannot be the fairest, as some items cannot be exposed due to the limited availability of recommendation slots. This means that the max score (most fair) cannot be reached by the measure, even if all recommended items are fair. In the datasets in Tab. 2.6, $km > n$ for $k \in \{10, 20\}$, but for some very large datasets like LFM-1b [196] or Foursquare NYC [241], $km < n$.

For II-D and AI-D, we show that the score cannot be the fairest, during single or multiple rounds of recommendations. For example, given $k = 1, m = 2, n = 3$, and considering all possible orders of recommendations for a single round of recommendation, the lowest scores for II-D and AI-D are $2/9$ and $1/18$ respectively. In the case of multiple recommendation rounds, the total number of slots is kmW, where W is the number of rounds. As an example, given $k = 1, m = 2, W = 2, n = 5$, the lowest possible II-D and AI-D are 0.06 and 0.01 respectively. So, in these cases, the lowest score (that should be the fairest) is not zero, even if the recommendations are made as fair as possible (close to random exposure).

This leads to an underestimation of fairness: the fairest value of the measure cannot be reached for some datasets (regardless of recommendation quality), so the measure is not evenly robust across datasets and can underestimate fairness.

2.3.1.3 Cause 3 (↑Jain, ↑Ent, ↓Gini, ↓Gini-w ↓II-D, ↓AI-D)

The third cause of non-realisability that we identify is that for Jain, Ent, Gini, and Gini-w, even when $km > n$, the measure still cannot reach the theoretical fairest value if the number of items is not an exact multiple of the recommendation slots, $n \nmid km$. E.g., if $km = 4$, $n = 3$, three slots can be filled with one unique item each, but no matter which item fills the last slot, one item will be recommended one more time than the rest. Likewise, the same applies for II-D and AI-D when the number of slots across all rounds (kmW) is not divisible by n. For example, when $k = m = W = 2, n = 3$, the minimum II-D is 0.02 and the minimum AI-D is 0.005. This limitation consequently leads to the same issue of robustness and underestimation of fairness, as described in *Cause 2*.

2.3.1.4 Cause 4 (↓Gini-w, ↓VoCD, ↓II-D, ↓AI-D)

The fourth case of non-realisability that we identify is that measures cannot reach the theoretical (un)fairest value, as the exact formulation of the max/min achievable score is unknown,[8] making the score hard to interpret. This happens because the most/least fair recommendation cannot be analytically determined due to parameters in the measure or item exposure being weighted by a non-uniform examination function. This causes the measure to have a range different from its theoretical range, i.e., $[0, 1]$, which we explain next for each measure.

↓Gini-w may not reach 0 or 1 due to non-uniform exposure. E.g., when $k = n = 3, m = 2$, the minimum and maximum ↓Gini-w for all possible recommendation lists are 0.0373 and 0.156 respectively. As $n|km$, this is separate from **non-realisability**, *Causes 1–3*.

For ↓VoCD, as $CD(i, i') \in [0, 1)$, ↓VoCD $\in [0, 1 - \beta)$. However, the score of ↓VoCD depends on item similarity, making the most unfair score unreachable. Even though it is impossible to formulate an exact achievable maximum value for ↓VoCD, we formally prove in App. 2.9.1.6 that $\forall \alpha \in [0, 2], \beta \in [0, 1)$, the maximum ↓VoCD, $\text{VoCD}_{\max} \leq \frac{m-1}{m} - \beta$. This is obtained when there is only one pair of similar items which is recommended 1 and m times each. E.g., $R^2_{u_1} = [i_1, i_2]$, $R^2_{u_2} = R^2_{u_3} = [i_1, i_3]$

[8]Here, 'unknown' is only related to the exact max/min formulation, as the maximum and minimum can always be computed by enumerating all possibilities, albeit being a costly process.

where $\text{VoCD}_{\max}$ is $2/3$, which happens when only i_1 and i_2 are similar. So, the most unfair score is non-realisable, as Eq. (2.6) depends on item-pair similarity.

↓II-D and ↓AI-D also may not reach 0 or 1, even in the context of multiple rounds of recommendations and having enough slots for all items. For example, when $k = m = W = 2, n = 3$, considering all possible ways of recommending items, the minimum values of II-D and AI-D are 0.02 and 0.005 respectively, while the maximum values are 0.187 for both. We posit this to be due to the exponential-like exposure.

A summary of situations producing the theoretical most (un)fair scores in existing measures is given in Tab. 2.3.

2.3.2 Limitation 2: Quantity-insensitivity (QF_{ori})

This limitation is part of the design choice by the authors of the original QF measure [264] based on a specific concept of fairness, which we explain next. Quantity-insensitivity means that the measure ignores how often an item is recommended across all users in a recommendation round. In economics, [7] states that 'sensitivity to transfer' is a basic criterion of an inequality measure. Similarly, we think that when exposure increases (or decreases) for an item, the fairness measure should be sensitive to the change. Meanwhile, QF_{ori} makes no distinction between items that are recommended once or more than once. To illustrate, consider these scenarios: 1) $R^2_{u_1} = [i_1, i_2], R^2_{u_2} = [i_2, i_3], R^2_{u_3} = [i_1, i_3]$ and 2) $R^2_{u_1} = R^2_{u_2} = [i_1, i_2], R^2_{u_3} = [i_1, i_3]$. Assuming $n = 5$, QF_{ori} would be 0.6 for both cases, even though in the second scenario i_1 is recommended more times than i_2, which is recommended more than i_3.

As a result of this design choice, the score does not reflect the repeated recommendations of the same item to many users, which may indicate unfairness (e.g., popularity bias). This is a design limitation that one should be aware of when using QF.

2.3.3 Limitation 3: Undefinedness (Ent_{ori})

We define the limitation of undefinedness as the measure giving an undefined value. In practice, this limitation renders the measure incomputable when encountering an (edge) case.[9] This is not negligible, as an assessment question for the design of (fairness) measures is related to how the measure responds to edge cases [175]. For Ent_{ori}, the case is related to the possibility of encountering the undefined value of $\log 0$ during computation. Undefinedness happens when there is at least one item from

[9]This is reminiscent of the completeness property in [153].

Table 2.3: Situations that produce the theoretical most (un)fair score in existing individual item fairness measures.

Measure	Most Unfair	Most Fair
Jain	no recommendation slots ($km = 0$)	all items recommended the same amount
QF	no items recommended ($\|R\| = 0$)	all items exposed, no matter how many times
Ent	no items in the dataset ($n = 0$)	all items recommended the same amount
Gini, Gini-w	a single item is recommended at all rank positions for all users	all items recommended the same amount (or same total exposure weight)
FSat	no items recommended ($\|R\| = 0$)	all items recommended $\geq \lfloor \frac{km}{n} \rfloor$ times
VoCD	not possible to deduce from the formula and description	all pairs of similar recommended items recommended similar amount with a normalised difference
II-D	not possible to deduce from the formula and description	exposure distribution of recommended items matches exposure given by random distribution
AI-D	not possible to deduce from the formula and description	exposure distribution of recommended items matches exposure given by random distribution, with the most possible number of unique items in top k

the dataset that does not appear in the recommendations at all,[10] which happens often because not all items in the datasets are guaranteed to be at the top k. For example, given $R^2_{u_1} = R^2_{u_2} = [i_1, i_2]$ and $I = \{i_1, i_2, i_3\}$, the value of $p(i_3) = 0$, and this entails $\log p(i_3) = \log 0$. Such a situation is common, as it will later be seen in Tab. 2.7, and therefore we do not consider this as an edge case. When the measure is incomputable for several models, its interpretation is less meaningful.

We exclude the case where no item is recommended to any users ($|R| = 0$), as this is a trivial case and it does not make much sense to evaluate fairness when there is

[10] $\log p(i)$ in Eq. 2.3 is undefined if item i is not in the recommendation list for any users ($p(i) = 0$).

no item being recommended. Regardless of the triviality, we identify some measures that are incomputable under this edge case: Jain_{ori}, Ent_{ori}, Gini_{ori}, $\text{Gini-w}_{\text{ori}}$, and VoCD_{ori}. Note that we do not consider these four measures to have the undefinedness limitation just for this reason.

2.3.4 Limitation 4: Always-fair (↑FSat)

We define the limitation of always-fair as the measure giving the fairest score regardless of the content of the recommendation list, under a specific condition depending on the particular measure. This happens for ↑FSat when $km < n$, as empirically discovered by [167]. The maximin share, in this case, is 0 and all items are deemed satisfied as per the definition in §2.2.2.5, regardless of the actual distribution of recommended items. While this is partly due to the design choice of FSat which is based on the maximin share, this means that ↑FSat will always be 1, rendering the measure unsuitable for use cases where $km < n$. Even though this limitation has been empirically identified before, there was no formal definition of it, and that is what we do here.

2.3.5 Limitation 5: Item-representation-dependence (↓VoCD)

We formally identify this limitation in this work, even though it is part of an intentional choice of the measure, as opposed to an accidental or unforeseen byproduct of the design. Item-representation-dependence means that the score of the measure varies according to how item representations are built (e.g., embedding, graphs). The max value of ↓VoCD depends on which item pairs are similar based on how items are represented. Even though the dependence on item representation is part of the design choice for VoCD, the limitation of this design should be taken into consideration when one chooses a fairness measure, e.g. ensuring a same way of representing items for a fair comparison.

Depending on how item representations are built, there may be different pairs of similar items in the set A, yielding different ↓VoCD scores. E.g., given $R^2_{u_1} = [i_1, i_2]$, $R^2_{u_2} = [i_1, i_3]$, if $A = \{(i_1, i_2)\}$ as per one item representation, ↓VoCD $= 0.5 - \beta$. However, if according to another item representation, $A = \{(i_2, i_3)\}$, then ↓VoCD $= 0$. While the former score represents a somewhat unfair RS, the latter denotes that the RS is fair. Note that one can use the same recommendation algorithm and the same dataset, but with different ways of representing the items, different VoCD scores may be obtained. Therefore, the limitation still holds even when the comparison is only performed within an algorithm and a dataset.

Table 2.4: Measures of individual item fairness and their theoretical limitations. We identify four different causes for non-realisability, denoted by C in this table.

Legend ●: we fully resolve the limitation ○: the limitation is unresolvable ✓: another measure resolves the limitation	Source	Jain [101]	QF [264]	Ent [200]	Gini [85]	Gini-w [60]	FSat [167]	VoCD [227]	IL-D [234]	AL-D [234]
non-realisability: cannot reach max/min score (cause number denoted by C)										
C1. Most unfair score is only given to an impossible scenario	us	●	●	●	●	●	●			
C2. Fewer recommendation slots compared to number of items	us	●	●	●	●	●	●		○	○
C3. Number of recommendation slots is indivisible by number of items	us	●		●	●	○			○	○
C4. Non-realisability due to unknown formulation of max/min score	us					○		○	○	○
quantity-insensitivity: ignores frequency of item recommendation	[264]		✓							
undefinedness: cannot be computed (undefined value)	us			●						
always-fair: gives fairest score regardless of recommendation contents	[167]						○			
item-representation-dependence: depends on how items are represented	us							○		

2.4 Resolving Limitations

We explain how we resolve each limitation or why it is unresolvable. For the remainder of this paper, we refer to the original version of an evaluation measure M as M_{ori}, and to our modified version of an evaluation measure M as M_{our}. When $\cdot_{\text{ori}}$ or $\cdot_{\text{our}}$ is not specified, we refer to both the original and modified version simultaneously.

2.4.1 Resolving Non-realisability (Limitation 1) and Undefinedness (Limitation 3)

We resolve causes 1, 2, and 3 of non-realisability via post-calculation correction of under/overestimated fairness scores based on the theoretical min and max values for a scenario with limited km recommendation slots. Specifically, we rescale the range of the measures to the actual theoretically achievable most fair and unfair values. The rescaled measures either retain the $[0, 1]$-range, or are now ranged in $[0, 1]$. To rescale, we compute the measure's upper and lower bounds when possible (see Tab. 2.5 and App. 2.9.1 for the derivations) by considering the most fair and unfair recommendation case for each measure, which we explain next.

For the measure that suffers from non-realisability due to *Causes 1–2* (and quantity-insensitivity) i.e., QF, we posit that the most unfair recommendation @k is when the same k unique items are recommended to each of the m users, resulting in the min exposure for items in I. Therefore, each of these k items is recommended m times. This is equivalent to the recommendation generated by Pop [182] that gives

Table 2.5: Most (un)fair score @k. Note that the Most Fair @k scores (except for Gini-w) remain the same as the theoretical fairest value (i.e., 0 or 1) when the number of recommendation slots is divisible by the number of items, $n \mid km$.

Measure	Most Unfair @k	Most Fair @k
$\uparrow$Jain$_{\text{ori}}$	$\frac{k}{n}$	$\frac{(km)^2}{n\left(n\left\lfloor\frac{km}{n}\right\rfloor^2+(km \bmod n)\left(2\left\lfloor\frac{km}{n}\right\rfloor+1\right)\right)}$
$\uparrow$QF$_{\text{ori}}$	$\frac{k}{n}$	$\min\left(\frac{km}{n}, 1\right)$
$\uparrow$Ent$_{\text{ori}}$	$\log k$	Eq. (2.13)
$\downarrow$Gini$_{\text{ori}}$	$1-\frac{k}{n}$	$\frac{(n-km \bmod n)(km \bmod n)}{kmn}$
$\downarrow$Gini-w$_{\text{ori}}$	Eq. (2.20)	Eq. (2.19) when $km \leq n$
$\uparrow$FSat$_{\text{ori}}$	$\frac{k}{n}$	1
$\downarrow$VoCD$_{\text{ori}}$	$\frac{m-1}{m}-\beta$	0

the same k most popular items to all users. We posit that the most fair case @k is when $\min(km, n)$ unique items are recommended to m users, i.e., the max number of items allowed by the recommendation slots is exposed.

For measures that suffer from non-realisability, due to *Causes 1–3* but not quantity-insensitivity (Jain, Ent, Gini, Gini-w, FSat), we consider the most fair recommendation to be when $(n - km \bmod n)$ items are exposed $\left\lfloor\frac{km}{n}\right\rfloor$ times and the rest $(km \bmod n)$ items are exposed $\left\lfloor\frac{km}{n}\right\rfloor + 1$ times. The most unfair case for Jain, Ent, Gini, Gini-w, and FSat is the same as QF.

We then perform min-max normalisation (hereafter referred to as normalisation) using the most unfair/fair bounds as the min/max possible value of the measure. The general process is: let $x_{\max}$ be the max possible value of a fairness score x, and $x_{\min}$ be the min possible value of x. The normalised score of x, denoted by x', is calculated using $x' = \frac{x-x_{\min}}{x_{\max}-x_{\min}}$. Note that this normalisation does not work when $x_{\max} = x_{\min}$ due to division by zero. This happens when the most unfair recommendation is equal to the fairest recommendation, which occurs when $k = n$. We therefore exclude this case from our corrections.

After normalisation, the measures quantify fairness of items by considering the following components: the recommendation list, the number of items in the dataset, and the most fair and most unfair recommendation. As a result, the most fair recommendation scenario and most unfair recommendation scenario are now mapped to the endpoints, instead of the unrealistic scenarios in Tab. 2.3. Next, we explain in detail the normalisation of each measure.

For $\uparrow\text{Jain}_{\text{ori}}$, we apply the following normalisation on Eq. (2.1): as the higher the score, the fairer, we use Eq. (2.10) as $x_{\max}$ and $x_{\min} = \frac{k}{n}$ because these are the most fair and unfair @k values. Eq. (2.10) simplifies to $y = \frac{km}{n}$ when $km < n$, and simplifies to 1 when $n \mid km$.

$$\text{Jain}_{\max} = \frac{(km)^2}{n\left(n\left\lfloor\frac{km}{n}\right\rfloor^2 + (km \bmod n)\left(2\left\lfloor\frac{km}{n}\right\rfloor + 1\right)\right)} \tag{2.10}$$

$$\text{Jain}_{\text{our}} = \frac{\text{Jain}_{\text{ori}} - \frac{k}{n}}{\text{Jain}_{\max} - \frac{k}{n}} \tag{2.11}$$

We normalise $\uparrow\text{QF}_{\text{ori}}$ similarly to the above. As the higher the QF score, the fairer, we use $x_{max} = \min(y, 1)$ and $x_{min} = \frac{k}{n}$.

$$\text{QF}_{\text{our}} = \begin{cases} \frac{\text{QF}_{\text{ori}} - \frac{k}{n}}{1 - \frac{k}{n}} = \frac{|R| - k}{n - k} & \text{if } km \geq n \\ \frac{\text{QF}_{\text{ori}} - \frac{k}{n}}{\frac{km}{n} - \frac{k}{n}} = \frac{|R| - k}{k(m-1)} & \text{otherwise} \end{cases} \tag{2.12}$$

We acknowledge that by performing this normalisation on QF_{ori}, there is now an additional way of measuring fairness. The new score can now be interpreted as "given that each item should be recommended at least once, how fair the recommendation is w.r.t. the most unfair and the fairest recommendation". The most (un)fair recommendation and the normalisation depend on the number of recommendation slots, and we argue that it is better to use this number instead of using the number of items in the datasets; in practice, the number of items shown to the users is almost always limited. However, if QF is meant to be used for detecting the percentage of items that are exposed, then the QF_{ori} may be used at the expense of needing additional information on what is the best possible QF score, to know the limit of how fair the recommendation can be.

For $\uparrow\text{Ent}_{\text{ori}}$, as the higher the score, the fairer, we use Eq. (2.13) as $x_{\max}$ when $km \geq n$ or $x_{\max} = \log km$ otherwise, and $x_{\min} = \log k$ for normalisation. Eq. (2.13) simplifies to $\log n$ when $n \mid km$.

$$\text{Ent}_{\max} = -(n - km \bmod n)\left(\frac{\left\lfloor\frac{km}{n}\right\rfloor}{km}\log\frac{\left\lfloor\frac{km}{n}\right\rfloor}{km}\right) \tag{2.13}$$

$$-(km \bmod n)\left(\frac{\left\lfloor\frac{km}{n}\right\rfloor + 1}{km}\log\frac{\left\lfloor\frac{km}{n}\right\rfloor + 1}{km}\right) \tag{2.14}$$

However, $\uparrow\text{Ent}_{\text{ori}}$ still suffers from **undefinedness**, which we resolve by restricting the sum over i in Eq. (2.3) to only recommended items:

$$\text{Ent}_{\text{def}} = -\sum_{i \in R} p(i) \log p(i) \tag{2.15}$$

where $p(i)$ is calculated via Eq. (2.3) and $p(i) > 0$ because $i \in R$. Performing normalization on Eq. (2.15), we obtain:

$$\text{Ent}_{\text{our}} = \begin{cases} \frac{\text{Ent}_{\text{def}} - \log k}{\text{Ent}_{\text{max}} - \log k} & \text{if } km \geq n \\ \frac{\text{Ent}_{\text{def}} - \log k}{\log km - \log k} = \frac{\text{Ent}_{\text{def}} - \log k}{\log m} & \text{otherwise} \end{cases} \tag{2.16}$$

For $\downarrow\text{Gini}_{\text{ori}}$, as the lower the score, the fairer, we use Eq. (2.17) as x_{min} and $x_{\text{max}} = 1 - \frac{k}{n}$. Eq. (2.17) simplifies to $1 - y$ when $km > n$, and simplifies to 0 when $n \mid km$.

$$\text{Gini}_{\text{min}} = \frac{(n - km \bmod n)(km \bmod n)}{kmn} \tag{2.17}$$

$$\text{Gini}_{\text{our}} = \frac{\text{Gini}_{\text{ori}} - \text{Gini}_{\text{min}}}{1 - \frac{k}{n} - \text{Gini}_{\text{min}}} \tag{2.18}$$

For $\downarrow\text{Gini-w}_{\text{ori}}$, we only consider cases where $km \leq n$ as finding the most fair recommendation list across all users for the other cases is analytically not possible. The problem does not have a closed form solution and computing the solution requires solving a constrained optimization that considers all possible permutations of recommendations across users. We use Eq. (2.19) as x_{min} when $km \leq n$, $x_{\text{min}} = 0$ otherwise, and Eq. (2.20) as x_{max} to normalise $\downarrow\text{Gini-w}_{\text{ori}}$.

$$\text{Gini-w}_{\text{min}} = \frac{\sum_{\ell=1}^{k} \sum_{j=n-\ell m+1}^{n-\ell m+m} (2j - n - 1) \log_{\ell+1} 2}{mn \sum_{\ell=1}^{k} \log_{\ell+1} 2} \tag{2.19}$$

$$\text{Gini-w}_{\text{max}} = \frac{\sum_{\ell=1}^{k} (n - 2\ell + 1) \log_{\ell+1} 2}{n \sum_{\ell=1}^{k} \log_{\ell+1} 2} \tag{2.20}$$

$$\text{Gini-w}_{\text{our}} = \begin{cases} \frac{\text{Gini-w}_{\text{ori}} - \text{Gini-w}_{\text{min}}}{\text{Gini-w}_{\text{max}} - \text{Gini-w}_{\text{min}}} & \text{if } km \leq n \\ \frac{\text{Gini-w}_{\text{ori}}}{\text{Gini-w}_{\text{max}}} & \text{otherwise} \end{cases} \tag{2.21}$$

For $\uparrow\text{FSat}_{\text{ori}}$, as the higher the score, the fairer, we normalise Eq. (2.5) using $x_{\max} = 1$ and $x_{\min} = \frac{k}{n}$.[11]

$$\text{FSat}_{\text{our}} = \frac{\text{FSat}_{\text{ori}} - \frac{k}{n}}{1 - \frac{k}{n}} \tag{2.22}$$

2.4.2 Resolving Quantity-insensitivity (Limitation 2)

While the quantity-insensitivity limitation is due to the design choice of the QF measure (§2.3.2), we reason that a solution to this limitation is needed in case one would like to use a measure that is equation-wise similar to QF_{ori}, but needs the measure to be sensitive to the change of exposure received by the item. Hence, the **quantity-insensitivity** limitation for $\uparrow\text{QF}_{\text{ori}}$ may simply be resolved by calculating $\uparrow\text{Jain}_{\text{ori}}$ instead, as $\uparrow\text{QF}_{\text{ori}}$ originates from $\uparrow\text{Jain}_{\text{ori}}$ (see Eq. 2.1 and 2.2). $\uparrow\text{Jain}_{\text{ori}}$ gives different weights between items that have been recommended once and more than once. Referring to the toy example given in §2.3.2 where $\uparrow\text{QF}_{\text{ori}}$ would be 0.6 for both cases, but $\uparrow\text{Jain}_{\text{ori}} \approx 0.514$ for the first scenario and $\uparrow\text{Jain}_{\text{ori}} = 0.6$ for the second scenario. $\uparrow\text{Jain}_{\text{ori}}$ returns a higher fairness score in the second scenario as the distribution of the frequency count of items is more balanced.

2.4.3 Unresolvable Limitations

We resolve three out of five limitations (see Tab. 2.4). The remaining limitations are unresolvable for the following reasons.

Non-realisability due to *Cause 4* cannot be resolved because no closed-form solutions have been discovered for the recommendations that produce the best and worst fairness scores for each measure. Computing the solution requires solving a constrained optimization problem that cannot be practically solved for large datasets such as RSs data. While the measures could theoretically be corrected in a similar manner to the ones in §2.4.1, it is only possible to do so after computing the most/least fair recommendation list, which is impractical.

Item-representation-dependence cannot be resolved because item representation is required by the measure to determine whether items are similar. It is avoidable if all items are considered similar to each other, but this is unrealistic. Moreover, to get comparable scores, all RSs should use the same representation of items, which cannot be guaranteed. This point is not handled in the original definition of VoCD

[11] VoCD, II-D, and AI-D can be normalised in a similar way after computationally approximating their empirical minimum and maximum values.

[227], where the only item representation considered is item embeddings. While it is possible to use the same external representation for item representation (e.g. similarity matrix based on tags, keyword, or other features) for different RSs, as commonly done with diversity metrics [269], the fairness score may still change depending on the representation used to determine the item similarity.
Always-fair cannot be resolved for FSat_{ori} because attempting to resolve this would require tampering with the definition of 'satisfied' based on the concept of maximin share. This definition of 'satisfied' is an integral part of the measure. Replacing the maximin share criterion with another requirement would turn FSat_{ori} into a different measure. Therefore, as this limitation is related to the concept of measure, we are unable to recommend a solution to the limitation without changing the design of the measure.

2.5 Empirical Analysis

We experimentally analyse the relevance and fairness of several recommenders and compare the original measures (§2.2) to our corrected versions (§2.4) for the six datasets shown in Tab. 2.6. We present the results for Lastfm and Ml-1m in this section, and for the other datasets in App. 2.9.2.

2.5.1 Experimental Setup

Dataset preprocessing. We use six freely available datasets (see Tab. 2.6) from [260]:[12] Lastfm;[13] Ml-1m [94]; Book-x [269]; Amazon-lb, Amazon-dm, and Amazon-is [156]. We remove users/items with < 5 interactions and use 80%/10%/10% to train/validate/test, with a user-based random split for Lastfm and Book-x (timestamps are not available), and a user-based temporal split for all other datasets, i.e., the last 10% of each user's interactions are in the test set. We convert ratings ≥ 3 on Ml-1m & Amazon-* and ratings ≥ 6 on Book-x to 1. We discard the rest of the ratings. We choose these thresholds as the ratings are from 1–5 in Ml-1m and Amazon-*, and 0–10 in Book-x. We do not convert for Lastfm as it uses implicit feedback, so all interactions have a value of 1. For duplicate values, we keep the last interaction.
Recommenders. For recommendation we use: Pop [182] (recommends k most popular items), item-based K-Nearest Neighbours (ItemKNN) [55], Sparse Linear Method (SLIM) [158], Bayesian Personalized Ranking (BPR) [185], Neural Graph

[12] `https://github.com/RUCAIBox/RecSysDatasets`
[13] `http://www.lastfm.com`

Collaborative Filtering (NGCF) [226], Neural Matrix Factorization (NeuMF) [95], and Variational Autoencoder with multinomial likelihood (MultiVAE) [135]. We use training batch sizes of 4096, Adam [117] as optimizer, and the RecBole library [260]. We train BPR, NGCF, NeuMF, and MultiVAE for 300 epochs, but use early stopping of 10 epochs and keep the model that produces the best NDCG@10 on the validation set. We tune hyperparameters on all models except Pop, with RecBole's hyperparameter tuning module. The hyperparameter search space and optimal hyperparameters are in App. 2.9.2.1. For all recommenders, when we generate the recommendation list for a user during testing, the items in the user's train or validation set are placed at the end of the user's list to avoid re-recommending them.

Measures. We evaluate models w.r.t. a) relevance-only measures (HR, MRR, Precision (P), Recall (R), MAP, NDCG), and b) individual item fairness measures, both the original and our corrected measures. All measures are computed at $k = 10$, unless otherwise stated. We evaluate on the full test set of items instead of a sample of them, as doing the latter is known to yield misleading results [120]. This leads to lower performance than reported when sampling the test set. Lastly, for Ent, we use the log base-n. For VoCD we choose the values of α and β such that VoCD maintains comparability with the other fairness measures: all recommended items are considered similar[14] ($\alpha = 2$) and thus A is the set of all possible pairs of different items in the top k, without any tolerance for coverage disparity ($\beta = 0$). We also choose this configuration to avoid reliance on similarity scores based on item embeddings. For II-D and AI-D, we use $\gamma = 0.8$ [234].

2.5.2 Analysis of Relevance and Fairness

We start by studying the relevance and fairness of several recommender models. We compare a) different recommender models w.r.t. relevance and fairness scores, and b) different evaluation measures, including corrected and uncorrected measures. The goal of a) is to study whether relevance and/or fairness scores vary between models, and to obtain a ranking of models that is used in subsequent analysis (§2.5.3). The goal of b) is to study measures based on diverse concepts of fairness, and highlight their differences (and similarity, if any).

The evaluation results are shown in Tab. 2.7 for Lastfm and Ml-1m, and in App. 2.9.2.2 for the other datasets.

[14] All recommended items will be treated as similar items as $sim(i, i') \geq 1-2 = -1$ and $sim(i, i') \in [-1, 1]$.

Table 2.6: Statistics of the datasets before and after our preprocessing.

dataset	**#users**	**#items**	**#interactions**	**sparsity (%)**
original (as provided by [260])				
Lastfm[13]	1,892	17,632	92,834	99.7217%
Ml-1m [94]	6,040	3,706	1,000,209	95.5316%
Book-x [269]	105,283	340,556	1,149,780	99.9968%
Amazon-lb [156]	416,174	12,120	574,628	99.9886%
Amazon-dm [156]	840,372	456,992	1,584,082	99.9996%
Amazon-is [156]	1,246,131	165,764	1,758,333	99.9991%
preprocessed (by us)				
Lastfm	1,859	2,823	71,355	98.6403%
Ml-1m	6,038	3,307	835,789	95.8143%
Book-x	5,639	7,455	91,385	99.7826%
Amazon-lb	1,644	791	16,765	98.7108%
Amazon-dm	11,750	9,462	116,681	99.8951%
Amazon-is	6,574	3,569	45,762	99.8050%

2.5.2.1 Discussion of recommendation models in Tab. 2.7

BPR has the highest relevance scores, while MultiVAE generally has the highest fairness scores. Between BPR and MultiVAE, we observe that the relevance scores are higher in BPR but the fairness scores are higher for MultiVAE. E.g., in Lastfm, NDCG = 0.223 for BPR and NDCG = 0.219 for MultiVAE. Meanwhile, the higher-is-better fairness scores range between $[0.078, 0.656]$ for BPR and $[0.132, 0.763]$ for MultiVAE. This is not always observed for all models. E.g., for ItemKNN and SLIM, better fairness tends to be accompanied by better relevance. Furthermore, other discrepancies exist: the relevance of ItemKNN and MultiVAE is on par, but their fairness scores e.g., the scores of ↑Jain of ItemKNN, are only half of those achieved by MultiVAE. Generally, the recommenders agree in relative ordering of scores, but some models have higher scores for $\cdot_{\text{our}}$ than $\cdot_{\text{ori}}$ and vice-versa, e.g., for Lastfm, $\downarrow\text{Gini}_{\text{ori}} > \downarrow\text{Gini}_{\text{our}}$ for ItemKNN but $\downarrow\text{Gini}_{\text{ori}} < \downarrow\text{Gini}_{\text{our}}$ for SLIM. Overall, we observe that:

- A recommender model that is the best in terms of relevance may also be relatively fair.

Table 2.7: Relevance (REL) and fairness (FAIR) scores of the recommender models on Lastfm and Ml-1m. The most relevant and most fair score per measure is in bold. ↑ means the higher the better, ↓ the lower the better. 'nan' stands for 'not a number'.

			Pop*	ItemKNN	SLIM	BPR	NGCF	NeuMF	MultiVAE
Lastfm	REL	↑ HR	0.236686	0.563206	0.520710	**0.603012**	0.598171	0.571813	0.597633
		↑ MRR	0.102296	0.320710	0.290459	**0.336577**	0.327790	0.301449	0.326032
		↑ P	0.029855	0.082948	0.075901	**0.091286**	0.090479	0.082679	0.091070
		↑ MAP	0.033572	0.129943	0.112303	**0.140626**	0.136694	0.120650	0.136794
		↑ R	0.078205	0.240326	0.210221	**0.262193**	0.260140	0.238121	0.261565
		↑ NDCG	0.063259	0.207209	0.183180	**0.223416**	0.219215	0.198248	0.219408
	FAIR	↑ Jain_{ori}	0.005350	0.050544	0.029023	0.080549	0.086601	0.096789	**0.134035**
		↑ Jain_{our}	0.001824	0.047434	0.025715	0.077714	0.083822	0.094103	**0.131692**
		↑ QF_{ori}	0.009564	0.432519	0.145590	0.428268	0.399221	0.462983	**0.683316**
		↑ QF_{our}	0.006043	0.430501	0.142552	0.426235	0.397085	0.461074	**0.682190**
		↑ Ent_{ori}	nan	nan	nan	nan	nan	nan	nan
		↑ Ent_{our}	0.096360	0.595274	0.439907	0.656154	0.660951	0.686602	**0.763463**
		↑ FSat_{ori}	0.009564	0.136734	0.071909	0.171803	0.178888	0.195537	**0.230960**
		↑ FSat_{our}	0.006043	0.133665	0.068610	0.168859	0.175969	0.192677	**0.228226**
		↓ Gini_{ori}	0.995080	0.908416	0.968311	0.884752	0.885149	0.864875	**0.780612**
		↓ Gini_{our}	0.998564	0.908251	0.970668	0.883591	0.884004	0.862877	**0.775067**
		↓ $\text{Gini-w}_{\text{ori}}$	0.995984	0.915773	0.972331	0.895154	0.896870	0.879730	**0.793962**
		↓ $\text{Gini-w}_{\text{our}}$	0.998747	0.918313	0.975028	0.897637	0.899358	0.882170	**0.796164**
		↓ VoCD_{ori}	0.669135	0.609061	0.704415	0.640846	0.656609	0.641465	**0.598510**
		↓ II-D_{ori}	**0.000970**	**0.000970**	**0.000970**	**0.000970**	**0.000970**	**0.000970**	**0.000970**
		↓ AI-D_{ori}	0.000668	0.000061	0.000114	0.000037	0.000034	0.000032	**0.000019**
Ml-1m	REL	↑ HR	0.273435	0.336204	0.343326	**0.348791**	0.337198	0.324114	0.331070
		↑ MRR	0.112995	0.136303	0.139593	**0.142428**	0.140765	0.133319	0.130503
		↑ P	0.045661	0.054190	0.053445	**0.055167**	0.055018	0.052302	0.051308
		↑ MAP	0.025940	0.036132	0.036830	**0.038389**	0.037460	0.033568	0.035284
		↑ R	0.040195	0.064935	0.071923	**0.073647**	0.067244	0.060861	0.068755
		↑ NDCG	0.056219	0.073897	0.075873	**0.078110**	0.076032	0.070162	0.072263
	FAIR	↑ Jain_{ori}	0.006867	0.023143	0.045463	**0.068296**	0.057974	0.059396	0.065298
		↑ Jain_{our}	0.003857	0.020192	0.042593	**0.065508**	0.055148	0.056576	0.062499
		↑ QF_{ori}	0.040822	0.188388	0.236166	0.444209	0.306018	0.384336	**0.485939**
		↑ QF_{our}	0.037913	0.185927	0.233849	0.442524	0.303913	0.382469	**0.484380**
		↑ Ent_{ori}	nan	nan	nan	nan	nan	nan	nan
		↑ Ent_{our}	0.187196	0.435739	0.552900	0.650556	0.606893	0.623249	**0.652672**
		↑ FSat_{ori}	0.020865	0.069549	0.114605	0.164500	0.147263	0.153311	**0.167523**
		↑ FSat_{our}	0.017895	0.066727	0.111920	0.161965	0.144677	0.150743	**0.164998**
		↓ Gini_{ori}	0.993229	0.970583	0.942655	0.893080	0.919816	0.908995	**0.888977**
		↓ Gini_{our}	0.996201	0.973246	0.944935	0.894681	0.921783	0.910813	**0.890521**
		↓ $\text{Gini-w}_{\text{ori}}$	0.994377	0.973088	0.948836	0.902403	0.927651	0.918188	**0.894798**
		↓ $\text{Gini-w}_{\text{our}}$	0.996731	0.975391	0.951082	0.904539	0.929846	0.920361	**0.896916**
		↓ VoCD_{ori}	0.789888	0.733042	0.737530	0.705762	0.724919	0.712774	**0.699980**
		↓ II-D_{ori}	**0.000828**	**0.000828**	**0.000828**	**0.000828**	**0.000828**	**0.000828**	**0.000828**
		↓ AI-D_{ori}	0.000381	0.000096	0.000049	0.000032	0.000038	0.000038	**0.000030**

*The scores of our FAIR measures for Pop are not 0 or 1, because in our experiment set-up, items from users' train or validation splits are excluded from the top k recommendation list.

- The recommender models mostly have a similar ordering of the scores of fairness measures: if a fairness measure has a higher value than another in a recommender model X, it is the same for recommender model Y.
- Some models achieve relatively similar relevance scores, but with a huge disparity between their fairness scores.

2.5.2.2 Discussion of fairness evaluation measures in Tab. 2.7

For the higher-is-better measures, Jain, QF, FSat, and Gini, the scores of the original measures and our measures are similar. Both ↑Jain and ↑QF should range in $[0, 1]$, but ↑Jain is very close to 0 i.e., (∼0.1 or less), and ↑QF scores are ∼0.7 (in Lastfm) and ∼0.5 (in Ml-1m). Similarly, the scores of ↓II-D and ↓AI-D are also very close to 0 while ↓Gini scores are closer to 1. While these are due to the different underlying fairness ideas between the measures, the big differences in scores may cause confusion, e.g. that a recommendation is very unfair based on ↑Jain or ↓Gini, or moderately fair based on ↑QF. For Lastfm and Ml-1m, we also see that the absolute scores for the same recommender, e.g., MultiVAE, follow the same order from the lowest to the highest: ↑Jain, ↑FSat, ↑QF, and ↑Ent. This indicates that ↑Jain tends to give lower scores (more unfair) than the other measures. We observe similar trends for ↓II-D and ↓AI-D, which tend to give lower scores (more fair) compared to other lower-is-better measures. The weighted ↓Gini-w is also more strict than the unweighted ↓Gini as ↓Gini-w tends to give more unfair scores than ↓Gini. We study further the strictness of these measures in §2.5.6.

We also observe that scores of ↓AI-D are hardly distinguishable. They differ only in the fourth or more decimal point for both Lastfm and Ml-1m. However, differences in other measures can be seen in the first or second decimal point. The small scores of AI-D may be due to the measure quantifying the disparity between item exposure and random exposure, which is very little when we have a large number of items. This finding suggests that when computing ↓AI-D, care should be taken to avoid rounding errors and failure to distinguish the scores due to the floating-point format.

For all datasets and models, the original Ent cannot be calculated because it returns NaN due to zero division errors. This happens because there are items in the dataset that are not recommended. Our corrected version of this measure (§2.4) does not suffer from this problem.

For the same dataset and k, regardless of the recommender, the II-D scores are always the same/constant. Due to the fixed amount of slots km, within a single recommendation round, the exposure values $E_{u,i} \in \{1, \gamma, \ldots, \gamma^k, 0\}$ (see Eq. 2.8) for all user-item pairs, and the number of user-item pair having a specific exposure

value $E_{u,i}$ is always m. Both of these properties lead to constant II-D scores, as II-D is calculated by taking a mean squared difference between each $E_{u,i}$ value and a constant value based on random expected exposure. When considering cases with multiple rounds of recommendations, the score of II-D may not remain constant anymore, as the user-item exposure values are aggregated across recommendation rounds, resulting in the possibility of $E_{u,i}$ having linear combinations of values from the set above. We have also illustrated in §2.3.1.4 that II-D is not constant in multiple recommendation rounds.

Overall, we observe that:

- The different fairness measures have different ranges in these experiments, even if theoretically they have the same range.
- The original Ent is always incomputable in the experiments, and our corrected Ent resolves the issue.
- II-D_{ori} remains constant for the same dataset, rendering this measure notably less meaningful under this single-round experimental set-up.
- Both ↓II-D_{ori} and ↓AI-D_{ori} have minuscule values, indicating near-perfect fairness even if this contradicts other fairness scores.

2.5.3 Correlation between Measures

When comparing different recommender models, sometimes the ranking of the models (e.g., from the most to least fair score) is more concerning than the absolute values of the measures that we have seen in Tab. 2.7. Motivated by this, we analyse the measures' correlation in order to study the agreement of model rankings based on different measures of relevance and fairness. We compare the following things: 1) the agreement between measures of the same type (relevance or fairness); 2) the agreement between measures of different types; 3) the agreement between the original measures and our corrections to the original measures; 4) the agreement between measures across different datasets. By performing this analysis, we also gain insights into how measures that capture different fairness concepts (dis)agree with one another.

We use Kendall's τ between measures to compute ranking agreement. Fig. 2.1–2.2 show the Kendall's τ values between relevance measures and fairness measures for Lastfm and Ml-1m (see App. 2.9.2.3 for the other datasets). The computation is as follows: for each dataset, we rank the models based on the most relevant or most fair scores. We omit Ent_{ori} as it produces NaN in Tab. 2.7 and we also omit II-D as

the scores for one dataset are the same across models. We compute the correlation significance and correct errors arising from multiple testings for a dataset, using the Benjamini-Hochberg (BH) procedure that is based on false discovery rate [17]. Upon correction, some correlations are still significant; these are indicated by an asterisk (*) in Fig. 2.1–2.2.[15]

We first analyse the correlation among measures of the same type. The relevance measures are highly correlated with each other: $[0.81, 1]$ for Lastfm and $[0.52, 1]$ for Ml-1m, as expected [230]. The fairness measures are also strongly correlated with each other: $[0.71, 1]$ for Lastfm and $[0.81, 1]$ for Ml-1m, except VoCD_{ori} for Lastfm, $[0.43, 0.71]$. This is expected from how the measures treat items when computing fairness; VoCD_{ori} only considers items in the recommendation list, while the remaining fairness measures consider all items in the dataset. For Ml-1m, all the computed correlations between fairness measures are significant after applying the BH procedure. On the other hand, after applying the same procedure for Lastfm, neither of QF nor VoCD, has significant correlations with the rest of the fairness measures, except for QF and Gini/Gini-w. It is also reasonable for QF to not correlate significantly with most of the measures, as it is the only measure insensitive to the difference in the number of times an item is recommended.

Interestingly, even though in Tab. 2.7 the scores of ↑Jain, ↑QF, ↑Ent, and ↑FSat occupy different parts of their range, these measures are highly correlated. The same goes for ↓Gini and ↓AI-D. This shows that even measures based on different concepts of fairness are still capable of producing similar rankings of models. Nevertheless, the absolute scores of the measures can be misinterpreted due to the measures occupying different parts of their range.

Our corrected fairness measures are always perfectly correlated with the original fairness measures (1 in both datasets). This is expected because our corrected versions are obtained by normalization, which does not change the relative order of the models.

Regarding the correlations between measures of different types, we see different trends between relevance and fairness measures for Lastfm and Ml-1m. In Lastfm, we see moderate correlations between fairness and relevance measures, $[0.33, 0.62]$, but these are lower for Ml-1m $[0.14, 0.52]$. These findings are expected as the fairness measures do not consider relevance. None of these correlations are significant after applying the BH procedure. Yet, some correlations between fairness and relevance measures are significant for Book-x and Amazon-is (App. 2.9.2.3).

[15] We also use two more conservative procedures separately to correct the errors: Bonferroni and Holm [99]. Upon correction we obtain no significant results for any tests across the six datasets.

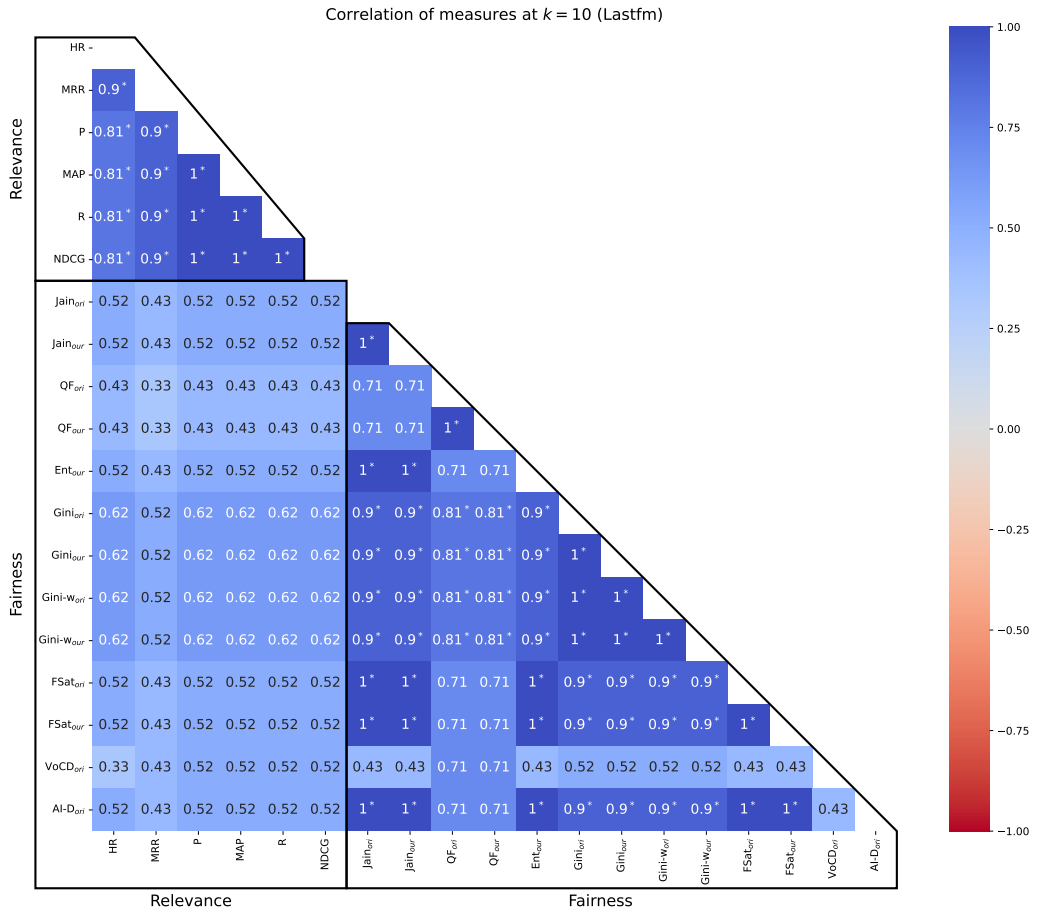


Figure 2.1: Correlation (Kendall's τ) between relevance and fairness measures for Lastfm. Asterisk (*) denotes a statistically significant correlation ($\alpha = 0.05$), after applying the Benjamini-Hochberg procedure.

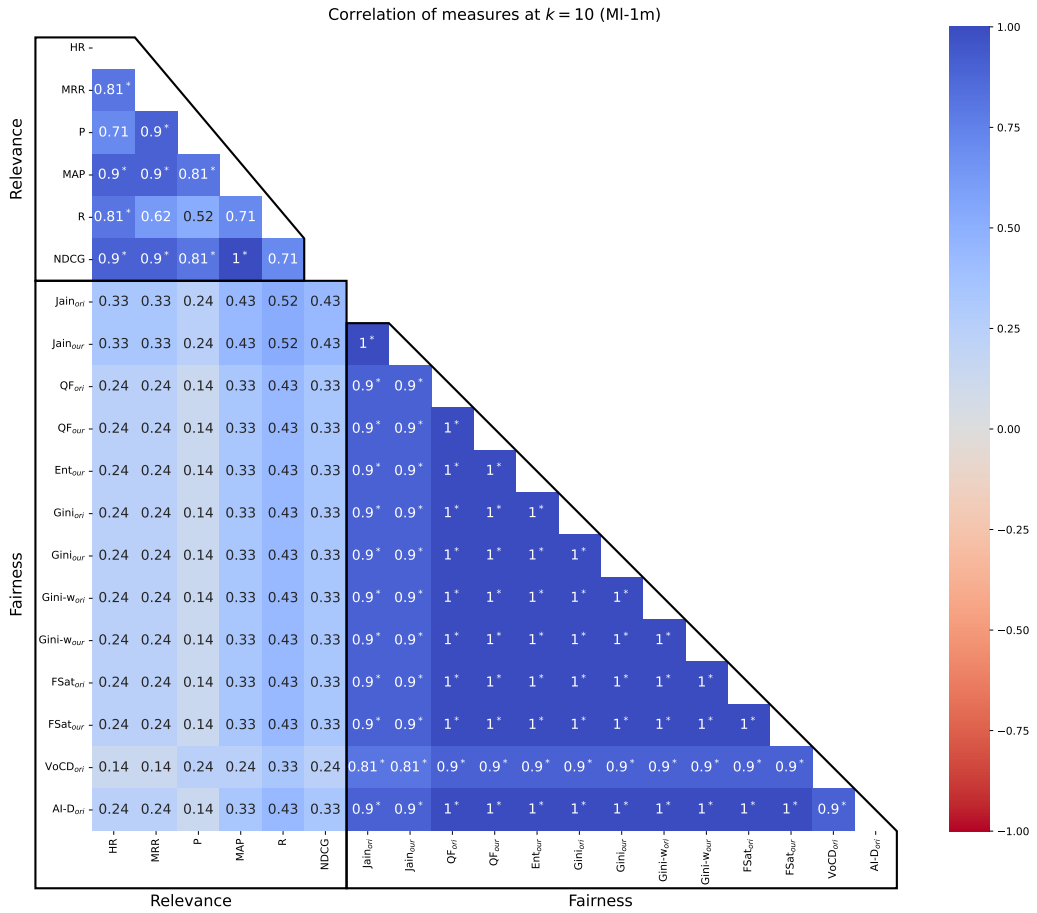


Figure 2.2: Correlation (Kendall's τ) between relevance and fairness measures for Ml-1m. Asterisk (*) denotes a statistically significant correlation ($\alpha = 0.05$), after applying the Benjamini-Hochberg procedure.

2.5.4 Max/min Achievable Fairness

The aim of this experiment is to quantify the extent to which the fairness measures can achieve their theoretical maximum and minimum fairness value (0 or 1) for different datasets and different $k \in \{1, 2, 3, 5, 10, 15, 20\}$. This relates to the **non-realisability** limitation (*Causes 1–3*). We experiment solely with the fairness measures for which we have resolved this limitation, namely Jain, QF, Ent, Gini, Gini-w, and FSat. We primarily compare the original (uncorrected) versions of these measures against the corrected ones. We use two settings: repeatable recommendation, where items in the train/val split can be re-recommended to users following practical cases in industry settings; and nonrepeatable recommendation, which is the typical setting for evaluating recommender systems in academic work. For each setting, we devise two recommenders: MostFair and MostUnfair. Repeatable MostFair aims to recommend each item in the dataset the same amount of times. However, this is impossible if $n \nmid km$ and in this case some items are recommended $\lfloor \frac{km}{n} \rfloor$ times while others $\lfloor \frac{km}{n} \rfloor + 1$ times. For Nonrepeatable MostFair, for each user we generate a list of *recommendable items*, defined as items in I that have not appeared in their corresponding train/val split. For one user at a time, we then recommend the least popular k recommendable items based on the current recommendation lists of all users. Repeatable MostUnfair recommends the same k items to each user. Nonrepeatable MostUnfair does the same, but if any of those k items is a non-recommendable item, the non-recommendable item is replaced by a recommendable item. The results of this experiment for Lastfm and Ml-1m are presented in Fig. 2.3–2.6 and for the remaining datasets in App. 2.9.2.4. We discuss the findings below.

Theoretical maximum fairness. For both nonrepeatable and repeatable settings, all original measures fail to achieve their theoretical maximum fairness values due to the **non-realisability** limitation (*Causes 2–3*). The scores of the original measures get closer to the theoretical maximum fairness values as k increases. However, these scores are still not equal to the theoretical maximum fairness value. In the original measures, having more slots due to larger k does not guarantee that the scores would be higher as well. E.g., the score of $\uparrow\text{Jain}_{\text{ori}}$ in Fig. 2.3 is higher at $k = 3$ compared to $k = 5$, because of the changing values of $km \bmod n$ for different values of k. Our corrected versions always reach their theoretical maximum fairness values for both repeatability settings, except for Gini-w_{our}. This behaviour is due to the unresolvable non-realisability (*Cause 4)* limitation for Gini-w. However, Gini-w_{our} can still reach the theoretical most fair value when $k = 1$ for Lastfm (Fig. 2.4), while the original version fails to do so.

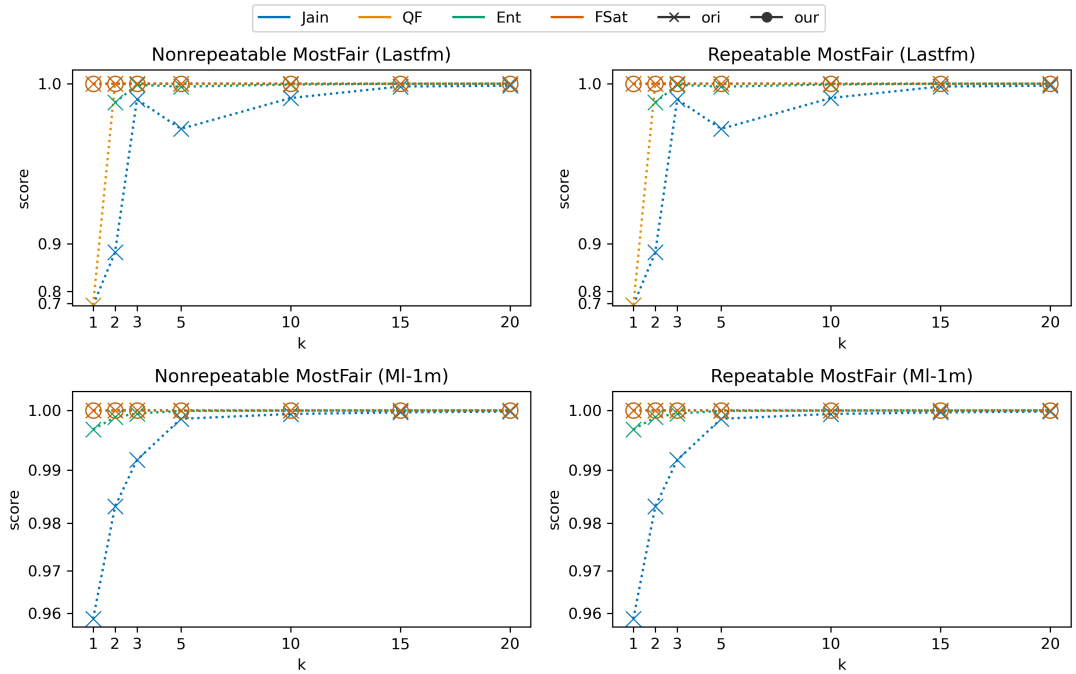


Figure 2.3: Most fair scores with varying k for higher-is-fairer fairness measures for Lastfm and Ml-1m. All scores from the corrected measures (denoted by 'our') measures overlap with each other.

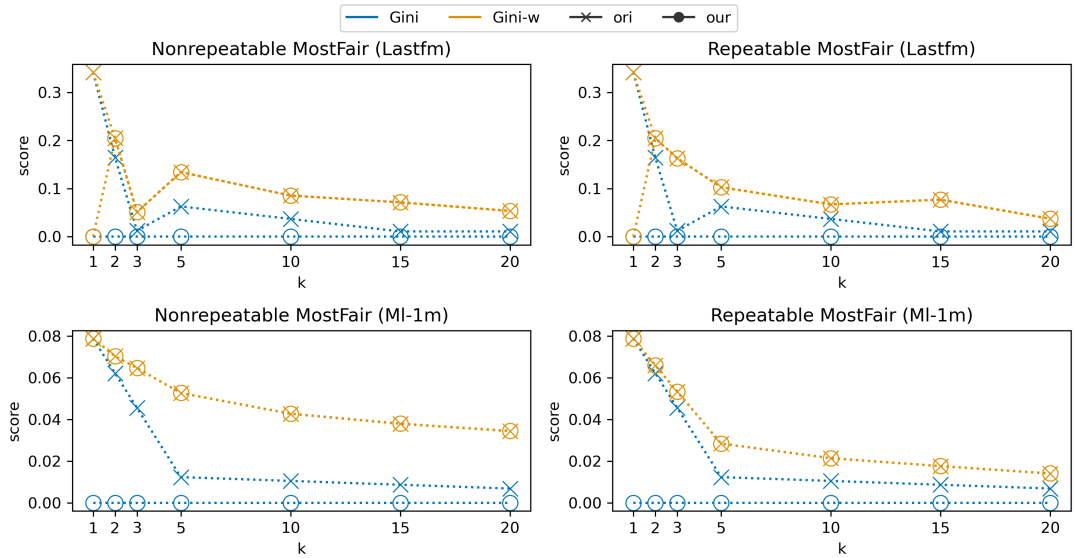


Figure 2.4: Most fair scores with varying k for lower-is-fairer fairness measures for Lastfm and Ml-1m.

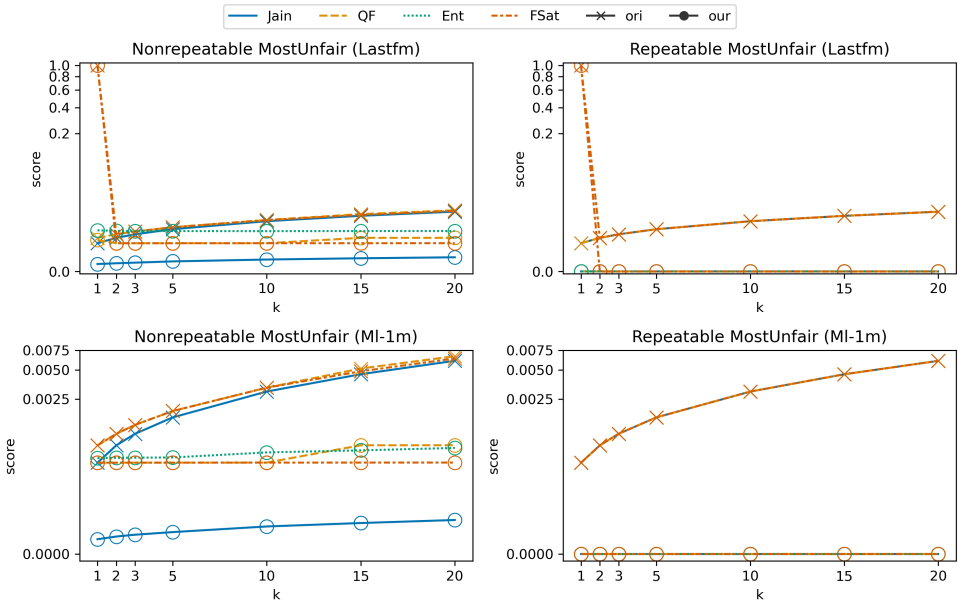


Figure 2.5: Most unfair scores with varying k for higher-is-fairer fairness measures for Lastfm and Ml-1m. On Repeatable MostUnfair, all scores from the corrected measures (denoted by 'our') overlap with each other for the shown values of $k > 1$ for Lastfm and for all shown values of k for Ml-1m.

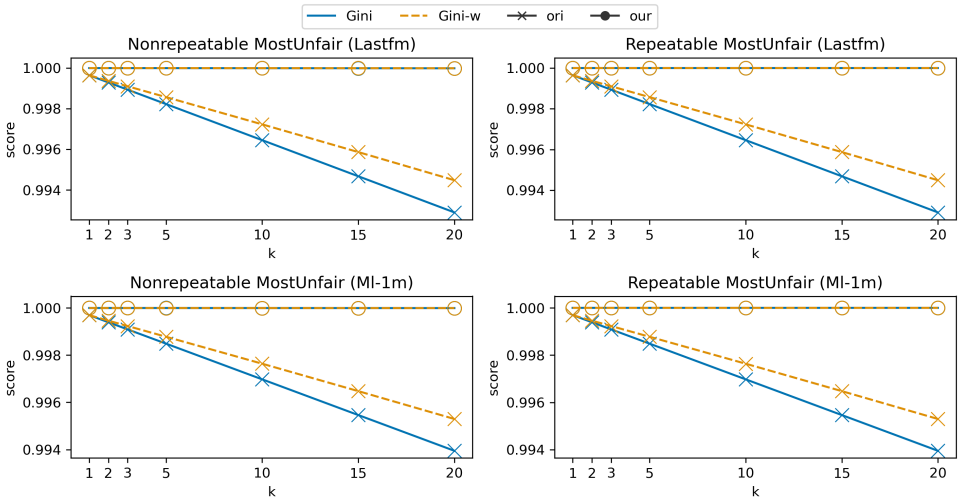


Figure 2.6: Most unfair scores with varying k for lower-is-fairer fairness measures for Lastfm and Ml-1m. On Repeatable MostUnfair, all scores from the corrected measures (denoted by 'our') overlap with each other for all shown values of k.

Theoretical minimum fairness. All original measures fail to reach the theoretical minimum values for all experimented values of k for all settings due to **non-realisability** limitation (*Cause 1*). This happens less frequently in our measures (Fig. 2.5). Our measures successfully achieve the theoretical minimum fair values under the repeatable settings, except for FSat in Lastfm (Fig. 2.5). This is because when $k = 1$, there are not enough slots for the items and due to the **always-fair** limitation that is unresolvable (§2.4.3), the score for ↑FSat is 1, which is not the theoretical minimum fair value. Additionally, the scores of the original measures diverge from the theoretical minimum fairness value with larger k. This happens to our measures only in the nonrepeatable setting because the normalization is done by assuming that any item can be recommended to any users. This assumption is not true in the nonrepeatable settings because some items cannot be re-recommended to some users. However, differently from the original measures, the scores of our measures diverge less as k increases.

Overall, while the difference in scores between the original measures and our versions is not large, our measures quantify the actual most (un)fair situations more accurately than the original measures. The difference between the original measures and our versions in the most unfair recommendation under the repeatable setting is $\frac{k}{n} - 0 = \frac{k}{n}$ for Jain, QF, Gini, and FSat; and $\log k$ for Ent (see Tab. 2.5). However, the difference would be greater in item-poor domains where n is small and therefore possibly close to k, e.g. insurance [22]. The scores of the original measures also change with k. This makes their interpretation harder because the distance between the original scores and the theoretical maximum/minimum fair score also changes without an intuitive pattern for different values of k, as seen in the $\uparrow\text{Jain}_{\text{ori}}$ scores in Fig. 2.3 which can increase or decrease as k increases. Furthermore, the original measures suffer particularly for low k values, which are the most important rank positions in real-life RSs. The scores of our measures rarely change with different values of k.

2.5.5 Sliding Window: Relevance and Fairness at Different Rank Positions

This experiment studies how relevance and fairness scores of all measures vary at decreasing rank positions. The experiment aims to observe 1) the change in relevance scores, if any, as items should ideally be placed in the ranks according to decreasing order of true relevance; and 2) whether and how the fairness scores change across different rank positions. Due to bias in recommenders, popular items tend to be given more exposure. Thus, we expect the relevance scores to decrease and the fairness

scores to become more fair at decreasing rank positions. We study how the above changes may generally differ between relevance measures and fairness measures, as well as between different fairness measures, including the ones with different fairness notions.

We conduct this experiment as follows. We use the runs from the BPR model, which is the best in our experiments. Given one run, we compute the measures for different sliding windows of rank positions in rankings 1–5, 2–6, and so on until 5–9. We reorder the recommended items such that items that were previously recommended at the top positions are now at the bottom positions when we change the window according to decreasing rank. The results for Lastfm and Ml-1m are presented in Fig. 2.7 and for the rest of the datasets in App. 2.9.2.5.

The following observations from Fig. 2.7 apply to both the original fairness measures and our corrected versions of these measures unless otherwise stated. All relevance scores decrease as rank decreases. The drop of relevance scores for Ml-1m ($[0.04, 0.23] \rightarrow [0.03, 0.20]$) is less extreme than in Lastfm ($[0.12, 0.48] \rightarrow [0.04, 0.25]$). This is partly because the test set of Lastfm has at most five relevant items per user, while on average, Ml-1m has many more. While relevance scores decrease, fairness measures show that fairness slightly increases down the rank, except for $\downarrow$VoCD$_{\text{ori}}$. The range of higher-is-better fairness measures increases from $[0.06, 0.65] \rightarrow [0.10, 0.71]$ for Lastfm and $[0.05, 0.65] \rightarrow [0.08, 0.71]$ for Ml-1m. The range of $\downarrow$Gini and $\downarrow$Gini-w, decreases from $[0.91, 0.92] \rightarrow [0.87, 0.88]$ for Lastfm and $[0.91, 0.92] \rightarrow [0.88, 0.89]$ for Ml-1m. $\downarrow$VoCD$_{\text{ori}}$ seems invariant to changes in the position window ($0.61 \rightarrow 0.60$ for Lastfm and $0.68 \rightarrow 0.67$ for Ml-1m). This may be because VoCD$_{\text{ori}}$ is the only measure that considers fairness exclusively for recommended items, and the recommended items differ a little in terms of the number of times they are recommended as rank decreases. $\downarrow$AI-D has even smaller changes in scores as the values are already minuscule in the first place, while $\downarrow$II-D is always constantly small for a dataset. The small values, compared to other measures, are due to these measures quantifying fairness using different concepts from other measures, i.e. comparing exposure to random exposure (also observed and explained in §2.5.2). The ranges of all fairness measures are roughly the same across datasets, but the range of relevance measures varies across datasets. This also holds for the datasets in App. 2.9.2.5. This may be due to the distribution of the recommended items being similar across datasets, and the distribution of the number of relevant items differing across datasets, as explained above for Lastfm and Ml-1m.

Fairness measures are also somewhat invariant to changes in relevance. This is anticipated as the equations of fairness measures are independent of relevance values.

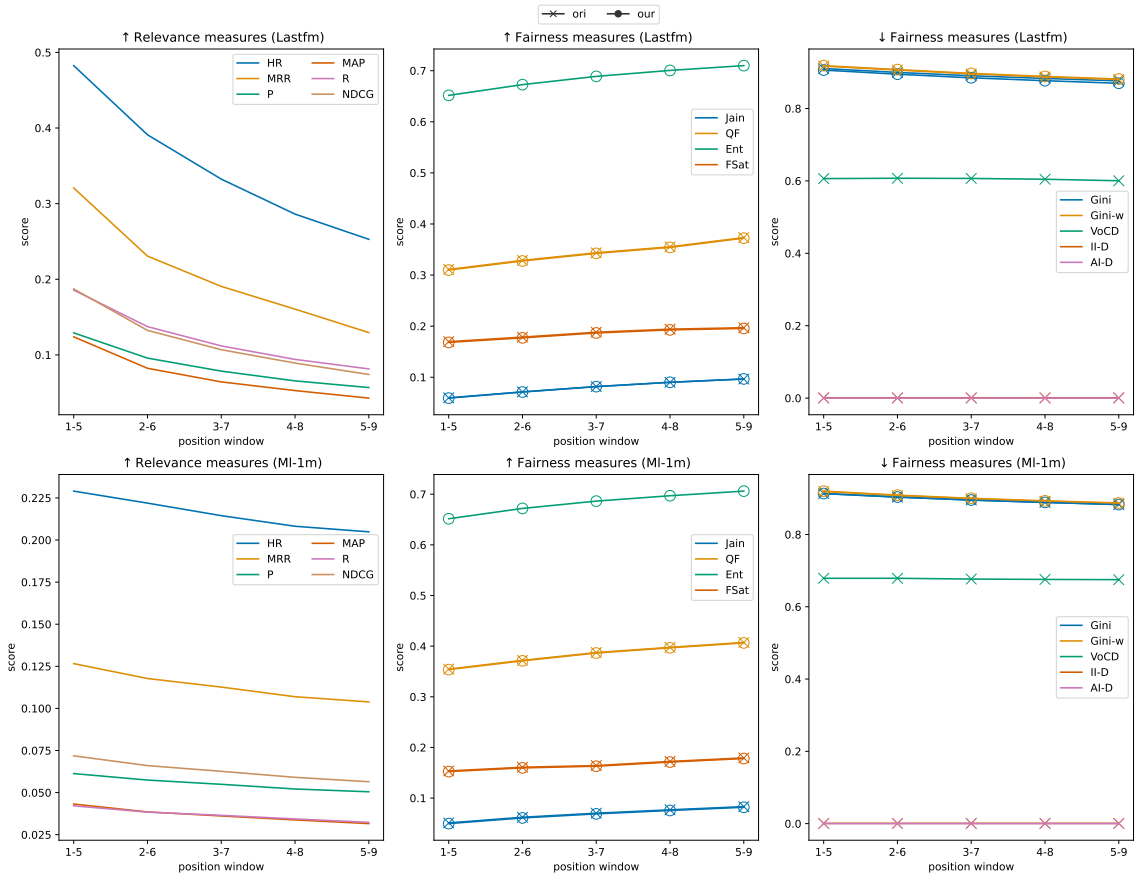


Figure 2.7: Sliding window evaluation for BPR model, on Lastfm and Ml-1m. Each row of figures is for one dataset, each column is for the different groups of measures (relevance, higher-is-better fairness, lower-is-better fairness measures). II-D and AI-D lines overlap.

2.5.6 Measure Strictness and Sensitivity through Artificial Insertion of Items

We have observed in §2.5.2 that different fairness measures vary in their strictness of quantifying fairness (e.g., some measures give scores close to the most fair values, and the opposite for others). It is however unknown how sensitive fairness measures are, given the change of the number of times an item is exposed in the recommendation list across all users. Therefore, the goal of this experiment is to study the strictness and sensitivity of the measures, and compare these aspects between measures of similar and different fairness concepts. Knowing the strictness and sensitivity of the measures matters as this affects how we interpret the scores of the measures. For example, if one uses a measure that tends to produce scores close to the most fair value, they must be aware that the score may not reflect fairness accurately.

As such, we devise an experiment to specifically study how the relevance measures, existing fairness measures and our corrected fairness measures scores change when we artificially control the fraction of jointly least exposed and relevant items in the recommendation list. We start with an initial recommendation list. We define a *least exposed (LE) item* as an item in the dataset with the least exposure, based on the current recommendation list.[16] An LE item in this experiment is therefore an item that has not appeared in the current recommendation list. We define a *relevant item* as per the labels of relevance.

From the initial recommendation list, we insert jointly LE and relevant items, one item at a time. We create a synthetic dataset with $m = 1000$ users and $n = 10000$ items. The number of items is exactly the number of recommendation slots km for a cut-off $k = 10$. We artificially generate a ranking of top k as follows. The artificial insertion of jointly LE and relevant items begins with the recommendation of the same 10 items $i_1, i_2, \ldots, i_{10}$ to all users. These items are irrelevant to each user except u_1, as we keep the recommendation list for u_1 the same throughout the experiment. This is because we want to keep the number of items exactly km where theoretically each item could be recommended exactly once and if we have to completely replace all m users' recommendation lists, we would need to have more than km items. We expect the relevance measures to give scores close to zero on this initial recommendation list as only u_1 has relevant items. We expect the fairness measures to give scores that are equal to or close to the theoretical most unfair scores.[17]

[16] This fairness concept is closely tied to all measures in this work, except for VoCD which concerns only items in the recommendation list, as opposed to in the dataset (Tab. 2.3).

[17] We say "close to" due to the **non-realisability** (*Cause 4*) limitation in some measures.

Let P be the fraction of items in the k that are artificially inserted by us. We vary from $P = 0$, the original recommendation where we have not inserted any items artificially, to $P = 1$ where all items in the k are jointly LE and relevant items that are artificially inserted by us. We increase P in steps of $1/k$. From the bottom of a user's recommendation list, we replace one item at a time with a known jointly LE and relevant item, until we end with a recommendation list of different km items across all users, that are all relevant only to the user to whom that item is recommended. At the end of the insertion process, each user is recommended exactly 10 relevant items, and those items are also fair w.r.t. the entire recommendation list for all users, considering all items in the dataset; item fairness is not defined w.r.t. a specific user. We expect the relevance measures to give scores of 1 on the final recommendation list and fairness measures to give scores that are (close to) the fairest scores.[17]

The results of this experiment are presented in Fig. 2.8. We see that all relevance measures increase as we add more relevant items.[18] The following observations apply to both the original fairness measures and to our corrected versions of these measures, unless otherwise specified. All fairness measures, except VoCD_{ori} and II-D_{ori}, indicate more fairness as we increase P, but with varying sensitivity, explained next. ↑Jain is one of the strictest fairness measures. Even when the proportion of LE items is 0.9 (item i_1 is recommended to all users, but the rest of the recommendation lists are filled with different items), the ↑Jain score is still close to 0, which translates to unfair while ↑QF, ↑Ent, and ↑FSat are 0.9, which is close to the fairest score of 1. The scores of QF_{ori} are exactly the same as FSat_{ori}, because all items in the recommendation list are recommended once, which is also the maximin share (defined in §2.2.2). $\uparrow\text{QF}_{\text{our}}$, $\uparrow\text{Ent}_{\text{our}}$, and $\uparrow\text{FSat}_{\text{our}}$ also give identical scores. This is expected as the increase in the scores is constant and proportional to the fraction of artificially inserted LE items, yet this is interesting as these three measures are based on three different fairness notions (QF being insensitive to the number of times an item is recommended, and FSat being based on maximin-shared fairness).

Meanwhile, the increase of fairness in ↓Gini and ↓Gini-w follows a non-linear trend, with ↓Gini-w being stricter than ↓Gini. The non-linear trend is also expected as Gini and Gini-w are based on the Lorenz curve, a graphical representation of the cumulative proportion of exposure to the cumulative proportion of items. We also see that $\text{Gini-w}_{\text{our}}$ is able to reach the theoretical most fair when the entire recommendation list consists of artificially inserted LE items, while $\text{Gini-w}_{\text{ori}}$ fails. $\downarrow\text{VoCD}_{\text{ori}}$ is insensitive to the insertion of items as it only considers fairness for recommended items. The number of times these items are recommended across all

[18]The relevance measures do not start from 0 when $P = 0$, as there is one user with ten relevant items.

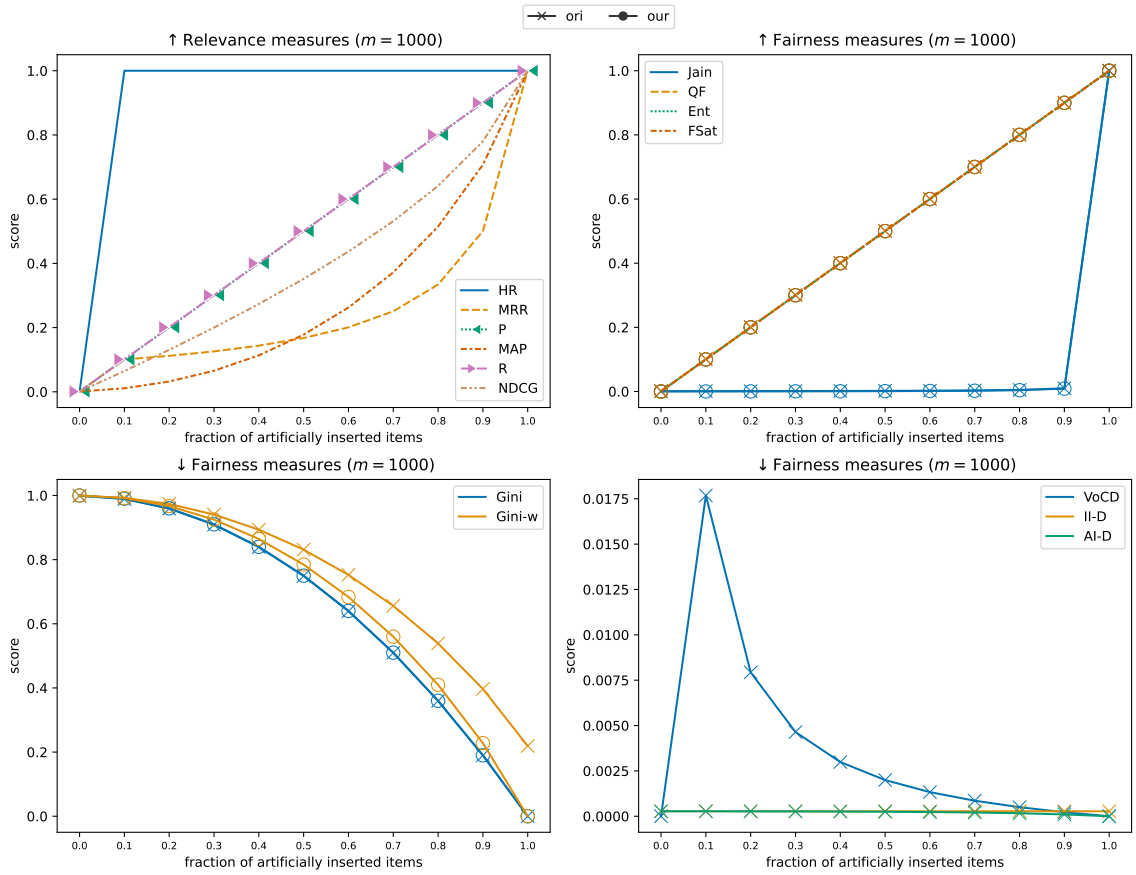


Figure 2.8: Results for jointly LE and relevant item insertion. All measures are at $k = 10$. $\mathrm{QF}_{\mathrm{ori}}$ and $\mathrm{FSat}_{\mathrm{ori}}$ overlap. $\mathrm{QF}_{\mathrm{our}}$, $\mathrm{FSat}_{\mathrm{our}}$, and $\mathrm{Ent}_{\mathrm{our}}$ also overlap.

users does not differ much in this set-up, therefore ↓VoCD_{ori} returns scores that are close to the fairest. Most notably, ↓II-D_{ori} and ↓AI-D_{ori} are very close to 0 (on the scale of 10^{-3} or even smaller) even when the same k items are recommended to all users. ↓II-D_{ori} remains constant, while ↓AI-D_{ori} is rather insensitive to the addition of LE and relevant items. The small scores are due to the measures quantifying fairness according to the closeness of item exposure with random exposure, while other measures have no such comparisons. Therefore, for these measures, the scales are not very meaningful, even though for AI-D, the scores still indicate improvement as we insert more LE items.

We see similar trends with $m \in \{100, 500\}$, but the change of the scores is most stable with $m = 1000$. As we increase the number of users (and items), the range of VoCD, II-D, and AI-D scores also becomes more compressed. In contrast, the range of the other measures remains similar. We also observe a similar but opposite trend of results when we artificially insert known irrelevant and multiple copies of items already in the recommendation list. Both of these results are in App. 2.9.2.6.

Overall, the artificial insertion experiment indicates that several measures respond linearly to the insertion of LE items i.e., ↑QF, ↑FSat, and ↑Ent, while the rest do so non-linearly. This can affect the interpretation of these scores, as we observe that it is generally harder to achieve a high fairness score in some measures. In some other measures, it is also easier to improve fairness when starting from a relatively fair situation, but much harder when starting from a completely unfair situation.

2.6 Related Work

Prior work [60, 147, 148, 167, 227, 234, 264] proposes exposure-based individual item fairness measures but does not provide a comprehensive analysis of the limitations of the measures. Our work differs from this because we extensively analyse individual item fairness measures specifically for RSs, identify novel and previously-known limitations in them, and address these limitations. Meanwhile, several other work uses individual item fairness measures that also takes into account the relevance of the item to users [23, 155, 191, 234, 268]. The investigation of these measures (of fairness and relevance) is reserved for our future work.

Amigó et al. [9] overview fairness measures in RSs and characterise them according to five dimensions. They focus on generalising the measures into broad categories and studying the relationship between multi-stakeholder fairness, whereas we analyse each individual item fairness measure both theoretically and empirically. We also focus on the relationship among individual item fairness measures.

Raj and Ekstrand [175] analyse fairness measures for provider-side group fairness in ranked outputs. They include one measure for individual item fairness, II-D (referred to as EED in [175], which was originally proposed by [56]), but only analyse it as a group fairness measure. They list some questions to assess the design of fairness measures, which we exploit in this work, e.g., the examination function used (Tab. 2.2), the ideal/fair criteria (Tab. 2.3), and if the measure incorporates relevance (§2.2.2). Both our work and theirs identify the limitations in the measures e.g., edge cases where zero values cause undefinedness in the measure computation (§2.3.3). They propose to use a small constant to avoid computing $\log 0$ in group fairness metrics, but we do not use the same approach as using a small constant can introduce noise in the measure computation.

Majumder et al. [144] examine classification measures for both individual and group fairness through empirical analysis and cluster the measures based on correlation values. They analyse the correlation among fairness measures but do not investigate the limitations of those measures. They find disagreements between the measures when labelling a model as fair or unfair. We did not do this mapping, as this may lead to loss of valuable information regarding the range and the actual values of the measures. However, we show that disagreements also exist between several individual item fairness measures in RSs.

All previous work finds that several fairness measures are highly correlated, and are insensitive to changes in data [9, 144, 175]. Our analysis in §2.5 confirms these findings in a different experimental set-up and sheds light onto additional limitations which have not been reported or corrected previously. Next, we provide practical guidelines for choosing among individual item fairness measures.

2.7 Discussion

2.7.1 Summary of Theoretical Corrections

In this paper, we critically analyse individual item fairness measures in RSs w.r.t. their limitations. We point out a total of five theoretical limitations in the measures, and identify that each measure suffers from two or more limitations. Some limitations are due to intentional design choices of the measure, while the remaining limitations go against some pre-defined desirable properties of (fairness) evaluation measures.

We posit that evaluation measures of fairness should have two important utilities: 1) to assess systems/models in isolation (i.e., evaluating 'how fair' a single system is, with one endpoint being the most unfair and the other being the most fair); 2) to compare different RSs and make a decision about whether and to what extent

system A is more fair than system B, based on the measure.[19] A measure should ideally be usable for both use cases. At the present stage, none of the individual item fairness measures is suitable for the first use case, hence the need to modify the current measures for it.[20]

Note that having limitation(s) does not mean that the measures are completely unusable. Some measures can still be used despite having limitations, as long as one is aware of these limitations. For instance, in the case of measures that empirically cannot reach the endpoints of $[0, 1]$, it is still possible to use the measure to compare fairness between two or more systems. Yet, evaluating fairness for a single system using such a measure is a challenge, as having a score of, for example, 0.6 does not always mean that there is still 40% room for improvement. This point should be kept in mind, especially given the common practice of interpreting a single score in comparison to the known range of that measure.

We also provide theoretical solutions to address the three resolvable limitations, and we argue why the remaining limitations cannot be resolved. The first set of solutions guarantees that the measures range in a bounded interval, e.g., $[0, 1]$, where both the theoretical minimum and maximum scores are achievable, with one endpoint corresponding to the most unfair recommendation list, and the other to the fairest recommendation list. This set of solutions considers the number of recommendation slots and the number of items in the dataset, and at the same time assures that the measures are well-defined for both common and edge cases. Our second solution ensures that the affected measure is sensitive to the change of exposure received by an item, thereby fulfilling a desired property of fairness measure.

2.7.2 Summary of Empirical Findings

Extensive empirical experiments were conducted to compute relevance and fairness scores for both the original measures and for our corrected versions of these measures. The experiments utilised six datasets and seven recommendation models, including state-of-the-art models and well-established baseline models. Even though the models are all trained to optimise for relevance, we discover that the fairest model is not necessarily the worst in terms of relevance scores. This was unexpected as the fairness measures used in this work are detached from relevance. However, we also

[19]Note that utility 1) is for assessment purposes, not a goal for model development; it may not be possible for relevant recommendations to be maximally fair at the same time.

[20]This is unlike relevance measures, where there is still room for choice, as several measures have reachable endpoints (e.g., NDCG, RR@k) [153].

see the common observation where several models have higher relevance scores, but exhibit lower fairness.

Our results empirically show that relevance measures and fairness measures have different ranges, which makes the interpretation of the fairness measures difficult. Further, the range of fairness measures is incomparable between different measures, and for some measures, this range is also not lower/upper-bounded empirically. Other noticeable observations include some fairness measures that tend to score much lower/higher compared to other measures, as well as uncorrected measures that are incomputable or produce constant values, given any recommendation list based on the same dataset. While the actual scores of the measures may differ, we found that most fairness measures have a high and significant agreement in ranking the recommenders from the most to least fair. The strong agreement is observed between the corrected and uncorrected measures, and even between some measures that are based on different fairness concepts. Altogether, our results show that:

- Despite limitations in quantifying the extent of fairness, the measures agreed in the ordering of models according to their individual item fairness.
- Some original measures do not reflect the absolute quantity or differences in fairness.
- The corrected measures are required to reliably use the measures for scenarios that may contain commonly occurring and edge cases.

2.7.3 Guidelines of the Appropriate Use of the Fairness Measures

Next, we summarise guidelines on using these fairness measures based on the above theoretical and empirical findings.

Use original fairness measures only to evaluate relative fairness. *Relative fairness* refers to comparing the relative ordering of fairness scores. The original fairness measures suffer from theoretical limitations that limit their usage in settings where data distributions or recommendation scenarios do not fulfil the theoretical premises of the original measures. Moreover, the original measures may be more difficult to interpret as their range and scaling do not always match the intuitive expectations of being between 0 and 1. While the original measures as proposed outside recommendation can be used as their ranges are known and can be easily interpreted, we advise using them to evaluate only relative fairness.

Use our corrected fairness measures to evaluate absolute fairness. *Absolute fairness* refers to measuring how close a model's recommendation is to the most (un)fair recommendation scenario. To evaluate absolute fairness, we recommend using our corrected measures for Jain, QF, Ent, Gini, and FSat. Our fairness measures are always perfectly correlated with the original measures, thus providing results that align with the original measures. Further, our measures are well-defined and have better interpretability w.r.t. how the minimum/maximum scores correspond to the most unfair/fair recommendation scenario, as shown in §2.5.4. Our fairness measures are highly correlated with each other, but because they operate on different scales, one should not deduce that a model is (un)fair based on the absolute fairness measurement scores.

Note that both FSat_{ori} and FSat_{our} should never be used when $km < n$ due to the unresolvable **always-fair** limitation, as the score will always be perfectly fair regardless of the recommendation. $\downarrow$Gini-w_{our} should preferably be used when one has equal or more items than recommendation slots ($km \leq n$), as the measure works ideally in that setting: a score of 0 means the recommendation is perfectly fair, while a score of 1 means the unfairest possible recommendation. For the remaining cases, which commonly happens in many public recommendation datasets for any cut-off k (Tab. 2.6), even if the most unfair recommendation entails a score of 1, the most fair recommendation is not mapped to a score of 0 in Gini-w_{our}. Yet, this is still better than Gini-w_{ori} which maps unrealistic scenarios to 0 and 1. For Ent, we recommend using our correction, since our correction avoids the **undefinedness** limitation, and would produce the same score as Ent_{ori} when all items are recommended.

We discourage using the rest of the measures due to their tendency to have scores that are not representative of fairness, e.g. scale mismatch between II-D/AI-D and the rest of the measures. Additionally, II-D should not be used for single-round recommendations as the scores are always constant.

2.8 Conclusions

We have presented a novel investigation into the theoretical and empirical limitations of current evaluation measures of individual item fairness in recommender systems. We have further amended these measures to correct their limitations or have argued why some limitations are impossible to resolve. Extensive experiments on real-life and synthetic data reveal novel insights on how individual item fairness measures should and should not be used.

In the present work, we solely concentrated on measures that quantify individual item fairness independently of recommendation performance. We reserve the analysis of fairness measures that are tied to relevance for our future work. Future work should investigate whether measures that aim to simultaneously quantify both recommendation performance, or relevance, and fairness suffer from similar limitations and empirical behaviours than the measures studied here. Other future work could further explore the relationship between individual item fairness and item group fairness [234], or fairness between users and items [9]. All measures studied here also assume that exposure is the key factor for fairness, while there might be other factors to consider for fairness, e.g. speed of the recommendation or the wait time from when an item is introduced to a system until it gets recommended.

The empirical studies could also be extended to account for the behaviour and performance of the measures using additional datasets. However, while we cannot exclude the possibility that the experiments on other settings, domains, or datasets could lead to new insights, it is unlikely that they would affect our conclusions and guidelines on the appropriate use of the studied measures. Future work could also utilize our corrected measures to optimize recommendation models for fairness to reveal whether the corrected measures could improve recommendation models as opposed to only measuring the performance of existing models.

Code Availability

Our source code (in Python 3.10) is publicly available on `https://github.com/theresiavr/individual-item-fairness-measures-recsys` and usable under the MIT License, with proper attribution to this work and possibly other related work. Other restrictions regarding the usability of the code may apply for the RecBole library [260].

Acknowledgements

The work is supported by the Algorithms, Data, and Democracy project (ADD-project), funded by Villum Foundation and Velux Foundation, as well as the Academy of Finland. We also thank the anonymous reviewers who have provided insightful comments and suggestions to improve earlier versions of the manuscript.

2.9 Appendix

2.9.1 Mathematical Workings for Bounds in Tab. 2.5

We provide the derivation of the obtained min/max achievable value of Jain, QF, Ent, Gini, FSat, and VoCD. The min/max achievable values are obtained from the most (un)fair recommendation scenarios described in §2.4.1.

2.9.1.1 Jain's Index

The most unfair case for ↑Jain produces Jain_{min} and the most fair case for ↑Jain produces Jain_{max}.

$$\text{Jain}_{\text{min}} = \frac{(km)^2}{n \sum_{i \in I} \left[\sum_{u \in U} 1_{R_u^k}(i) \right]^2} = \frac{(km)^2}{n(km^2)} = \frac{k}{n}$$

$$\begin{aligned}
\text{Jain}_{\text{max}} &= \frac{(km)^2}{n \sum_{i \in I} \left[\sum_{u \in U} 1_{R_u^k}(i) \right]^2} \\
&= \frac{(km)^2}{n \left((n - km \bmod n) \left\lfloor \frac{km}{n} \right\rfloor^2 + (km \bmod n) \left(\left\lfloor \frac{km}{n} \right\rfloor + 1 \right)^2 \right)} \\
&= \frac{(km)^2}{n \left(n \left\lfloor \frac{km}{n} \right\rfloor^2 - (km \bmod n) \left\lfloor \frac{km}{n} \right\rfloor^2 + (km \bmod n) \left(\left\lfloor \frac{km}{n} \right\rfloor^2 + 2 \left\lfloor \frac{km}{n} \right\rfloor + 1 \right) \right)} \\
&= \frac{(km)^2}{n \left(n \left\lfloor \frac{km}{n} \right\rfloor^2 + (km \bmod n) \left(2 \left\lfloor \frac{km}{n} \right\rfloor + 1 \right) \right)}
\end{aligned}$$

2.9.1.2 Qualification Fairness

The most unfair case for ↑QF produces QF_{min} and the most fair case for ↑QF produces QF_{max}.

$$\text{QF}_{\text{min}} = \frac{(k \cdot 1)^2}{n(k \cdot 1^2)} = \frac{k}{n}$$

When there are not enough recommendation slots for all items, $km < n$:

$$\mathrm{QF}_{\max} = \frac{(km \cdot 1)^2}{nkm(1^2)} = \frac{km}{n}$$

When $km \geq n$, all items can be recommended, hence $\mathrm{QF}_{\max} = \frac{n}{n} = 1$.

2.9.1.3 Entropy

The most unfair case for ↑Ent produces $\mathrm{Ent}_{\min}$ and the most fair case for ↑Ent produces $\mathrm{Ent}_{\max}$. The first term in $\mathrm{Ent}_{\max}$ comes from $n - km \bmod n$ that are each recommended $\left\lfloor \frac{km}{n} \right\rfloor$ times and the second comes from $km \bmod n$ items that are each recommended $\left\lfloor \frac{km}{n} \right\rfloor + 1$ times.

$$\mathrm{Ent}_{\min} = -k \cdot \frac{1}{k} \log \frac{1}{k} = -\log k^{-1} = \log k$$

$$\begin{aligned}\mathrm{Ent}_{\max} = &-(n - km \bmod n)\left(\frac{\left\lfloor \frac{km}{n} \right\rfloor}{km} \log \frac{\left\lfloor \frac{km}{n} \right\rfloor}{km}\right) \\ &-(km \bmod n)\left(\frac{\left\lfloor \frac{km}{n} \right\rfloor + 1}{km} \log \frac{\left\lfloor \frac{km}{n} \right\rfloor + 1}{km}\right)\end{aligned}$$

2.9.1.4 Gini Index

The most unfair case for ↓Gini produces $\mathrm{Gini}_{\max}$ and the most fair case for ↓Gini produces $\mathrm{Gini}_{\min}$.

$$
\begin{aligned}
\text{Gini}_{\max} =& \frac{\sum_{j=n-k+1}^{n} (2j-n-1)m}{nk(m)} \\
=& \frac{\sum_{n-k+1}^{n} 2j - n \sum_{n-k+1}^{n} 1 - \sum_{n-k+1}^{n} 1}{nk} \\
=& \frac{2\sum_{n-k+1}^{n} j - n(n-(n-k+1)+1) - (n-(n-k+1)+1)}{nk} \\
=& \frac{2\frac{(n-k+1+n)(n-(n-k+1)+1)}{2} - nk - k}{nk} \\
=& \frac{(2n-k+1)(k) - nk - k}{nk} \\
=& \frac{(k)(2n-km+1-n-1)}{nk} \\
=& \frac{n-k}{n} \\
=& 1 - \frac{k}{n}
\end{aligned}
$$

To derive $\text{Gini}_{\min}$, we use the pairwise difference formula of Gini:

$$
\text{Gini} = \frac{\sum_{(i,i')} \left| \sum_{u \in U} 1_{R_u^k}(i) - \sum_{u \in U} 1_{R_u^k}(i') \right|}{2n^2 \bar{x}}
$$

where $\bar{x}$ is the average number of times an item is recommended, for the most fair case, calculated as follows:

$$
\bar{x} = \frac{(n - km \bmod n)\left\lfloor \frac{km}{n} \right\rfloor + (km \bmod n)(\left\lfloor \frac{km}{n} \right\rfloor + 1)}{n} = \frac{n\left\lfloor \frac{km}{n} \right\rfloor + km \bmod n}{n}
$$

We simplify the numerator and denominator separately, for clarity in the proof. To simplify the numerator, in the most fair case, there are $2(n - km \bmod n)(km \bmod n)$ pairs of items with an absolute difference of $\left(\left\lfloor \frac{km}{n} \right\rfloor + 1\right) - \left\lfloor \frac{km}{n} \right\rfloor = 1$, which is the difference of the number of times the items are recommended. The rest of the pairs

have 0 differences. Hence, the numerator is $2(n - km \bmod n)(km \bmod n)$. Putting everything together:

$$\begin{aligned}
\text{Gini}_{\text{min}} &= \frac{2(n - km \bmod n)(km \bmod n)}{2n^2 \left(\frac{n \left\lfloor \frac{km}{n} \right\rfloor + km \bmod n}{n} \right)} \\
&= \frac{(n - km \bmod n)(km \bmod n)}{n \left(n \left\lfloor \frac{km}{n} \right\rfloor + km \bmod n \right)} \\
&= \frac{(n - km \bmod n)(km \bmod n)}{n \left(n \left\lfloor \frac{km}{n} \right\rfloor + km - n \left\lfloor \frac{km}{n} \right\rfloor \right)} \\
&= \frac{(n - km \bmod n)(km \bmod n)}{kmn}
\end{aligned}$$

The most unfair case for ↓Gini-w produces Gini-w_{max} and the most fair case for ↓Gini produces Gini-w_{min}.

To obtain Gini-w_{max}, we derive the numerator and denominator separately using Eq. 2.4. We first compute the numerator of Gini-w_{max}, considering that the items with the least to the most exposure are as follows: the first $n - k$ items are with zero exposure (as they are not present in the top k), one item is exposed m times at position k, one item is exposed m times at position $k-1$, and so on, until one last item that is exposed m times at the top of the recommendation list. The exposure received by those items respectively are $0, m \log_{k+1} 2, m \log_k 2, \ldots, m \log_2 2$. Therefore, the numerator of Gini-w_{max} can be written as $m \sum_{\ell=1}^{k} (n - 2\ell + 1) \log_{\ell+1} 2$. Meanwhile, the denominator of Gini-w_{max} is simply n times the total exposure received by the items: $mn \sum_{\ell=1}^{k} \log_{\ell+1} 2$. Putting the numerator and denominator together:

$$\text{Gini-w}_{\text{max}} = \frac{m \sum_{\ell=1}^{k} (n - 2\ell + 1) \log_{\ell+1} 2}{mn \sum_{\ell=1}^{k} \log_{\ell+1} 2} = \frac{\sum_{\ell=1}^{k} (n - 2\ell + 1) \log_{\ell+1} 2}{n \sum_{\ell=1}^{k} \log_{\ell+1} 2}$$

To obtain Gini-w_{min}, we also derive the numerator and denominator separately using Eq. 2.4. Note that for Gini-w_{min} we only consider cases where $km \leq n$ due to the unresolvable limitation of non-realisability, *Cause 4* (§2.4.3). First, we explain how to obtain the numerator. With the restriction of $km \leq n$, to make the recommendation the fairest, the km items that are recommended must be unique, leaving

$n - km$ items exposed. Thus, the items with the least to the most exposure are as follows: the first $n - km$ items receive zero exposure, the next m items will be recommended once each at position k, another set of m items each at position $k-1$, and so on until the last set of m items that will each be recommended at the top of the recommendation list. The numerator of Gini-w_{min} can then be calculated as follows:

$$\begin{aligned}
&\sum_{j=(n-km)+1}^{n-km+m} (2j-n-1)\log_{k+1} 2 \\
&+ \sum_{j=(n-km+m)+1}^{n-km+2m} (2j-n-1)\log_{k} 2 \\
&+ \dots \\
&+ \sum_{j=(n-m)+1}^{n} (2j-n-1)\log_{2} 2 \\
&= \sum_{\ell=1}^{k} \sum_{j=n-\ell m+1}^{n-\ell m+m} (2j-n-1)\log_{\ell+1} 2
\end{aligned}$$

As for the denominator, it is obtained the same way as in Gini-w_{max}, resulting in $mn \sum_{\ell=1}^{k} \log_{\ell+1} 2$ as the total exposure received by the items remains the same for the same cut-off k and the number of user m. Putting the numerator and denominator together:

$$\text{Gini-w}_{\text{min}} = \frac{\sum_{\ell=1}^{k} \sum_{j=n-\ell m+1}^{n-\ell m+m} (2j-n-1)\log_{\ell+1} 2}{mn \sum_{\ell=1}^{k} \log_{\ell+1} 2}$$

2.9.1.5 Fraction of Satisfied Items

The most unfair case for ↑FSat produces FSat_{min} and the most fair case for ↑FSat produces FSat_{max}. Note that $k \leq n$ thus $1 \geq \frac{k}{n} \Leftrightarrow m \geq \frac{km}{n} \geq \left\lfloor \frac{km}{n} \right\rfloor$.

$$\text{FSat}_{\text{min}} = \frac{1}{n}\left(k \cdot \delta\left(m \geq \left\lfloor \frac{km}{n} \right\rfloor\right)\right) = \frac{k}{n}$$

$$\begin{aligned}\mathrm{FSat}_{\max} &= \frac{1}{n}(n - km \bmod n) \cdot \delta\left(\left\lfloor \frac{km}{n} \right\rfloor \geq \left\lfloor \frac{km}{n} \right\rfloor\right) \\ &\quad + \frac{1}{n}(km \bmod n) \cdot \delta\left(\left\lfloor \frac{km}{n} \right\rfloor + 1 \geq \left\lfloor \frac{km}{n} \right\rfloor\right) \\ &= \frac{1}{n}[(n - km \bmod n) + km \bmod n] = \frac{n}{n} = 1\end{aligned}$$

2.9.1.6 Proofs for the maximum value of VoCD

First, we prove in Theorem 2.1 that when there is only one pair of similar items, the maximum VoCD value can be obtained when the two items are recommended 1 time and m times each. We then use Theorem 2.1 to show that when there are more than one pair of similar items, the maximum VoCD value does not increase (Theorem 2.2).

Theorem 2.1. *If there is only one pair of $(i, i') \in A$, $VoCD_{\max}$ is obtained when $\sum_{u \in U} 1_{R_u^k}(i) = 1$ and $\sum_{u \in U} 1_{R_u^k}(i') = m$*

Proof. We prove by contradiction: if an item i is recommended, $1 \leq \sum_{u \in U} 1_{R_u^k}(i) \leq m$. Suppose $\exists x_1, x_2$ where $1 < x_1 < x_2 < m$ such that $\frac{m-1}{m} < \frac{x_2 - x_1}{x_2} \Leftrightarrow \frac{-1}{m} < \frac{-x_1}{x_2} \Leftrightarrow \frac{1}{m} > \frac{x_1}{x_2} \Leftrightarrow m < \frac{x_2}{x_1}$. However, $x_2 < m \Leftrightarrow \frac{x_2}{x_1} < \frac{m}{x_1} < m$, which contradicts the previous inequality, so it must be $x_1 = 1$ and $x_2 = m$. □

Theorem 2.2. *The max VoCD score does not increase with $|A|$.*

Proof. Case 1: for each item pair in A that is disjoint from the other item pairs, e.g., (i_1, i_2) and (i_3, i_4), by Theorem 2.1, the maximum score for each pair and hence the average score of those pairs, is still $\frac{m-1}{m} - \beta$. Case 2: suppose the pairs are not disjoint, e.g., $(i_1, i_3), (i_2, i_3)$, and i_1, i_2, i_3 are recommended x_1, x_2, x_3 times respectively, where $1 \leq x_1 \leq x_2 \leq x_3 \leq m$. We show that for Case 2, the maximum value does not increase.

Note that $x_3 \leq m \Leftrightarrow -\frac{1}{x_3} \leq -\frac{1}{m} \Leftrightarrow -\frac{x_1}{2x_3} \leq -\frac{x_1}{2m} \leq -\frac{1}{2m}$ and $-\frac{x_2}{2x_3} \leq -\frac{1}{2m}$. Thus, $\frac{1}{2}\left(\frac{x_3 - x_1}{x_3} + \frac{x_3 - x_2}{x_3}\right) = 1 - \frac{x_1}{2x_3} - \frac{x_2}{2x_3} \leq 1 - \frac{1}{m}$. □

2.9.2 Extended Results of Experiments

2.9.2.1 Experimental set-up

We first report the hyperparameter search space for each recommender (Tab. 2.8) and the optimal hyperparameters for each model and each dataset (Tab. 2.9).

Table 2.8: Hyperparameter search space for each recommender.

	Hyperparameter search space
ItemKNN	k: [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 300, 400, 500, 1000] shrink: [0.0, 0.5, 1.0]
SLIM	alpha: [0.2, 0.5, 0.8, 1.0] l1 ratio: [0.01, 0.02, 0.05, 0.1, 0.5]
BPR	embedding size: [16, 32, 64, 128, 256, 512, 1024, 2048, 4096] lr: [5e-5, 1e-4, 5e-4, 1e-3, 5e-3, 1e-2, 5e-2]
NGCF	dropout prob: [0.1, 0.2] embedding size: [64, 128, 256, 512] hidden size: [[64, 64, 64], [128, 128, 128], [256, 256, 256]] lr: [5e-4, 1e-3, 5e-3]
NeuMF	dropout prob: [0.1, 0.2] hidden size: [[128, 64], [128, 64, 32], [64, 32, 16], [32, 16, 8]] lr: [5e-4, 1e-3, 5e-3, 1e-2]
MultiVAE	dropout prob: [0.1, 0.2, 0.5] hidden size: [[100], [300], [600]] latent dimension: [64, 128, 256] lr: [5e-4, 1e-3, 5e-3, 1e-2]

Table 2.9: Optimal hyperparameters for each recommender and each dataset.

	ItemKNN	SLIM	BPR	NGCF	NeuMF	MultiVAE
Amazon-lb	k: 20, shrink: 1.0	alpha: 0.2, l1 ratio: 0.01	embedding size: 4096, lr: 0.0001	dropout prob: 0.2, embedding size: 256, hidden size: [128,128,128], lr: 0.005	dropout prob: 0.2, hidden size: [128,64,32], lr: 0.005	dropout prob: 0.5, hidden size: [600], latent dim: 128, lr: 0.005
Lastfm	k: 300, shrink: 1.0	alpha: 0.2, l1 ratio: 0.01	embedding size: 2048, lr: 0.0005	dropout prob: 0.2, embedding size: 512, hidden size: [256,256,256], lr: 0.001	dropout prob: 0.2, hidden size: [32,16,8], lr: 0.001	dropout prob: 0.5, hidden size: [600], latent dim: 64, lr: 0.0005
Ml-1m	k: 1000, shrink: 1.0	alpha: 0.2, l1 ratio: 0.01	embedding size: 2048, lr: 0.0001	dropout prob: 0.1, embedding size: 256, hidden size: [256,256,256], lr: 0.0005	dropout prob: 0.2, hidden size: [64,32,16], lr: 0.0005	dropout prob: 0.1, hidden size: [600], latent dim: 64, lr: 0.01
Book-x	k: 20, shrink: 1.0	alpha: 0.2, l1 ratio: 0.01	embedding size: 4096, lr: 0.0001	dropout prob: 0.1, embedding size: 512, hidden size: [256,256,256], lr: 0.0005	dropout prob: 0.1, hidden size: [32,16,8], lr: 0.001	dropout prob: 0.5, hidden size: [600], latent dim: 128, lr: 0.0005
Amazon-is	k: 250, shrink: 1.0	alpha: 0.2, l1 ratio: 0.01	embedding size: 4096, lr: 0.0001	dropout prob: 0.1, embedding size: 512, hidden size: [256,256,256], lr: 0.001	dropout prob: 0.2, hidden size: [128,64,32], lr: 0.001	dropout prob: 0.5, hidden size: [600], latent dim: 128, lr: 0.005
Amazon-dm	k: 10, shrink: 1.0	alpha: 0.2, l1 ratio: 0.01	embedding size: 4096, lr: 0.0001	dropout prob: 0.2, embedding size: 512, hidden size: [256,256,256], lr: 0.001	dropout prob: 0.2, hidden size: [128,64,32], lr: 0.0005	dropout prob: 0.2, hidden size: [600], latent dim: 256, lr: 0.01

2.9.2.2 Analysis of relevance and fairness

We present the performance scores of the recommender systems on the Amazon-* and Book-x datasets in Tab. 2.10 & 2.11. The scores of the original version of Ent cannot be calculated due to zero divisions error for the same reasons explained in §2.5.2. The constant scores of II-D have also been explained in the same section. The best relevance and fairness scores are bolded.

BPR generally performs the best in relevance, with the exception of Amazon-lb where NeuMF is best, while ItemKNN gives the best fairness scores. Other trends observed on Amazon-* and Book-x are similar to that on Lastfm and Ml-1m.

2.9.2.3 Correlation between measures

We show the Kendall's Tau values between relevance measures and fairness measures for the Amazon-* and Book-x datasets in Fig. 2.9–2.12.

2.9.2.4 Max/min achievable fairness

The results of the max/min achievable fairness experiment for Amazon-* and Book-x are in Fig. 2.13–2.16.

2.9.2.5 Sliding window: relevance and fairness at different rank positions

The results of the sliding window experiment for Amazon-* and Book-x are in Fig. 2.17–2.18.

2.9.2.6 Measure strictness and sensitivity through artificial insertion of items

We present in Fig. 2.19 the extended results of artificially inserting least exposed (LE) and relevant items for $m = \{100, 500\}$ for relevance measures and fairness measures. The changes in scores are less stable compared to $m = 1000$, but the general trends are the same.

We also experiment with the artificial insertion of multiple copies of items that are already in the recommendation list and the insertion of irrelevant items, using a similar methodology. We refer to the insertion of multiple item copies as inserting the *most exposed (ME) items*, as in this experiment we aim to maximise exposure of as few items as possible. This is done by iteratively inserting a copy of several items that currently have the most exposure, one copy at a time. We swap the starting and ending recommendation list of the artificial insertion of LE and relevant items such

that at the end of the experiments, only k unique items are in the recommendation list. These k items will get the most exposure, while the rest of the items in the dataset get zero exposure. The item replacement is still done from the bottom of the recommendation list. In Fig. 2.20, we see that the trends of the measures are similar, but the opposite to that of the artificial insertion of LE and relevant items.

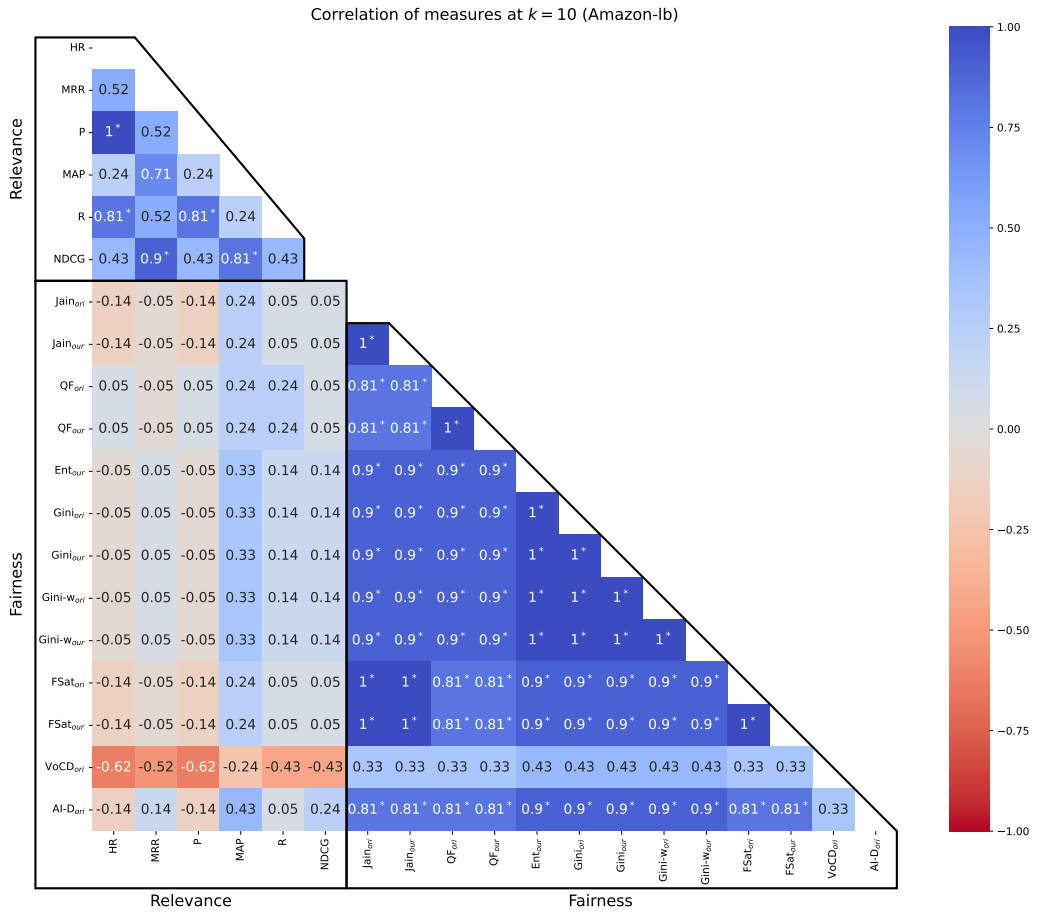


Figure 2.9: Correlation (Kendall's τ) between relevance and fairness measures for Amazon-lb. Asterisk (*) denotes a statistically significant correlation ($\alpha = 0.05$), after applying the Benjamini-Hochberg procedure.

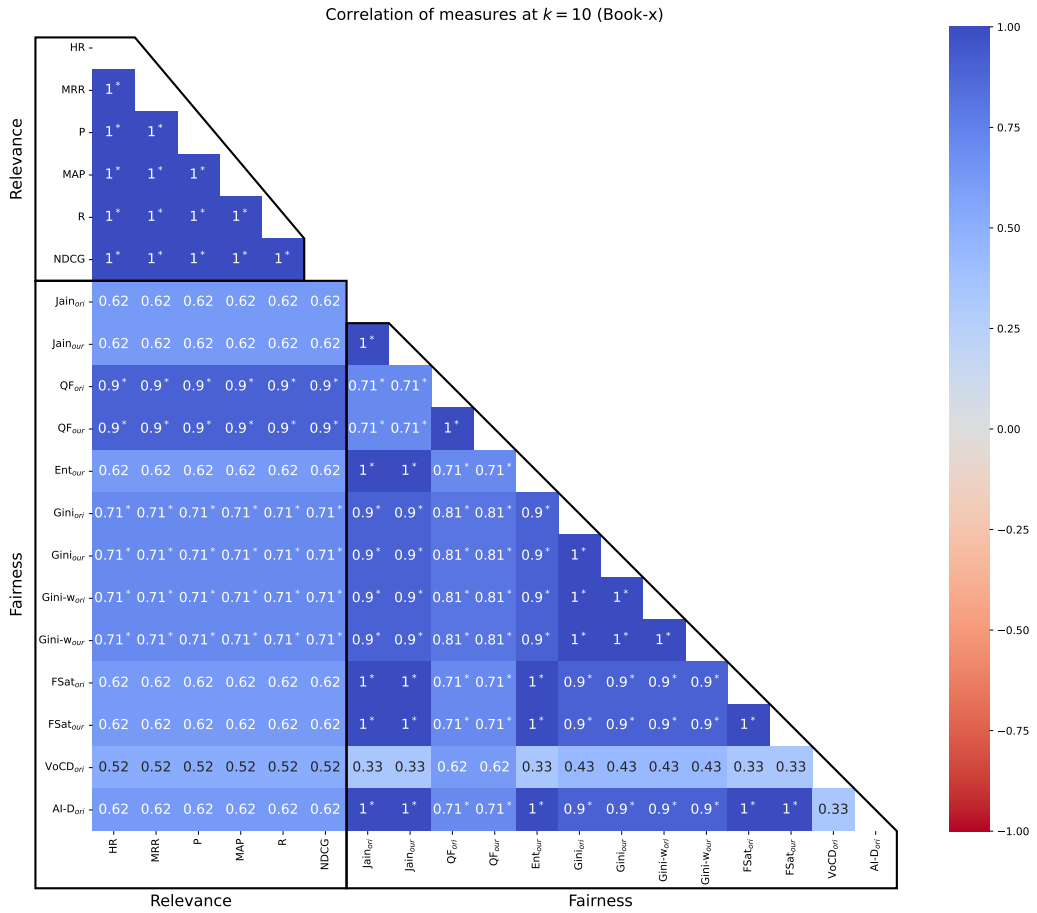


Figure 2.10: Correlation (Kendall's τ) between relevance and fairness measures for Book-x. Asterisk (*) denotes a statistically significant correlation ($\alpha = 0.05$), after applying the Benjamini-Hochberg procedure.

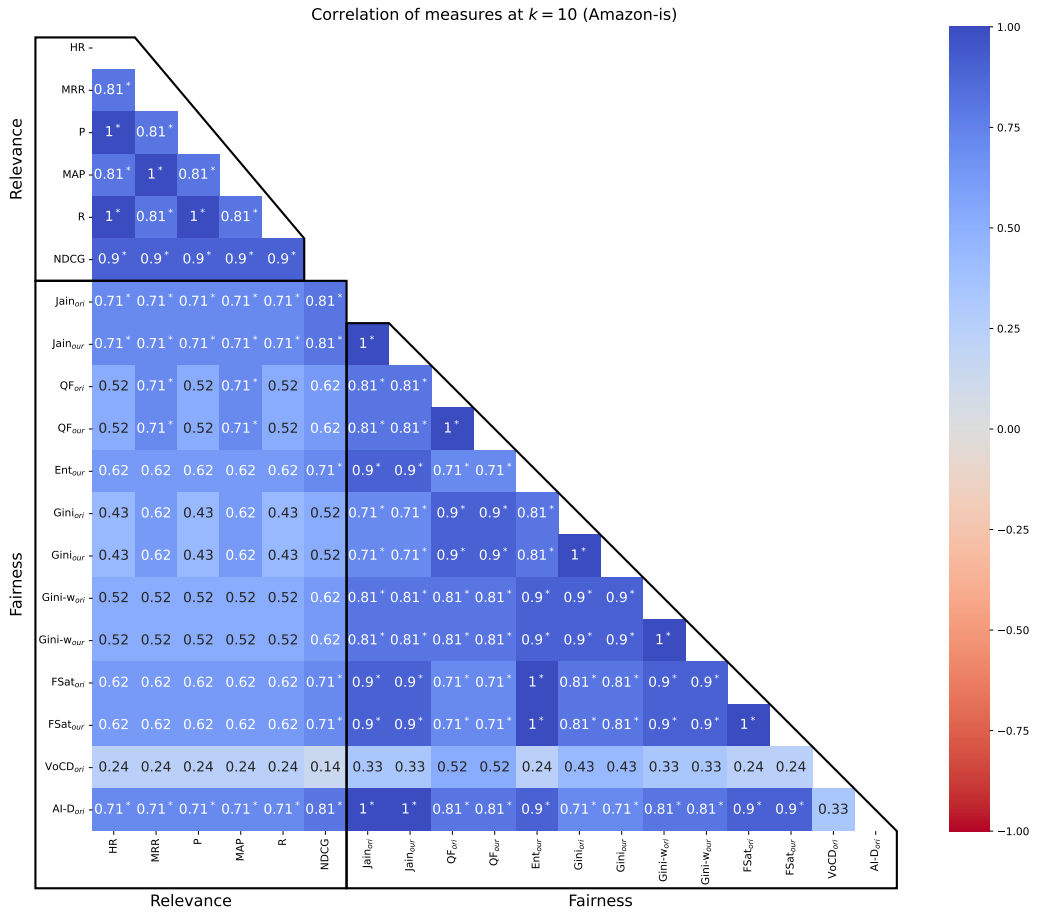


Figure 2.11: Correlation (Kendall's τ) between relevance and fairness measures for Amazon-is. Asterisk (*) denotes a statistically significant correlation ($\alpha = 0.05$), after applying the Benjamini-Hochberg procedure.

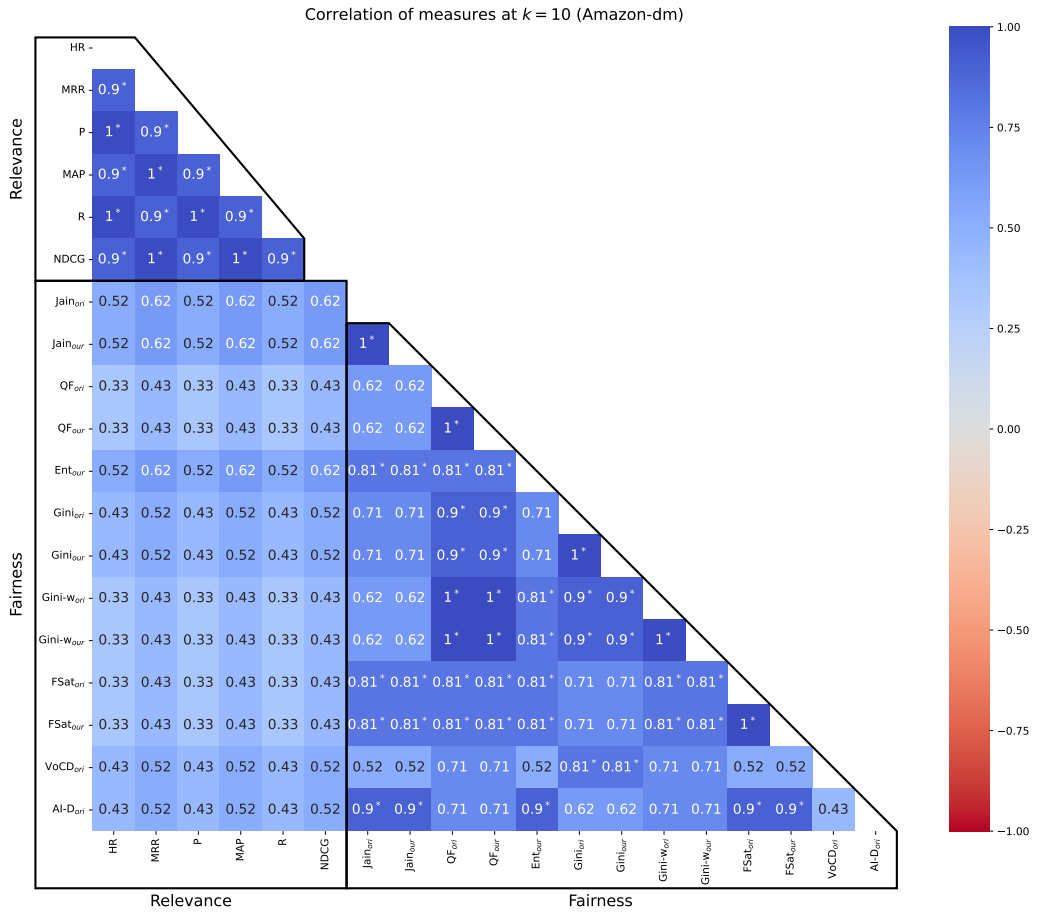


Figure 2.12: Correlation (Kendall's τ) between relevance and fairness measures for Amazon-dm. Asterisk (*) denotes a statistically significant correlation ($\alpha = 0.05$), after applying the Benjamini-Hochberg procedure.

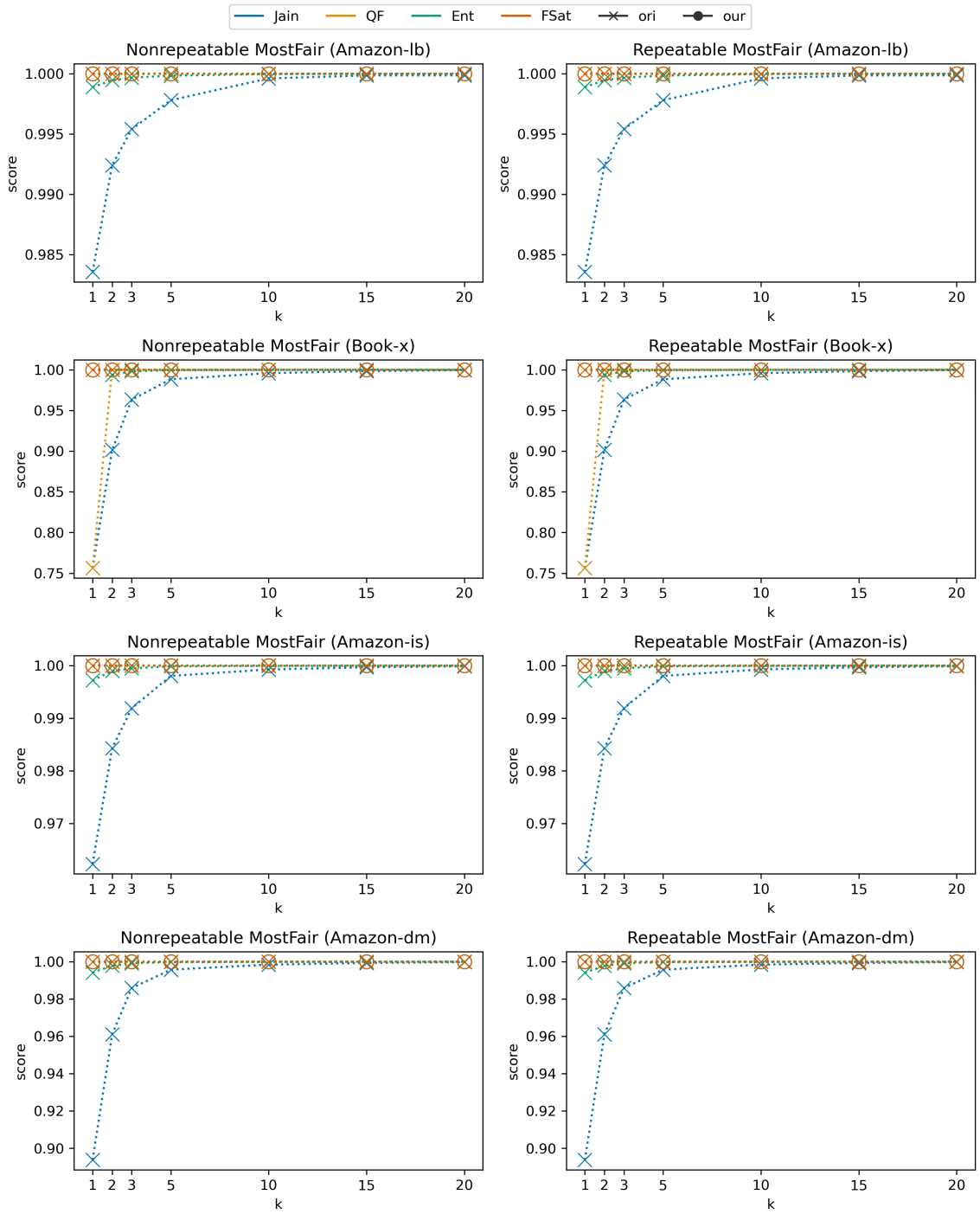


Figure 2.13: Most fair scores with varying k for higher-is-fairer fairness measures on Amazon-* and Book-x.

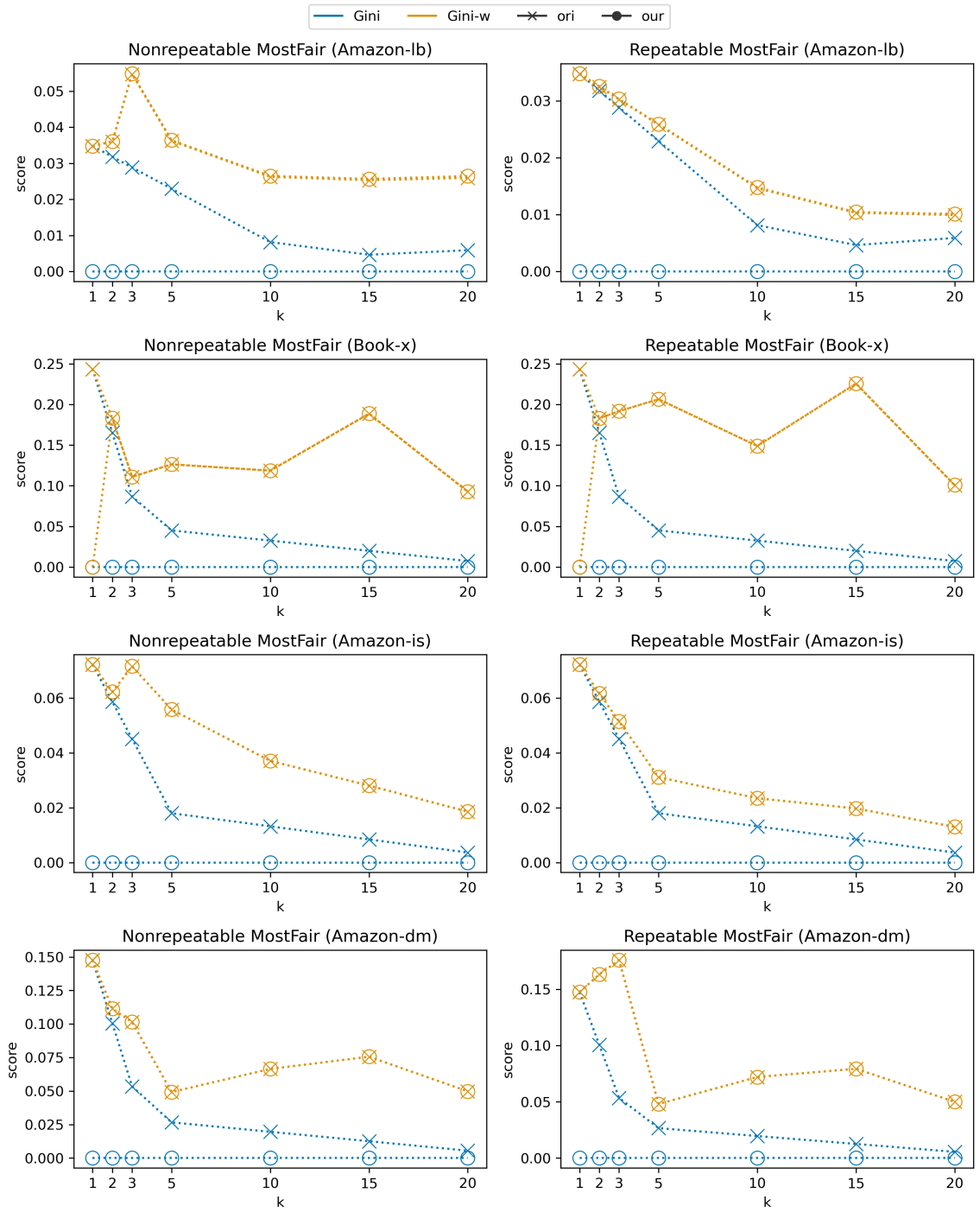


Figure 2.14: Most fair scores with varying k for lower-is-fairer fairness measures on Amazon-* and Book-x.

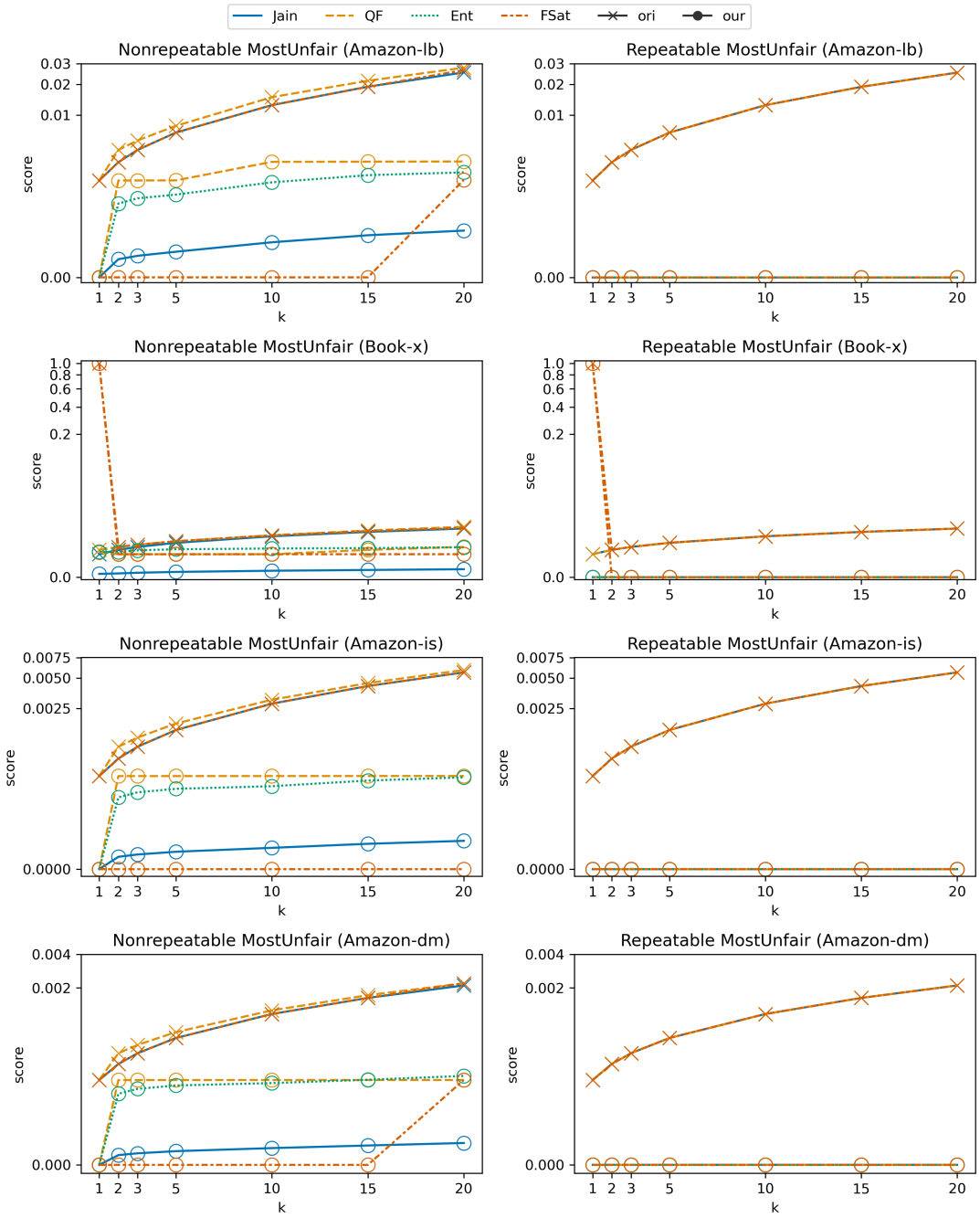


Figure 2.15: Most unfair scores with varying k for higher-is-fairer fairness measures on Amazon-* and Book-x.

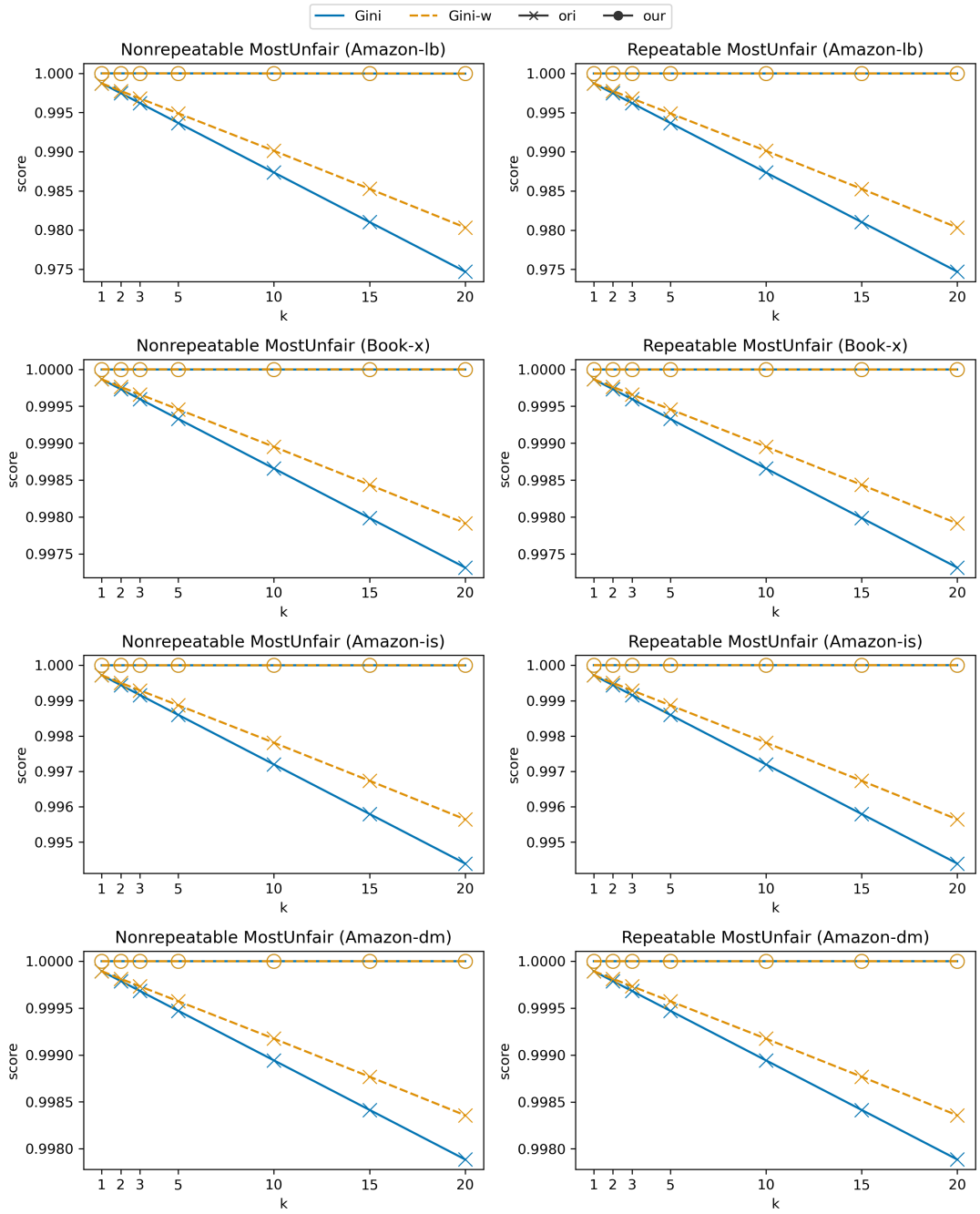


Figure 2.16: Most unfair scores with varying k for lower-is-fairer fairness measures on Amazon-* and Book-x.

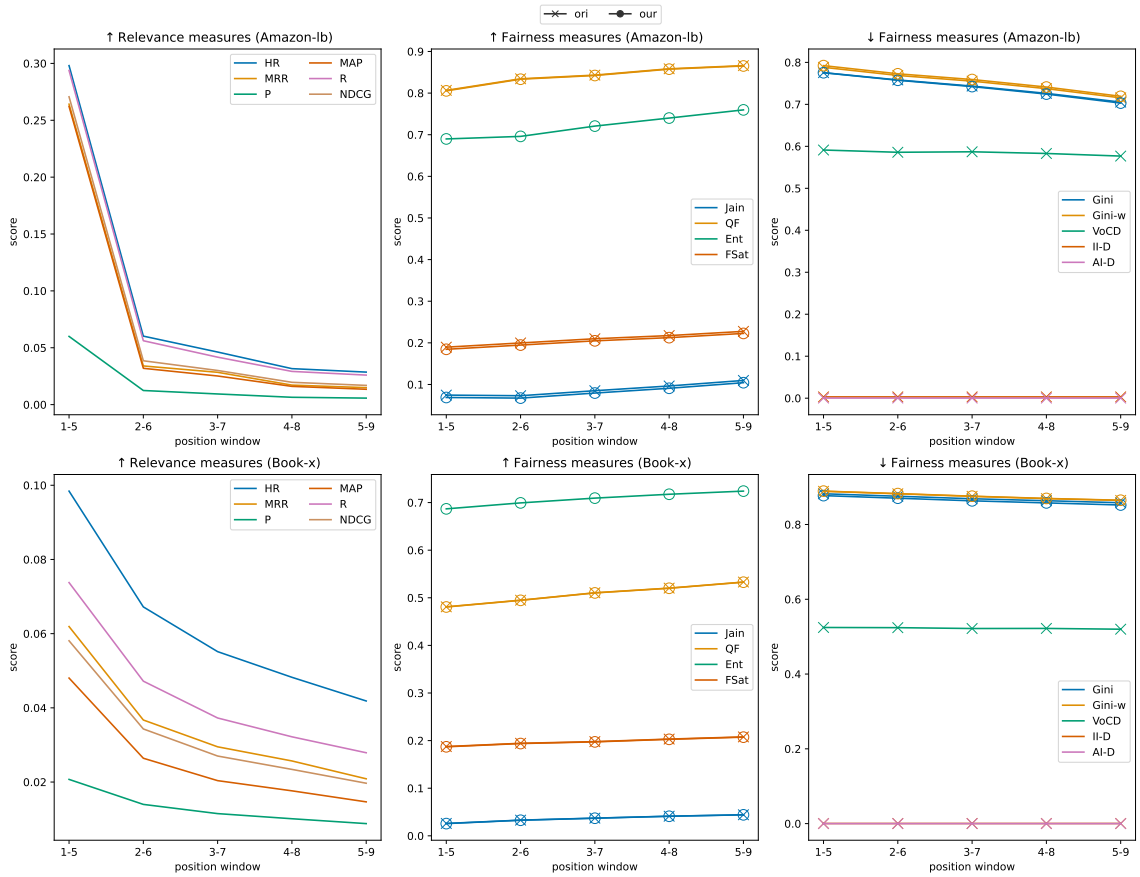


Figure 2.17: Sliding window evaluation for BPR model, on Amazon-lb and Book-x. Each row is for one dataset, each column is for the different groups of measures (relevance, higher-is-better fairness, lower-is-better fairness measures). II-D and AI-D lines overlap.

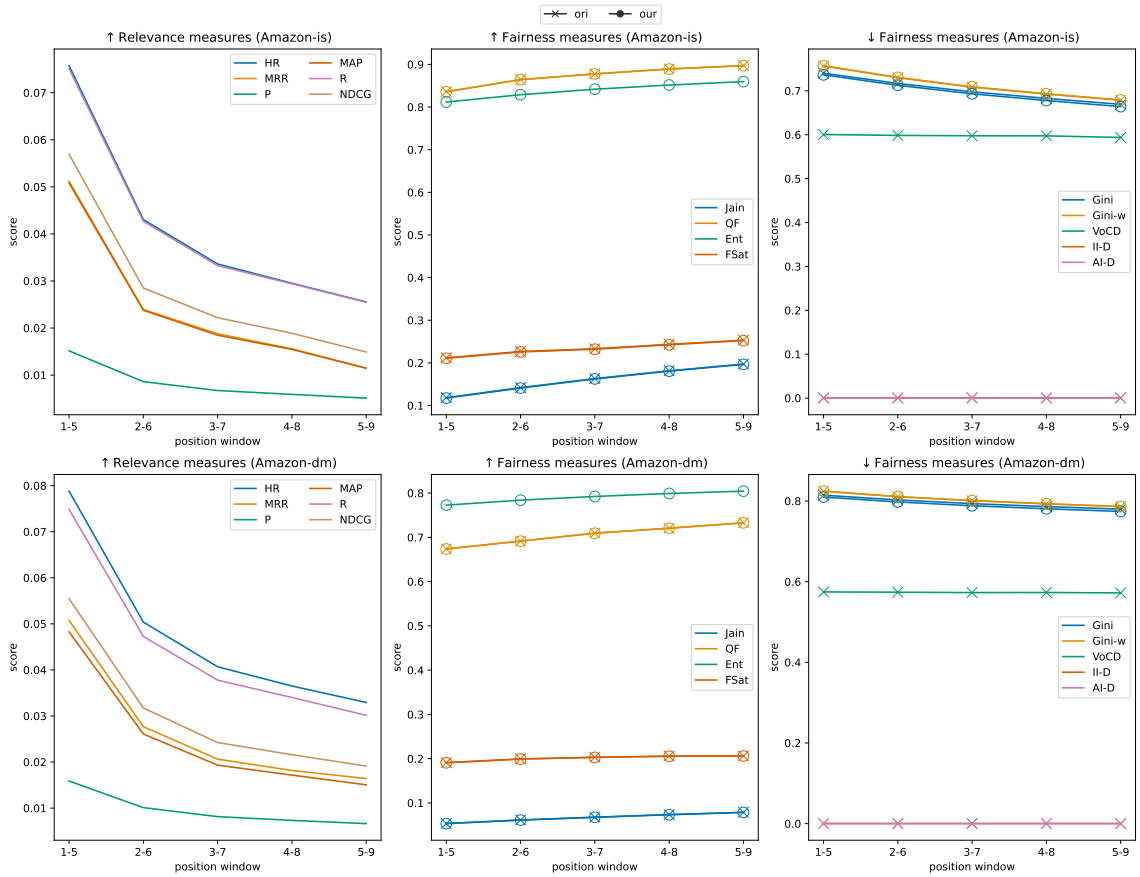


Figure 2.18: Sliding window evaluation for BPR model, on Amazon-is and Amazon-dm. Each row is for one dataset, each column is for the different groups of measures (relevance, higher-is-better fairness, lower-is-better fairness measures). II-D and AI-D lines overlap.

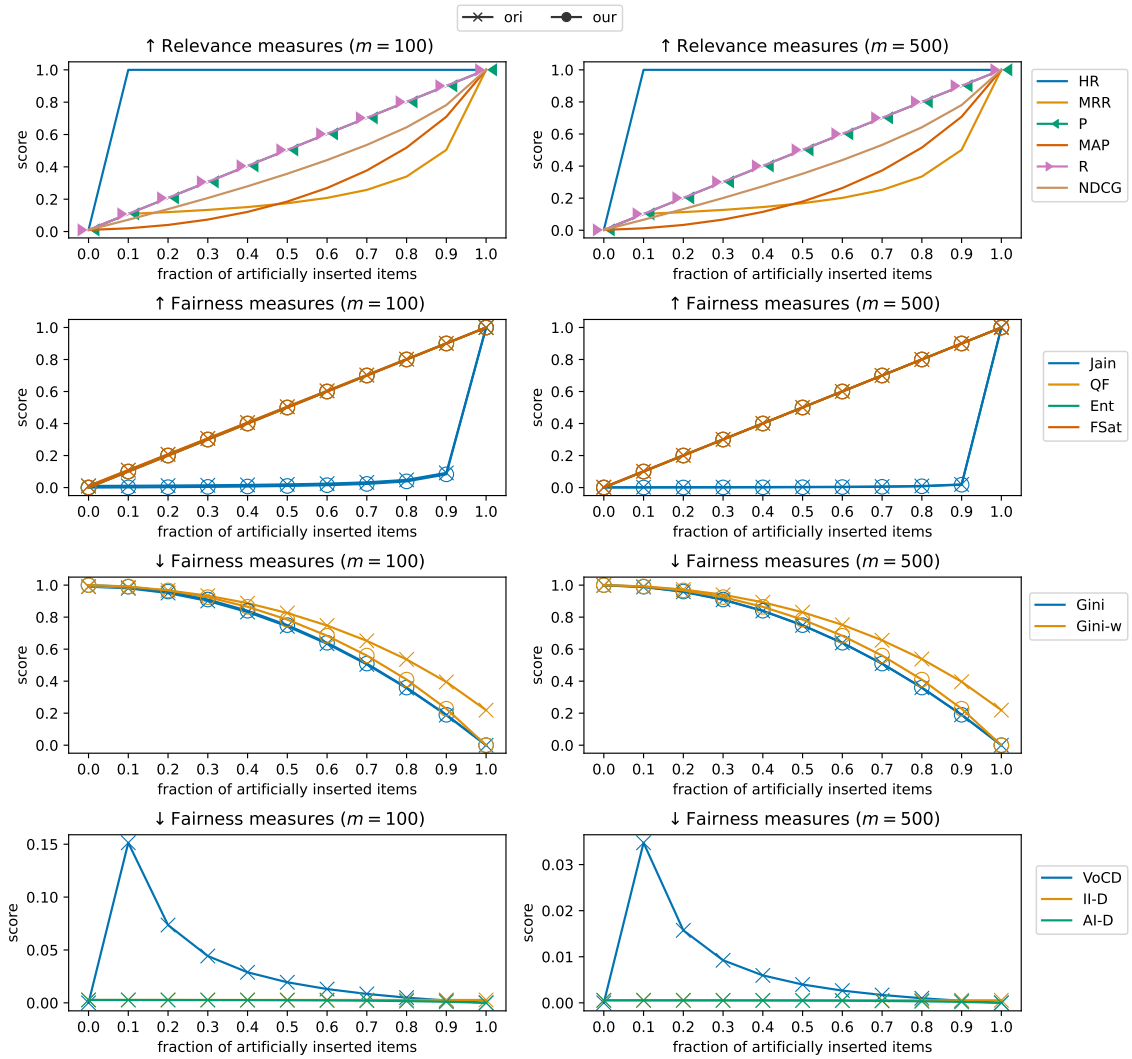


Figure 2.19: Results for jointly LE and relevant item insertion for $m \in \{100, 500\}$. All measures are calculated at $k = 10$. $\mathrm{QF}_{\mathrm{ori}}$ and $\mathrm{FSat}_{\mathrm{ori}}$ overlap. $\mathrm{QF}_{\mathrm{our}}$, $\mathrm{FSat}_{\mathrm{our}}$, and $\mathrm{Ent}_{\mathrm{our}}$ also overlap.

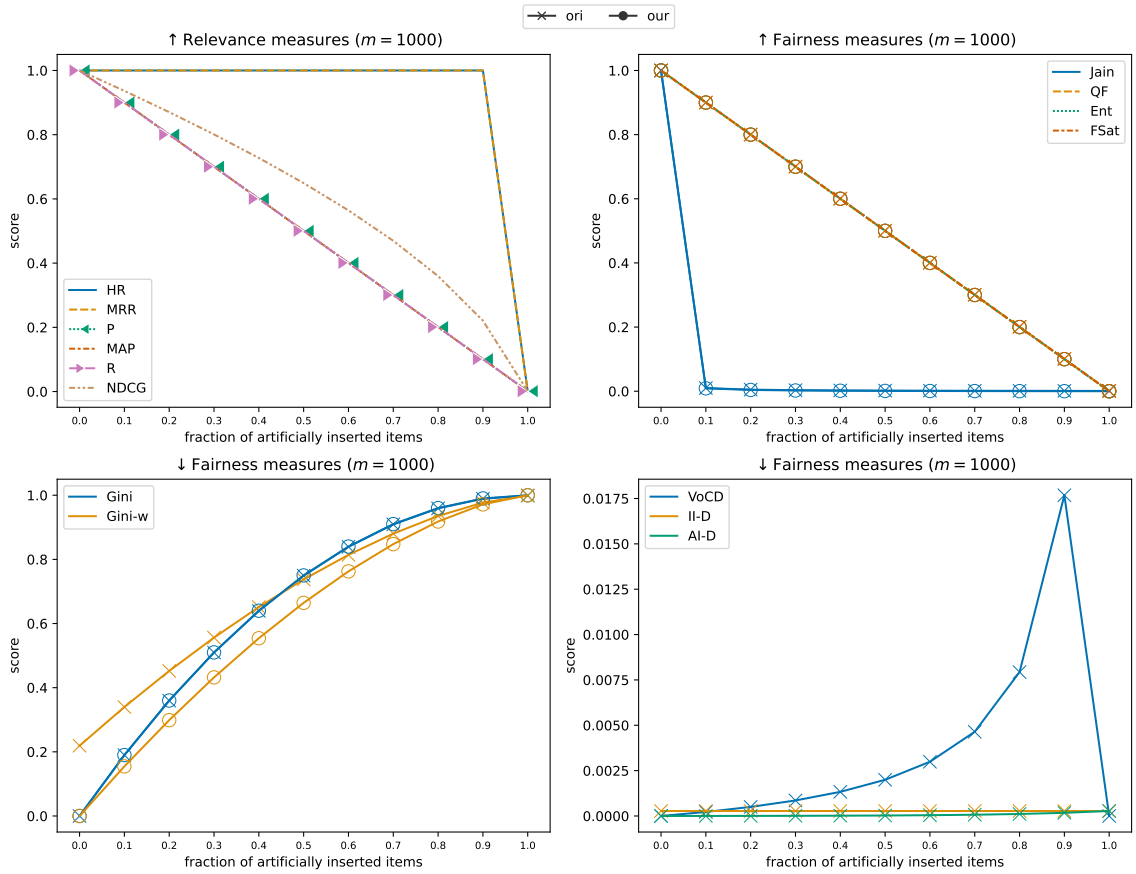


Figure 2.20: Results for jointly most exposed (ME) and irrelevant item insertion. All measures are calculated at $k = 10$. $\mathrm{QF}_{\mathrm{ori}}$ and $\mathrm{FSat}_{\mathrm{ori}}$ overlap. $\mathrm{QF}_{\mathrm{our}}$, $\mathrm{FSat}_{\mathrm{our}}$, and $\mathrm{Ent}_{\mathrm{our}}$ also overlap.

Table 2.10: Relevance (REL) and fairness (FAIR) scores of the recommender models for Amazon-lb and Book-x. The most relevant and most fair score per measure is in bold. ↑ means the higher the better, ↓ the lower the better. 'nan' stands for 'not a number'.

			Pop*	ItemKNN	SLIM	BPR	NGCF	NeuMF	MultiVAE
Amazon-lb	REL	↑ HR	0.257908	0.309611	0.313869	0.320560	0.305961	**0.325426**	0.321776
		↑ MRR	0.222141	0.265047	**0.270739**	0.267188	0.263003	0.267511	0.264885
		↑ P	0.025912	0.031022	0.031752	0.032360	0.030718	**0.032968**	0.032664
		↑ MAP	0.221568	0.264694	**0.268219**	0.265084	0.261340	0.263527	0.261339
		↑ R	0.252251	0.308698	0.306767	0.315237	0.302616	**0.318375**	0.314852
		↑ NDCG	0.228580	0.275196	**0.278262**	0.277727	0.271526	0.277666	0.275024
	FAIR	↑ Jain_{ori}	0.017661	0.188400	0.050735	0.134711	**0.193167**	0.060450	0.028521
		↑ Jain_{our}	0.005085	0.178079	0.038596	0.123681	**0.182909**	0.048439	0.016089
		↑ QF_{ori}	0.024020	**0.978508**	0.379267	0.921618	0.967130	0.790139	0.386852
		↑ QF_{our}	0.011524	**0.978233**	0.371319	0.920615	0.966709	0.787452	0.379001
		↑ Ent_{ori}	nan	nan	nan	nan	nan	nan	nan
		↑ Ent_{our}	0.094992	**0.822309**	0.460701	0.731135	0.812621	0.564546	0.310875
		↑ FSat_{ori}	0.022756	0.285714	0.134008	0.212389	**0.294564**	0.158028	0.074589
		↑ FSat_{our}	0.010243	0.276569	0.122919	0.202305	**0.285531**	0.147247	0.062740
		↓ Gini_{ori}	0.983735	**0.580725**	0.915792	0.710528	0.606176	0.841397	0.955542
		↓ Gini_{our}	0.996300	**0.584731**	0.926915	0.717292	0.610723	0.850940	0.967509
		↓ $\text{Gini-w}_{\text{ori}}$	0.986481	**0.605205**	0.914661	0.731609	0.628605	0.861775	0.959218
		↓ $\text{Gini-w}_{\text{our}}$	0.996317	**0.611240**	0.923782	0.738904	0.634873	0.870368	0.968782
		↓ VoCD_{ori}	**0.509398**	0.560071	0.654539	0.617237	0.590049	0.645770	0.661114
		↓ II-D_{ori}	**0.003439**	**0.003439**	**0.003439**	**0.003439**	**0.003439**	**0.003439**	**0.003439**
		↓ AI-D_{ori}	0.002443	**0.000158**	0.000610	0.000280	0.000165	0.000728	0.001385
Book-x	REL	↑ HR	0.034581	0.115269	0.059762	**0.130342**	0.103564	0.085299	0.089910
		↑ MRR	0.012664	0.063709	0.034535	**0.066125**	0.041422	0.037474	0.039687
		↑ P	0.003476	0.012999	0.006349	**0.014276**	0.010977	0.008973	0.009576
		↑ MAP	0.008713	0.047170	0.023412	**0.050778**	0.032601	0.029294	0.031962
		↑ R	0.022909	0.082607	0.037784	**0.097505**	0.078578	0.064687	0.069501
		↑ NDCG	0.013074	0.059732	0.029559	**0.065820**	0.045847	0.039839	0.042925
	FAIR	↑ Jain_{ori}	0.001420	**0.375894**	0.002297	0.036191	0.020391	0.030978	0.050532
		↑ Jain_{our}	0.000080	**0.376670**	0.000961	0.035046	0.019157	0.029804	0.049469
		↑ QF_{ori}	0.002414	**0.899799**	0.019584	0.671630	0.662777	0.519920	0.587928
		↑ QF_{our}	0.001075	**0.899664**	0.018267	0.671189	0.662324	0.519275	0.587374
		↑ Ent_{ori}	nan	nan	nan	nan	nan	nan	nan
		↑ Ent_{our}	0.014182	**0.916559**	0.156894	0.688733	0.623813	0.646908	0.716425
		↑ FSat_{ori}	0.002012	**0.392622**	0.017170	0.158820	0.120724	0.145942	0.188062
		↑ FSat_{our}	0.000672	**0.391807**	0.015850	0.157690	0.119543	0.144795	0.186971
		↓ Gini_{ori}	0.998615	**0.548876**	0.996714	0.853353	0.877575	0.894874	0.848862
		↓ Gini_{our}	0.999955	**0.534458**	0.997988	0.849604	0.874675	0.892579	0.844955
		↓ $\text{Gini-w}_{\text{ori}}$	0.998906	**0.582320**	0.996386	0.863144	0.889082	0.904719	0.860833
		↓ $\text{Gini-w}_{\text{our}}$	0.999954	**0.582930**	0.997431	0.864049	0.890015	0.905668	0.861736
		↓ VoCD_{ori}	0.663365	0.564473	0.720722	0.570963	**0.537809**	0.600639	0.611070
		↓ II-D_{ori}	**0.000368**	**0.000368**	**0.000368**	**0.000368**	**0.000368**	**0.000368**	**0.000368**
		↓ AI-D_{ori}	0.000350	**0.000001**	0.000176	0.000013	0.000027	0.000015	0.000009

*The scores of our FAIR measures for Pop are not 0 or 1, because in our experiment set-up, items from users' train or validation splits are excluded from the top k recommendation list.

Table 2.11: Relevance (REL) and fairness (FAIR) scores of the recommender models for Amazon-is and Amazon-dm. The most relevant and most fair score per measure is in bold. ↑ means the higher the better, ↓ the lower the better. 'nan' stands for 'not a number'.

			Pop*	ItemKNN	SLIM	BPR	NGCF	NeuMF	MultiVAE
Amazon-is	REL	↑ HR	0.031792	0.084271	0.028902	**0.100548**	0.095224	0.077122	0.081838
		↑ MRR	0.011671	0.047550	0.023075	**0.054403**	0.048366	0.042206	0.041294
		↑ P	0.003179	0.008427	0.002890	**0.010055**	0.009522	0.007712	0.008184
		↑ MAP	0.011593	0.047313	0.022880	**0.054002**	0.048093	0.041905	0.041036
		↑ R	0.031462	0.083916	0.028572	**0.099825**	0.094717	0.076387	0.080993
		↑ NDCG	0.016224	0.055902	0.024254	**0.064820**	0.059064	0.050052	0.050436
	FAIR	↑ Jain_{ori}	0.002925	**0.493178**	0.002958	0.167723	0.101140	0.061059	0.093401
		↑ Jain_{our}	0.000124	**0.492108**	0.000156	0.165503	0.098685	0.058463	0.090919
		↑ QF_{ori}	0.004483	**0.989633**	0.014570	0.956851	0.869151	0.919866	0.856823
		↑ QF_{our}	0.001686	**0.989604**	0.011801	0.956729	0.868783	0.919640	0.856420
		↑ Ent_{ori}	nan	nan	nan	nan	nan	nan	nan
		↑ Ent_{our}	0.012261	**0.939484**	0.024132	0.832303	0.746126	0.741374	0.766085
		↑ FSat_{ori}	0.003642	**0.373494**	0.008686	0.223032	0.171477	0.170636	0.204819
		↑ FSat_{our}	0.000843	**0.371734**	0.005901	0.220849	0.169149	0.168306	0.202585
		↓ Gini_{ori}	0.997136	**0.445874**	0.997066	0.679338	0.798275	0.763940	0.769841
		↓ Gini_{our}	0.999937	**0.439697**	0.999866	0.676963	0.797837	0.762944	0.768941
		↓ $\text{Gini-w}_{\text{ori}}$	0.997738	**0.476640**	0.997518	0.703146	0.818197	0.784878	0.784863
		↓ $\text{Gini-w}_{\text{our}}$	0.999926	**0.477685**	0.999705	0.704688	0.819991	0.786599	0.786584
		↓ VoCD_{ori}	0.592760	**0.523208**	0.787225	0.615085	0.640122	0.616732	0.648353
		↓ II-D_{ori}	**0.000768**	**0.000768**	**0.000768**	**0.000768**	**0.000768**	**0.000768**	**0.000768**
		↓ AI-D_{ori}	0.000738	**0.000002**	0.000705	0.000011	0.000019	0.000036	0.000020
Amazon-dm	REL	↑ HR	0.022809	0.087660	0.005702	**0.108596**	0.093787	0.073872	0.079660
		↑ MRR	0.009252	0.048030	0.004607	**0.054654**	0.043576	0.033947	0.036754
		↑ P	0.002289	0.008928	0.000570	**0.011004**	0.009489	0.007489	0.008051
		↑ MAP	0.008463	0.044812	0.004106	**0.051930**	0.041533	0.032263	0.034956
		↑ R	0.020528	0.081430	0.005123	**0.102600**	0.089059	0.069847	0.075436
		↑ NDCG	0.011454	0.054259	0.004447	**0.064525**	0.053219	0.041538	0.044960
	FAIR	↑ Jain_{ori}	0.001092	**0.358718**	0.001080	0.068624	0.078606	0.038114	0.121317
		↑ Jain_{our}	0.000036	**0.358605**	0.000023	0.067745	0.077754	0.037155	0.120578
		↑ QF_{ori}	0.001902	**0.957620**	0.003593	0.843585	0.720672	0.843691	0.879624
		↑ QF_{our}	0.000846	**0.957575**	0.002539	0.843419	0.720377	0.843525	0.879496
		↑ Ent_{ori}	nan	nan	nan	nan	nan	nan	nan
		↑ Ent_{our}	0.008686	**0.940649**	0.008735	0.779117	0.768888	0.764866	0.840448
		↑ FSat_{ori}	0.001480	**0.404143**	0.002219	0.187804	0.199535	0.197738	0.246354
		↑ FSat_{our}	0.000423	**0.403512**	0.001164	0.186945	0.198688	0.196890	0.245556
		↓ Gini_{ori}	0.998925	**0.462579**	0.998932	0.778288	0.814128	0.771164	0.702994
		↓ Gini_{our}	0.999982	**0.452327**	0.999989	0.774693	0.811289	0.767418	0.697811
		↓ $\text{Gini-w}_{\text{ori}}$	0.999154	**0.493364**	0.999132	0.791188	0.828961	0.788203	0.718667
		↓ $\text{Gini-w}_{\text{our}}$	0.999979	**0.493772**	0.999957	0.791841	0.829646	0.788854	0.719261
		↓ VoCD_{ori}	0.668615	**0.532741**	0.796493	0.615043	0.643785	0.611526	0.616569
		↓ II-D_{ori}	**0.000290**	**0.000290**	**0.000290**	**0.000290**	**0.000290**	**0.000290**	**0.000290**
		↓ AI-D_{ori}	0.000282	**0.000000**	0.000278	0.000004	0.000004	0.000008	0.000002

*The scores of our FAIR measures for Pop are not 0 or 1, because in our experiment set-up, items from users' train or validation splits are excluded from the top k recommendation list.

Chapter 3

Can We Trust Recommender System Fairness Evaluation? The Role of Fairness and Relevance

Abstract

Relevance and fairness are two major objectives of recommender systems (RSs). Recent work proposes measures of RS fairness that are either independent from relevance (fairness-only) or conditioned on relevance (joint measures). While fairness-only measures have been studied extensively, we look into whether joint measures can be trusted. We collect all joint evaluation measures of RS relevance and fairness, and ask: How much do they agree with each other? To what extent do they agree with relevance/fairness measures? How sensitive are they to changes in rank position, or to increasingly fair and relevant recommendations? We empirically study for the first time the behaviour of these measures across 4 real-world datasets and 4 recommenders. We find that most of these measures: i) correlate weakly with one another and even contradict each other at times; ii) are less sensitive to rank position changes than relevance- and fairness-only measures, meaning that they are less granular than traditional RS measures; and iii) tend to compress scores at the low end of their range, meaning that they are not very expressive. We counter the above limitations with a set of guidelines on the appropriate usage of such measures, i.e., they should be used with caution due to their tendency to contradict each other and of having a very small empirical range.

3.1 Introduction

Recent increased focus on fairness in recommender systems (RSs) has led to studies on how to evaluate different notions of fairness in RS. A recent survey [228] shows that prior work on fairness evaluation in RS mainly focuses on group fairness (e.g., [9, 175, 252]), but less so on individual fairness. Individual fairness is commonly understood as treating similar individuals similarly [63]. Unlike group fairness evaluation, evaluating individual fairness does not require information on sensitive attributes (e.g., gender, age) to identify protected groups [124]. Such information is often unavailable due to privacy and legal issues. Further, intersectionality between different group characteristics complicates group fairness [44, 65]. Individual fairness is known to lead into group fairness, but not vice versa [25]. Overall, individual fairness gives a broader view by assessing distribution across all individuals in the population [124]. For all these reasons, we focus on individual fairness, particularly individual item fairness, which is typically broadly defined w.r.t. exposure received by items, i.e., how uniform the exposure distribution between items is [177]. Yet, fairness beyond exposure also matters, i.e., the exposure should be proportional to item relevance [19, 168, 209].

Individual item fairness is measured by measures that (i) are detached from relevance (fairness-only measures, defined by exposure); or (ii) are conditioned on relevance (joint measures considering exposure w.r.t. relevance). Measures of type (i) have been extensively analysed [177], but to our knowledge, this is not the case for measures of type (ii). The growing number of measures of type (ii) necessitates a thorough look into their usage in RS evaluation.

We present a comprehensive study into the empirical properties of all joint measures of individual item fairness and relevance, motivated by the question of how much can we practically trust these measures, particularly: **RQ1.** To what extent do the joint measures agree with existing relevance- and fairness-only measures? **RQ2.** To what extent do the joint measures agree with each other? **RQ3.** How sensitive are the joint measures across decreasing rank positions? and **RQ4.** How sensitive are the joint measures given increasingly fair and relevant recommendations?

We identify some alarming limitations in the measures, and we reflect on their best usage in practice. This is the first in-depth study on individual item fairness measures that consider relevance in RS.

3.2 Individual Item Fairness & Relevance

We present the notation (§3.2.1) and all existing joint evaluation measures of individual item fairness and relevance (§3.2.2).

3.2.1 Notation and Definitions

Given a set of n items, $I = \{i_1, i_2, \ldots, i_n\}$, and a set of m users, $U = \{u_1, u_2, \ldots, u_m\}$, an ordered list of the n items is created for each $u \in U$. This list is created in each recommendation round w, where $w \in \{1, 2, \ldots, W\}$; a round means an occurrence in which a user receives a list of recommendations. If an item i is *relevant* to user u, we write $r_{u,i} = 1$, otherwise $r_{u,i} = 0$. Relevance can also be denoted as real values, $r_{u,i} \in [0, 1]$. The list of user u's top k recommended items in round w is $L_{u,w}$ and the rank position of item i in user u's recommendation list is $z(u, i, w)$. For cases with only one recommendation round, user u's list of top k recommended items is L_u and the rank position of item i for user u is $z(u, i)$.

While several different definitions of fairness exist, the definitions commonly used in prior work on individual item fairness are closely linked to item exposure [9, 177, 252]. An item is *exposed* when it is recommended at the top k to a user. The probability of a user seeing an item exposed to them can be modelled using various *examination functions*, $e(\cdot)$. Examination functions typically assume that the viewing probability depends only on the position $z(u, i, w)$ or $z(u, i)$. This is a common choice across all measures in §3.2.2.

The examination functions used in this work are shown in Tab. 3.1: the linear examination function, e_{li} and its min-max normalised version $\tilde{e}_{\text{li}}$ apply a linear discount to each rank position up to k [23]. Meanwhile, discounts based on Discounted Cumulative Gain (DCG) [102] and Rank-Biased Precision (RBP) [154] are used in e_{DCG} [161, 205] and e_{RBP} [103, 234] respectively. The parameter γ in e_{RBP} is the user's patience, i.e., the probability of the user examining the next ranked item. The user patience parameter is commonly set at e.g., $\gamma \in \{0.8, 0.9\}$ [103, 234]. In the inverse examination function e_{inv}, the inverse of the rank position is used as a discount factor [191]. Overall, we use three types of examination functions (linear, discounted, and inverse), which assume that item exposure diminishes with decreasing ranking either linearly, or with an increasing penalty that is either proportional to the rank decrease or the inverse of the rank. Generally, the most punishing is the inverse function, and the least punishing is the DCG-based discount function.

Table 3.1: Examination functions used in this work. $\tilde{e}$ is the min-max normalised examination function.

	equation	measure	reference
linear	$e_{\text{li}}(u,i,w) = k+1-z(u,i,w)$ $\tilde{e}_{\text{li}}(u,i,w) = \frac{e_{\text{li}}(u,i,w)-1}{k-1} = \frac{k-z(u,i,w)}{k-1}$	IAA	[23]
DCG	$e_{\text{DCG}}(u,i,w) = 1/\log_2(z(u,i,w)+1)$	IFD	[161, 205]
RBP	$e_{\text{RBP}}(u,i,w) = \gamma^{z(u,i,w)-1}$	HD, II-F, AI-F	[103, 234]
inverse*	$e_{\text{inv}}(p) = 1/p$	MME, IBO/IWO	[191]

*In the published version, the inverse examination function has been written as $e_{\text{inv}}(u,i) = 1/z(u,i)$. The correct formulation of the function depends directly on the recommendation slot at position p.

3.2.2 Joint Measures of Fairness and Relevance

We present measures that evaluate fairness considering relevance (Fair+Rel or joint measures henceforth). To our knowledge, we include all Fair+Rel measures for RSs published up to October 2023. Each measure uses an exposure function, which is linked to the fairness of item distribution in the recommendation and, therefore, measures item fairness jointly with relevance. We use ↑ for measures where the higher the score, the fairest the recommendation, and vice versa for ↓. All measures–except HD–are defined for multiple recommendation rounds or stochastic rankings, where a distribution over rankings is considered [19].

3.2.2.1 Inequity of Amortized Attention (IAA) [19]

↓IAA[1] measures fairness as the aggregated difference between item exposure and its relevance in a series of rankings that have been generated by a stochastic process [19]. The intuition behind IAA is that for a sequence of rankings to be fair, items should be allocated exposure proportional to their relevance to the user. The item position is a proxy of its exposure level. IAA was modified in [23] to account for multiple recommendation rounds (stochastic rankings):

$$\text{IAA} = \frac{1}{m}\sum_{u \in U} \text{IAA}(u) \tag{3.1}$$

[1]This measure is called IAA in [175] and L1-norm in [228].

$$\mathrm{IAA}(u) = \frac{1}{n}\frac{1}{W}\sum_{i\in I}\left|\sum_{w=1}^{W} 1_{L_{u,w}}(i)\cdot \tilde{e}_{\cdot}(u,i,w) - \tilde{r}(u,i,w)\right| \tag{3.2}$$

In [23], $\tilde{e}_{(\cdot)}(u,i,w)$ is the min-max normalised linear examination function $\tilde{e}_{\mathrm{li}}(u,i,w)$ (see Tab. 3.1) and $\tilde{r}(u,i,w) \in [0,1]$ is the min-max normalised relevance value of item i for user u in round w, $r_{u,i,w}$.[2] Both the min and max relevance values are taken from the values associated with all items for each user per round, i.e., $\min_{i\in I} r_{u,i,w}$. This value is the aggregated relevance over all items for user u in round w; the higher the relevance, the closer to 1. The higher the relevance value differs from item exposure, the more unfair. The range of IAA is $[0,1]$.

3.2.2.2 Individual Fairness Disparity (IFD) [161, 205]

↓IFD is the average pairwise difference of the combined value of item exposure and item merit. Merit is defined as a function of relevance.[3] Similar to IAA, IFD follows the principle of allocating exposure to an item based on its relevance. While IAA computes the difference between the exposure and relevance of each item, IFD computes the disparity of exposure allocation between item pairs. Based on how exposure and merit are combined, two variations of IFD exist: $\mathrm{IFD}_{\div}$, where item exposure is divided by item relevance [205], and $\mathrm{IFD}_{\times}$, where the division is replaced by multiplication [161]. The term $\mathrm{IFD}_{(\cdot)}$ or IFD refers to the measure in general. The two versions slightly differ in the pairwise difference computation, the formation of set of item pairs, and the exposure weighting scheme.[4] Both IFD versions have been used to measure fairness in ranking [161, 205, 242, 243].

$$\mathrm{IFD}_{(\cdot)} = \frac{1}{m}\sum_{u\in U}\mathrm{IFD}_{(\cdot)}(u) \tag{3.3}$$

$$\mathrm{IFD}_{\div}(u) = \frac{1}{|H_u|}\sum_{(i,i')\in H_u}\max\{0, J_{\div}(u,i) - J_{\div}(u,i')\} \tag{3.4}$$

$$\mathrm{IFD}_{\times}(u) = \frac{1}{n(n-1)}\sum_{i\in I}\sum_{i'\in I\setminus i}\left[J_{\times}(u,i) - J_{\times}(u,i')\right]^2 \tag{3.5}$$

$$J_{\div}(u,i) = \frac{1}{W}\sum_{w=1}^{W} e_{\mathrm{DCG}}(u,i,w)/r_{u,i,w} \tag{3.6}$$

[2]Note that the normalised exposure value for a recommended item at k is zero.
[3]We use the item relevance value as the item merit, as per [205].
[4]Exposure is weighed proportional to e_{DCG} in [205]; to simplify, we use e_{DCG} directly.

$$J_{\times}(u,i) = \frac{1}{W}\sum_{w=1}^{W} r_{u,i,w} \cdot 1_{L_{u,w}}(i) \cdot e_{\text{DCG}}(u,i,w) \tag{3.7}$$

$J_{(\cdot)}(u,i)$ is the function combining the expected exposure and relevance of item i for user u and $H_u = \{(i,i') \in I \mid r_{u,i} \geq r_{u,i'} > 0\}$. The range of IFD$_{\div}$ is $[0,\infty)$ and it is 0 only when the exposure received by each relevant item is exactly proportional to its relevance [205]. The range of IFD$_{\times}$ is $[0,\infty)$ based on empirical results [242].

3.2.2.3 Hellinger Distance (HD) [103]

↓HD has been used as a measure of individual item fairness in top k contextual bandits, by quantifying the difference between the relevance- and click-distributions of the top k items sorted according to (ground truth) relevance [103]. The click probability is based on user patience, system-allocated item exposure, and item relevance. A recommendation is fair based on HD when the click probability of an item is proportional to the relevance probability of that item. To compute the relevance and click distributions, a list of top k items is created for each user by sorting items based on their (ground truth) relevance; this list is the reference list used in the next step. Another list of items is created based on system prediction and used to get the click probability. For each item in the reference list, we compute its click probability based on its order in the second list. Next, the relevance probabilities of items at the same position in the reference list are aggregated across users and similarly for the click probabilities. For each rank position, two aggregated values are obtained: relevance and click. The aggregated values are the inputs to the distance metric (Eq. (3.8)).

$$\text{HD} = \frac{1}{\sqrt{2}}\sqrt{\sum_{p=1}^{k}\left(\sqrt{q'_p} - \sqrt{c'_p}\right)^2} \tag{3.8}$$

$$q'_p = \frac{1}{m}\sum_{u\in U}\sum_{i\in I}\delta\left(z^*(u,i) = p\right) \cdot r'_{u,i} \tag{3.9}$$

$$c'_p = \frac{1}{m}\sum_{u\in U}\frac{c^*_{u,p}}{\sum_{\ell=1}^{k} c^*_{u,\ell}} \tag{3.10}$$

$$c^*_{u,p} = \sum_{i\in I}\delta\left(z^*(u,i) = p\right) \cdot c^{full}_{u,i} \tag{3.11}$$

$$c^{full}_{u,i} = c'_{u,p} \text{ if } \exists p : z(u,i) = p\,, \text{ otherwise } 0 \tag{3.12}$$

$$c_{u,p} = \sum_{i \in L_u} \delta\left(z(u,i) = p\right) \cdot r_{u,i} \cdot \gamma \; e_{\text{RBP}}(u,i) \cdot s_{u,p} \tag{3.13}$$

$$s_{u,p} = \prod_{1 \leq j < p} 1 - \sum_{i \in L_u} \delta(z(u,i) = j) \cdot r_{u,i} \tag{3.14}$$

where q'_p and c'_p are the normalised relevance and click probability of the item at position j respectively, where click depends on both relevance and exposure. The position of item i based on ground-truth relevance is $z^*(u,i)$. The click probability of user u for item at position p, $c_{u,p}$ depends on $s_{u,p}$, the probability that items before position p were irrelevant to the user, and the user patience $\gamma \; e_{\text{RBP}}(u,i)$. $r'_{u,i} = r_{u,i} / \sum_{i \in I} r_{u,i}$ is the user-wise normalised relevance value of item i to user u, and $c'_{u,p} = c_{u,p} / \sum_{p=1}^{k} c_{u,p}$ is the user-wise normalised click probability. The value of $\delta(\cdot) = 1$ when the expression $\cdot$ is True and 0 otherwise. HD ranges between $[0, \infty)$.

3.2.2.4 Mean Max Envy (MME) [191]

↓MME uses the concept of envy-freeness, where a recommendation is fair when each item is not disadvantaged by its own exposure allocation compared to being allocated the exposure of any other item. In other words, MME computes unfairness as the disadvantage suffered by the item, if the exposure allocation of an item is swapped with another item. The disadvantage is computed based on an impact score that uses exposure and relevance: given full recommendation lists (size n) across all users, we swap each item i with another item i' and compute the impact score before and after the swap for all rank positions and users. If the score of the item i before the swap is greater or equal to its score after the swap, we have envy-freeness for item i w.r.t. item i'. MME thus computes the average maximum difference of impact imposed if item i is replaced with another item i'. E.g., let $L_{u_1} = [i_1, i_2]$, $L_{u_2} = [i_1, i_3]$ and let us swap item i_3 with i_1. Item i_3 will be exposed to both users at the top position, like i_1 did, and then impact is recomputed. MME is computed as follows:

$$\text{MME} = \frac{1}{n} \sum_{i \in I} \left\{ \max_{i' \in I} Imp_i(i') - Imp_i(i) \right\} \tag{3.15}$$

$$Imp_i(i') = \sum_{u \in U} \sum_{p=1}^{k} r_{u,i} \cdot e_{\text{inv}}(p) \cdot X_{u,i',p} \tag{3.16}$$

$$X_{u,i',p} = \frac{1}{W} \frac{1}{m} \sum_{w=1}^{W} 1_{L_{u,w}}(i') \cdot \delta(z(u,i',w) = p) \tag{3.17}$$

where $Imp_i(i')$ is the impact when we allocate the exposure of item i' to item i, $X_{u,i',p}$ is the probability that item i' is recommended to user u at position p in W rounds of recommendations, and $e_{\text{inv}}(p)$ is the exposure weight of item at position p, where the weight is discounted with the inverse examination function (see Tab. 3.1).[5] MME ranges within $[0, \infty)$.

3.2.2.5 Item Better-Off (IBO) & Item Worse-Off (IWO) [191]

↑IBO and ↓IWO use the principle of *dominance over uniform ranking*, where fairness means each item has a better impact (as defined in MME) under the current ranking policy, than if it were under the uniform random ranking policy π_{unif}, which samples all possible permutations of items uniformly at random. IBO/IWO measures the percentage of items for which our current ranking policy increases/decreases impact by at least 10% compared to π_{unif}[6]:

$$\text{IBO} = \frac{100}{|I^-|} \sum_{i \in I^-} \delta \left(Imp_i(i) \geq 1.1 \cdot Imp_i^{unif} \right) \tag{3.18}$$

$$\text{IWO} = \frac{100}{|I^-|} \sum_{i \in I^-} \delta \left(Imp_i(i) \leq 0.9 \cdot Imp_i^{unif} \right) \tag{3.19}$$

$$Imp_i^{unif} = \frac{1}{m} \frac{1}{n} \sum_{p=1}^{k} \frac{1}{p} \cdot \sum_{u \in U} r_{u,i} \tag{3.20}$$

where I^- is the set of items with at least one user that finds the item relevant. This ensures that the set of items that cause $\delta(\cdot) = 1$ in IBO is disjoint from that in IWO.[7] $Imp_i(i)$ is as per Eq. (3.16) and Imp_i^{unif} is the impact if item i is exposed according to π_{unif} using $e_{\text{inv}}(p)$ as examination function (see Tab. 3.1).[8] Note that the above definitions are modifications to the formulation of [191] to avoid computational issues that result in division by zero (undefinedness limitation [177]).[9] IBO/IWO ranges between $[0, 100]$.

[5]This has been changed from the original text "$e_{\text{inv}}(u, i)$ is the exposure weight of item i' to user u", as e_{inv} depends directly on p.

[6]In [191], 10% is hard coded, but this can be a variable. We also use 10%.

[7]We exclude items with no relevant users, as for these items $Imp_i(i) = Imp_i^{unif} = 0$, causing the same items being considered 'better-off' and 'worse-off' at the same time.

[8]This has been changed from the original text "$e_{\text{inv}}(u, i)$", as the function depends directly on position p.

[9]We move the divisor Imp_i^{unif} from the left-hand side to the right-hand side.

3.2.2.6 Individual-user-to-individual-item fairness (II-F) [234]

↓II-F was first defined by [56] to quantify unfairness as the disparity between system exposure and target exposure in individual queries and individual items. II-F was redefined by [234] for RSs as:

$$\text{II-F} = \frac{1}{m}\frac{1}{n}\sum_{u\in U}\sum_{i\in I}\left(E_{u,i} - E^*_{u,i}\right)^2 \tag{3.21}$$

$$E_{u,i} = \frac{1}{W}\sum_{w=1}^{W} 1_{L_{u,w}}(i)\cdot e_{\text{RBP}}(u,i,w) \tag{3.22}$$

$$E^*_{u,i} = \frac{r_{u,i}}{|R^*_u|}\cdot\frac{1-\gamma^{|R^*_u|}}{1-\gamma} \text{ if } |R^*_u| > 0\,, \text{ otherwise } 0 \tag{3.23}$$

where $E_{u,i}$ is the expected exposure of i to u as per a stochastic ranking policy. $E^*_{u,i}$ is the expected exposure of i to u as per an ideal stochastic ranking policy, which assumes that relevant items get equal expected exposure [56]. Thus, the recommendation is fair based on II-F if the system exposure matches the exposure allocated to items under an ideal ranking policy. The examination function based on RBP (see Tab. 3.1) is used in $E_{u,i}$ and the equation of $E^*_{u,i}$ is derived based on the same examination function [234]. $|R^*_u|$ is the number of relevant items for user u. II-F ranges between $[0,1]$.

3.2.2.7 All-users-to-individual-item fairness (AI-F) [234]

↓AI-F evaluates how much RSs under/overexpose an item to all users as the mean deviation of overall system exposure over target exposure:

$$\text{AI-F} = \frac{1}{n}\sum_{i\in I}\left(\frac{1}{m}\sum_{u\in U}E_{u,i} - \frac{1}{m}\sum_{u\in U}E^*_{u,i}\right)^2 \tag{3.24}$$

where $E_{u,i}$, $E^*_{u,i}$ are as per Eq. (3.22)–(3.23). Similar to II-F, AI-F also quantifies fairness based on how close the system exposure is to the target exposure. In II-F, this disparity is computed individually between each user-item pair, while in AI-F item exposure is first aggregated across users prior to computing the difference in exposure. Due to this aggregation, AI-F would have a better fairness score than II-F when there is a greater number of unique items in the recommendation, as opposed to having the same few items exposed to all users. The range of AI-F is $[0,1]$.

Table 3.2: Statistics of the preprocessed datasets.

dataset	#users (m)	#items (n)	#interactions	sparsity (%)
Lastfm [32]	1,859	2,823	71,355	98.64%
Amazon-lb [156]	1,054	791	12,397	98.51%
QK-video [247]	4,656	6,423	51,777	99.83%
ML-10M [94]	49,378	9,821	5,362,685	98.89%

3.3 Experimental Setup

We study the above Fair+Rel measures across different recommenders and datasets. Our general experimental setup follows, and we provide the description of the experiments in §3.4.[10]

Datasets. We use four real-world datasets of varying sizes and domains: Lastfm (music) [32], Amazon Luxury Beauty, i.e., Amazon-lb (e-commerce) [156], QK-video (videos) [247], and ML-10M (movies) [94]. QK-video is as provided by [247], and the rest are as provided by [260]. For QK-video, we use only the 'sharing' interactions.

Preprocessing. We keep only users and items with at least 5 interactions (5-core filtering). When there are duplicate interactions, we keep the most recent one. Ratings equal/above 3 are converted to 1, and the rest are discarded for Amazon-lb and ML-10M, as their ratings range between $[1, 5]$ and $[0.5, 5]$ respectively. No conversions are done for Lastfm and QK-video as they only have implicit feedback. Tab. 3.2 presents statistics of the preprocessed datasets.

Data splits. Global temporal splits [151] with a ratio of 6:2:2 form the train/-val/test sets from the preprocessed datasets for Amazon-lb and ML-10M. Global random splits with the same ratio are used for Lastfm and QK-video as they have no timestamps. Only users with at least 5 interactions in the train set are kept in all splits.

Recommenders. We use four well-known top k recommenders: item-based K-Nearest Neighbour (ItemKNN) [55], Bayesian Personalised Ranking (BPR), [185], Variational Autoencoder with multinomial likelihood (MultiVAE) [135], and Neighbourhood-enriched Contrastive Learning (NCL) [139]. We train BPR, MultiVAE, and NCL using RecBole [260] for 300 epochs with early stopping. The configuration with the best NDCG@10 during validation is taken as the final model.[11] During

[10] Our code: https://github.com/theresiavr/can-we-trust-recsys-fairness-evaluation.

[11] The hyperparameter search space and best values are in the code repository.

testing, all unobserved items are selected as candidates for recommendation and each user's train/val items are excluded from their own recommendations.

Fair re-ranker. As the models are not directly optimised for fairness, we use a re-ranker to obtain fairer recommendations. The top k' items are re-ranked to provide exposure to items that were outside the top k, where k' is ideally larger than the cut-off $k = 10$. In RS datasets, normally there are very few relevant items per user, so k' should not be too big (e.g., 100). We choose $k' = 25$ for all datasets and models. The re-ranking is done per user with COMBMNZ (CM) [125] as a robust rank fusion method.[12] CM fuses two lists of scores, one based on relevance and one based on fairness, to create a new ranking for each user. The relevance-based score is the min-max normalised predicted relevance score. The fairness-based score is first obtained from the coverage score of each top k' items based on their appearance in the top k. Then, we compute 1 minus the normalised coverage to allocate higher score for items with lower exposure, thus increasing fairness. The combined scores are sorted to generate the final fused ranking of relevance and fairness.

Measures. Recommendation models are evaluated using all the joint measures of relevance and fairness (Fair+Rel) presented in §3.2.2.[13] As comparison to the joint measures, we evaluate relevance only (Rel) with: Hit Rate (HR), MRR, Precision (P), Recall (R), MAP, and NDCG. We also evaluate fairness only (Fair) with:[14] Jain Index (Jain) [101, 264], Qualification Fairness (QF) [264], Entropy (Ent) [167, 200], Fraction of Satisfied Items (FSat) [167], and Gini Index (Gini) [85, 147]. Unless otherwise stated, all measures are computed at $k = 10$.

3.4 Empirical Analysis

We present the evaluation results of all Fair+Rel, Rel, and Fair measures, in §3.4.1. We study their correlation in §3.4.2, their sensitivity across different top-k positions in §3.4.3 and across increasing levels of relevance and fairness in §3.4.4.

[12]Other re-rankers exist but do not suit our setup, e.g., [227] requires computing item similarity, but true similarity is challenging to obtain [63, 216].

[13]For IAA, the ground truth relevance is used to compute the relevance score. For HD, $\gamma = 0.9$ as per [103]. For II-F and AI-F, $\gamma = 0.8$ as per [234]. Note that IBO/IWO are normalised to [0,1] for consistency with the other measures.

[14]We use the modified versions of these measures as per [177].

3.4.1 Evaluation Results of All Measures

Tab. 3.3 shows the scores of all Fair+Rel, Rel, and Fair measures, per dataset and recommender/re-ranking. ↑ means the higher the score, the better, and vice versa for ↓. Overall we observe the following.

Best model agreement. We aim to study whether the measures agree on the same best model. We note two main trends. First, for all datasets, the best model based on Rel measures is always different from the one based on Fair measures, except for QF in Amazon-lb. This means that the fairest model is not necessarily the best in terms of relevance. Second, while all Rel measures agree on the same best model per dataset (except MRR and MAP for Amazon-lb) and all the Fair measures always agree on the same best model (except QF), the Fair+Rel measures disagree on the best model. Occasionally, some Fair+Rel measures agree with another more often (e.g., IBO with IWO, or IAA with HD and II-F, or MME with AI-F and sometimes IFD), but there is no overall consistency. The agreement between some joint measures may be due to their similar formulations: both IBO/IWO are the fractions of items with an impact score greater/lower than a threshold; MME/AI-F aggregate exposure across users prior to computing the exposure difference, while IAA/HD/II-F do not; and MME/IFD are pairwise measures.

Range of scores. We identify three issues on the score range of the Fair+Rel measures: (1) extremely small scales for several joint measures; (2) scale mismatch between single-aspect measures and joint measures; and (3) scale mismatch between joint measures. About (1), for all datasets and models, several ↓Fair+Rel scores are extremely small ($\leq 10^{-3}$), and these scores do not allow to distinguish across models per dataset. For example, $\text{IFD}_\times$ is always close to 0 across all datasets, as the term Eq. (3.7) is often 0 due to the low number of relevant items per user.[15] For MME and II-F/AI-F, Eq. (3.16) and Eq. (3.23) often result in 0 for the same reason as $\text{IFD}_\times$. About (2), while the above Fair+Rel scores differ in the fourth or later decimal point, the differences in the Rel and Fair scores are in the second decimal point or before. E.g., the NDCG (Rel score) of MultiVAE-CM and NCL for Lastfm differs by ∼0.16 and their Jain (Fair score) differs by ∼0.14. These examples imply non-negligible differences, but the joint scores of IAA/$\text{IFD}_\times$/MME/II-F/AI-F only differ by $\leq 10^{-3}$, which may seem negligible.[16] These inconsistencies in the difference of magnitude make the scores hard to understand. About (3), we see large gaps in the score range of all joint measures, e.g., between IWO, HD, and AI-F, despite all of

[15]For all four datasets, the median number of relevant items per user is at most 46.

[16]We use NDCG and Jain as they tend to be stricter* than some other single-aspect measures used in this work. *Note that this footnote has been changed from the original text "... as they are more sensitive to changes than HR and QF" for clarity.

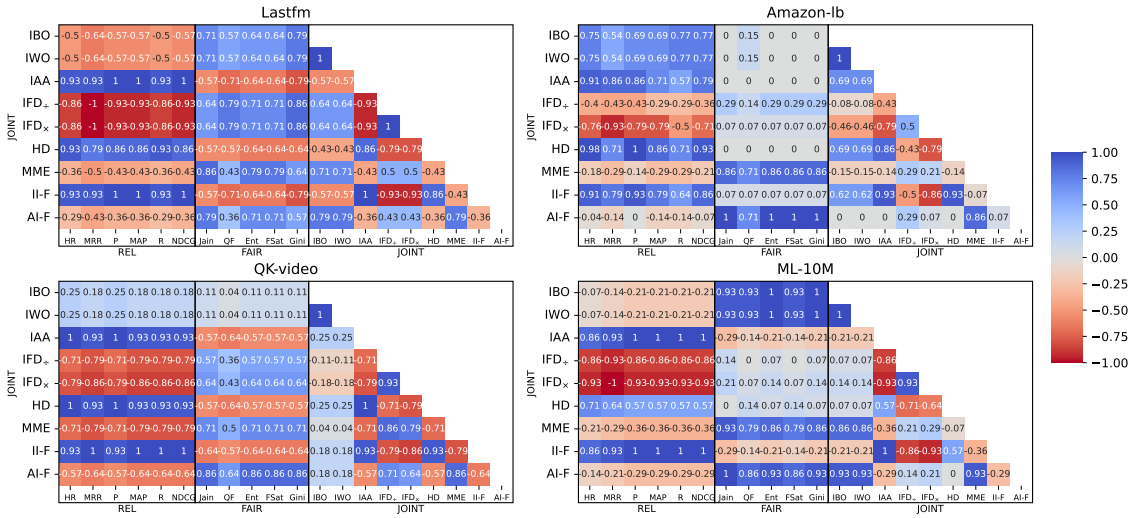


Figure 3.1: Kendall's τ correlation between joint Fair+Rel measures, Rel, and Fair measures.

them being lower-is-better measures. E.g., in ML-10M, ↓IWO ≈ 1 (very unfair) based on its theoretical $[0, 1]$-range, ↓HD is about a quarter of the ↓IWO score (somewhat fair), while ↓AI-F ≈ 0 (extremely fair). This discrepancy causes confusion in score interpretation.

Finally, we group all Fair+Rel measures into 3 clusters: (i) IAA/HD/II-F, which align more with Rel measures; (ii) IFD/MME/AI-F, which align more with Fair measures; and (iii) IBO/IWO, which do not consistently align with any single-aspect measure. Within the same cluster, especially in (i), measures often have large differences in their score ranges (up to $\Delta \approx 0.7$).

3.4.2 Correlation between Measures (RQ1 & RQ2)

We compute Kendall's τ correlation between the orderings of the recommenders produced by the scores of each measure, to study how much the Fair+Rel measures agree among themselves but also with Rel-only and Fair-only measures, when ranking recommenders (Fig. 3.1). As Rel and Fair measures do not always correlate strongly with each other [177], we do not expect Fair+Rel measures to correlate strongly with either Rel or Fair measures.

RQ1. Agreement of joint and single-aspect measures. Overall, there is no consistent correlation between Rel and Fair+Rel measures. IBO/IWO's correlations vary wildly ($\tau \in [-0.64, 0.77]$); IAA, HD, and II-F have moderate-to-strong

Table 3.3: Relevance (Rel), fairness (Fair), and joint Fair+Rel scores at $k = 10$ without and with re-ranking the top $k' = 25$ items using COMBMNZ (CM). Bold marks the most relevant/fair score per measure. The score 0.000 does not mean the scores are exactly 0; this is due to the measures having small scores ($< 10^{-3}$ and rounding to 3 d.p.

		model	ItemKNN		BPR		MultiVAE		NCL	
		re-ranker	-	CM	-	CM	-	CM	-	CM
Lastfm	Rel	↑ HR	0.765	0.581	0.773	0.587	0.778	0.523	**0.793**	0.571
		↑ MRR	0.484	0.270	0.492	0.280	0.476	0.232	**0.503**	0.260
		↑ P	0.172	0.089	0.178	0.092	0.176	0.076	**0.184**	0.087
		↑ MAP	0.137	0.053	0.141	0.058	0.138	0.045	**0.148**	0.050
		↑ R	0.218	0.114	0.224	0.119	0.224	0.098	**0.234**	0.110
		↑ NDCG	0.245	0.119	0.252	0.126	0.247	0.102	**0.261**	0.115
	Fair	↑ Jain	0.042	0.094	0.058	0.140	0.097	**0.222**	0.082	0.215
		↑ QF	0.474	**0.679**	0.362	0.528	0.517	0.678	0.453	0.657
		↑ Ent	0.589	0.735	0.610	0.740	0.707	**0.826**	0.671	0.810
		↑ FSat	0.129	0.216	0.147	0.228	0.202	**0.321**	0.178	0.286
		↓ Gini	0.904	0.790	0.910	0.818	0.839	**0.696**	0.872	0.728
	Fair+Rel	↑ IBO	0.209	0.256	0.208	0.253	0.261	0.278	0.242	**0.292**
		↓ IWO	0.791	0.744	0.792	0.747	0.739	0.722	0.758	**0.708**
		↓ IAA	0.004	0.004	0.004	0.004	0.004	0.004	**0.004**	0.004
		↓ $\text{IFD}_{\div}$	0.074	0.053	0.075	0.054	0.073	**0.049**	0.076	0.052
		↓ $\text{IFD}_{\times}$	0.000	0.000	0.000	0.000	0.000	**0.000**	0.000	0.000
		↓ HD	0.099	0.177	0.104	0.174	0.095	0.203	**0.092**	0.177
		↓ MME	0.001	0.001	0.001	0.001	0.001	0.001	0.001	**0.001**
		↓ II-F	0.001	0.002	0.001	0.002	0.001	0.002	**0.001**	0.002
		↓ AI-F	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.000**
Amazon-lb	Rel	↑ HR	**0.046**	0.016	0.011	0.021	0.039	0.014	0.034	0.011
		↑ MRR	0.020	0.011	0.003	0.007	**0.023**	0.004	0.022	0.003
		↑ P	**0.005**	0.002	0.001	0.002	0.004	0.002	0.004	0.001
		↑ MAP	0.006	0.004	0.002	0.004	0.006	0.003	**0.006**	0.001
		↑ R	**0.013**	0.005	0.005	0.010	0.010	0.008	0.012	0.003
		↑ NDCG	**0.011**	0.005	0.003	0.006	0.010	0.004	0.011	0.002
	Fair	↑ Jain	0.271	**0.431**	0.223	0.359	0.035	0.097	0.026	0.080
		↑ QF	**0.650**	0.612	0.549	0.594	0.222	0.286	0.229	0.310
		↑ Ent	0.802	**0.839**	0.747	0.809	0.418	0.558	0.371	0.534
		↑ FSat	0.370	**0.438**	0.314	0.376	0.114	0.152	0.091	0.138
		↓ Gini	0.665	**0.598**	0.747	0.660	0.949	0.899	0.959	0.910
	Fair+Rel	↑ IBO	**0.062**	0.029	0.019	0.038	0.029	0.029	0.038	0.024
		↓ IWO	**0.938**	0.971	0.981	0.962	0.971	0.971	0.962	0.976
		↓ IAA	**0.011**	0.011	0.011	0.011	0.011	0.011	0.011	0.011
		↓ $\text{IFD}_{\div}$	0.005	0.003	0.003	**0.002**	0.005	0.002	0.005	0.003
		↓ $\text{IFD}_{\times}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.000**
		↓ HD	**0.580**	0.630	0.661	0.626	0.597	0.653	0.598	0.667
		↓ MME	0.001	**0.001**	0.001	0.001	0.003	0.001	0.004	0.001
		↓ II-F	**0.006**	0.006	0.006	0.006	0.006	0.006	0.006	0.006
		↓ AI-F	0.000	**0.000**	0.000	0.000	0.001	0.000	0.002	0.000

Table 3.3 (continued): (REL), fairness (FAIR), and joint FAIR+REL scores at $k = 10$ without and with re-ranking the top $k' = 25$ items using COMBMNZ (CM). Bold marks the most relevant/fair score per measure. The score 0.000 does not mean the scores are exactly 0; this is due to the measures having small scores ($< 10^{-3}$ and rounding to 3 d.p.

		model	ItemKNN		BPR		MultiVAE		NCL	
		re-ranker	-	CM	-	CM	-	CM	-	CM
QK-video	REL	↑ HR	0.040	0.047	0.099	0.045	0.109	0.061	**0.130**	0.077
		↑ MRR	0.013	0.013	0.039	0.015	0.039	0.021	**0.048**	0.024
		↑ P	0.004	0.005	0.011	0.005	0.012	0.006	**0.014**	0.008
		↑ MAP	0.005	0.005	0.017	0.006	0.018	0.009	**0.022**	0.010
		↑ R	0.014	0.019	0.043	0.019	0.051	0.027	**0.061**	0.033
		↑ NDCG	0.009	0.010	0.029	0.011	0.031	0.016	**0.038**	0.019
	FAIR	↑ Jain	0.483	**0.589**	0.081	0.379	0.012	0.032	0.020	0.071
		↑ QF	**0.901**	0.790	0.625	0.823	0.100	0.163	0.201	0.365
		↑ Ent	0.933	**0.937**	0.755	0.903	0.420	0.547	0.507	0.674
		↑ FSat	0.443	**0.547**	0.212	0.382	0.052	0.090	0.077	0.150
		↓ Gini	0.472	**0.442**	0.807	0.570	0.982	0.959	0.966	0.902
	FAIR+REL	↑ IBO	0.033	0.038	**0.054**	0.036	0.031	0.036	0.043	**0.054**
		↓ IWO	0.967	0.962	**0.946**	0.964	0.969	0.964	0.957	**0.946**
		↓ IAA	0.001	0.001	0.001	0.001	0.001	0.001	**0.001**	0.001
		↓ IFD$_{\div}$	0.009	**0.007**	0.014	0.008	0.014	0.009	0.015	0.010
		↓ IFD$_{\times}$	0.000	**0.000**	0.000	0.000	0.000	0.000	0.000	0.000
		↓ HD	0.576	0.560	0.490	0.565	0.478	0.535	**0.457**	0.519
		↓ MME	0.000	**0.000**	0.000	0.000	0.000	0.000	0.000	0.000
		↓ II-F	0.001	0.001	0.001	0.001	0.001	0.001	**0.001**	0.001
		↓ AI-F	0.000	**0.000**	0.000	0.000	0.000	0.000	0.000	0.000
ML-10M	REL	↑ HR	0.487	0.443	0.512	0.386	0.417	0.387	**0.521**	0.402
		↑ MRR	0.282	0.225	0.299	0.185	0.237	0.191	**0.302**	0.203
		↑ P	0.137	0.105	0.146	0.088	0.107	0.096	**0.154**	0.094
		↑ MAP	0.089	0.060	0.095	0.047	0.067	0.054	**0.101**	0.052
		↑ R	0.022	0.018	0.025	0.012	0.020	0.016	**0.026**	0.013
		↑ NDCG	0.150	0.113	0.160	0.092	0.119	0.100	**0.167**	0.100
	FAIR	↑ Jain	0.011	0.027	0.037	**0.115**	0.003	0.006	0.024	0.069
		↑ QF*	0.044	0.068	0.145	**0.216**	0.014	0.025	0.086	0.132
		↑ Ent	0.407	0.514	0.596	**0.716**	0.238	0.324	0.519	0.638
		↑ FSat*	0.044	0.068	0.145	**0.216**	0.014	0.025	0.086	0.132
		↓ Gini	0.987	0.971	0.945	**0.879**	0.997	0.993	0.969	0.930
	FAIR+REL	↑ IBO	0.031	0.046	0.069	**0.091**	0.012	0.018	0.054	0.074
		↓ IWO	0.969	0.954	0.931	**0.909**	0.988	0.982	0.946	0.926
		↓ IAA	0.008	0.009	0.008	0.009	0.009	0.009	**0.008**	0.009
		↓ IFD$_{\div}$	0.018	0.012	0.019	0.011	0.016	**0.010**	0.020	0.012
		↓ IFD$_{\times}$	0.000	0.000	0.000	**0.000**	0.000	0.000	0.000	0.000
		↓ HD	0.221	0.255	0.226	0.262	0.265	0.273	**0.218**	0.257
		↓ MME	0.001	0.001	0.001	**0.001**	0.003	0.001	0.001	0.001
		↓ II-F	0.000	0.000	0.000	0.000	0.000	0.000	**0.000**	0.000
		↓ AI-F	0.000	0.000	0.000	**0.000**	0.000	0.000	0.000	0.000

*For ML-10M, QF ≡ FSat, as QF is computed based on the % of recommended items from all items, which in this case is equivalent to FSat.

positive correlations ($\tau \in [0.57, 1]$); IFD and MME have weak-to-strong negative correlations ($\tau \in [-1, -0.29]$ for IFD and $\tau \in [-0.79, -0.14]$ for MME); and AI-F has non-positive correlations ($\tau \in [-0.64, 0]$).

The correlations between Fair and Fair+Rel measures are inconsistent. The correlations of IBO/IWO vary largely again, albeit less than with Rel measures. IAA/HD/II-F have two distinct trends across groups of datasets: they have negative moderate-to-strong correlations ($\tau \in [-0.79, -0.57]$) for Lastfm and QK-video, but weak correlations for Amazon-lb and ML-10M ($\tau \in [-0.29, 0.14]$). Similarly, IFD has high correlations for Lastfm and QK-video (except with QF for QK-video), $\tau \in [0.57, 0.86]$, and weak or zero correlations for the other datasets ($\tau \in [0, 0.29]$). Conversely, MME and AI-F have strong correlations except with QF for Lastfm ($\tau \in [0.5, 1]$).

Note that Fair+Rel measures strongly agreeing with Rel measures do not always strongly disagree with Fair measures, and vice versa. E.g., IAA/HD/II-F strongly correlates with Rel measures for Amazon-lb, but they correlate weakly with Fair measures.

RQ2. Agreement between joint measures. Overall we find that the three clusters of joint measures identified in §3.4.1 show strong positive correlations between measures inside the same cluster and strong negative correlations between measures from different clusters. E.g., IBO always perfectly correlates with IWO, due to their similar formulation. IAA, HD, and II-F agree strongly with one another, $\tau \in [0.57, 1]$. $\text{IFD}_{\div}$ correlates highly with $\text{IFD}_{\times}$, $\tau \in [0.5, 1]$, as their formulations are similar. MME always agrees strongly with AI-F, $\tau \in [0.79, 0.93]$. IFD sometimes has moderate-to-strong correlations with MME and AI-F, $\tau \in [0.43, 0.86]$ for Lastfm and QK-video, but the correlations are weaker for Amazon-lb and ML-10M, $\tau \in [0.07, 0.29]$. In contrast, IAA/HD/II-F strongly disagrees with IFD, $\tau \in [-0.93, -0.5]$ except for $\text{IFD}_{\div}$ in Amazon-lb ($\tau = -0.43$).[17]

Based on the above, we conclude that: IBO/IWO has inconsistent relationships with single-aspect and joint measures; IAA/HD/II-F do not align with fairness; and IFD/MME/AI-F highly disagree with relevance (even if IFD sometimes disagrees with Fair measures too). Among the joint measures, IBO/IWO weakly correlate with the single-aspect measures for QK-video, and similarly with $\text{IFD}_{\div}$ for Amazon-lb, but this is not consistent. We thus argue that no joint measures reliably account for both relevance and fairness.

[17]The published version reported " $\tau \in [-0.71, -0.5]$"; the lower bound has been corrected and the conclusion remains the same.

3.4.3 Measure Sensitivity at Different Ranks (RQ3)

We now study how sensitive the joint measures are at decreasing rank positions, compared to Rel and Fair measures. When moving down the rank, Rel scores are known to decrease while Fair scores are known to improve [177]. For this analysis, we use only the runs of the non-reranked NCL model as it generally has the best Rel scores. We compute all measures at $k = 5$ for each sliding window, where the windows consist of items at decreasing rank positions: 1–5, 2–6, . . . , 5–9. Fig. 3.2 shows the results for Lastfm, Amazon-lb, and ML-10M, which represent the overall trends in all our datasets; results for QK-video are shown in the appendix (in our code repository).

We find that, as expected, as we move down the rank, Rel overall decreases and Fair improves. However, the joint measures are notably less sensitive to changes in rank position. Changes with decreasing rank position in the single-aspect scores are up two magnitudes greater than in the joint measures, and the latter do not reflect these differences to the same scale. We posit that the insensitivity is due to the effect of changing relevance being masked by that of fairness and vice versa. This masking makes the scores hard to interpret. Further, the very small scores of ↓IAA, $\text{IFD}_\times$, MME, II-F, and AI-F imply extremely fair recommendations (we explain the reasons for this in §3.4.1), even if Rel and Fair scores are low. Thus, these joint measures do not account well for relevance and fairness simultaneously. Last, we note that as we move down the rank, IAA/HD/II-F worsen, IFD/MME/AI-F improve, and IBO/IWO are inconsistent across datasets. This follows the three groups of joint measures discussed above.

3.4.4 Artificial Insertion of Items (RQ4)

Lastly, we study how sensitive the joint measures are to different proportions of relevant items and item fairness in the ranking. Assessing this sensitivity is important as it affects score interpretation; if a joint measure is unresponsive to significant changes in relevance and fairness distribution, its score may not reflect both the fairness and relevance of the recommendations accurately.

We start with a recommendation list having the worst Rel and Fair scores, and gradually insert more relevant and fair items to it (we explain 'fair items' below). We observe how the joint measures respond to these changes, compared to the Rel and Fair measures.

We cannot use real-life datasets for this analysis, so we build a synthetic dataset with $m = 1000$ and $n = 10000$, and artificially generate rankings of items per user, as per [177]. The initial ranking contains the same $k = 10$ items for all users, to

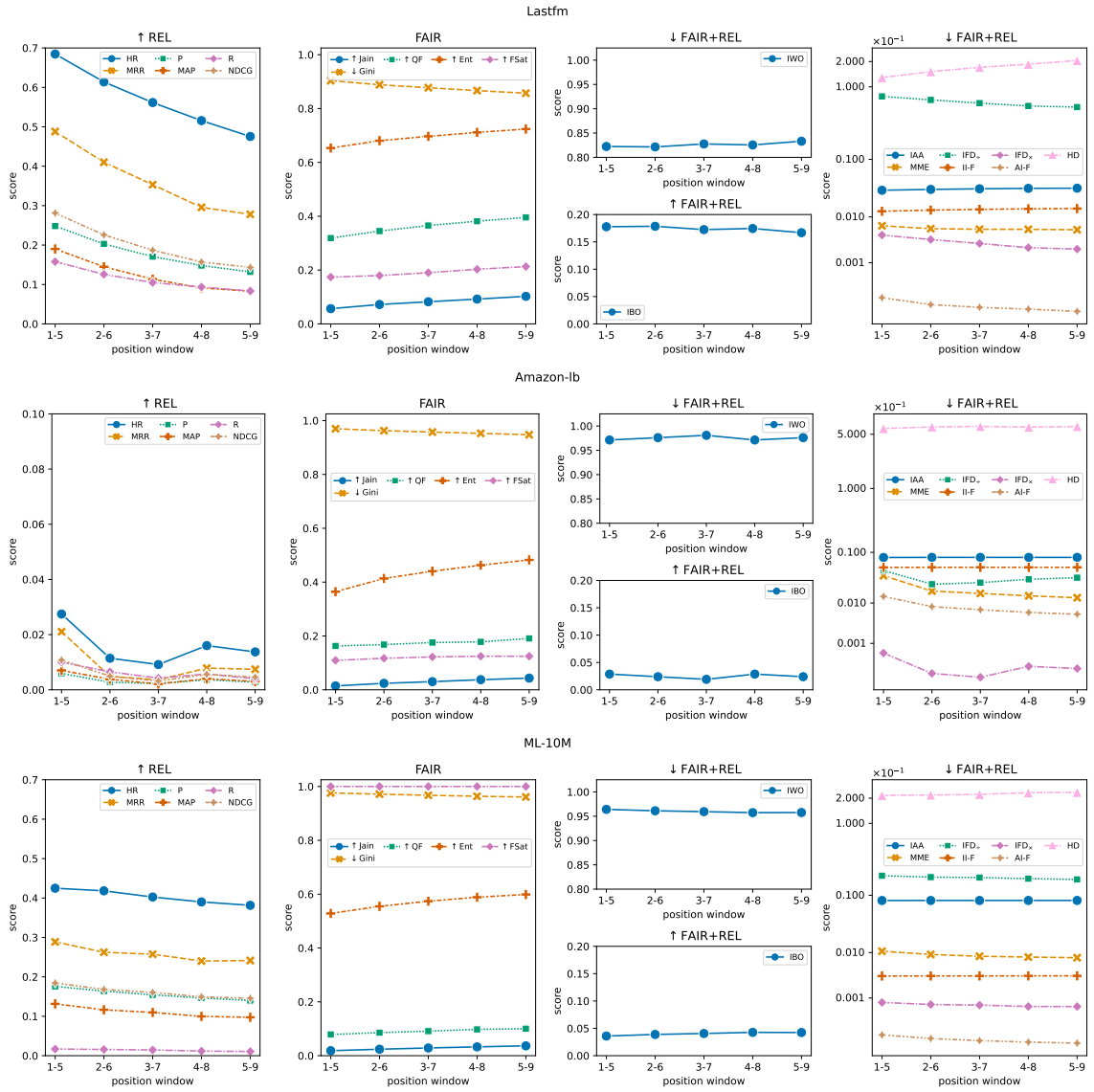


Figure 3.2: Sliding window evaluation ($k = 5$) of NCL for Lastfm, Amazon-lb, and ML-10M. The last column is in exponential scale.

whom these items are irrelevant, except for one user.[18] In each iteration, an item from the bottom of each user's top k ranking is replaced by a relevant item having less exposure (hence more fair). The final ranking thus contains km unique items across all users, where each item is relevant only to the user that receives that item in the top k. We expect all measures to initially score the worst possible, and then gradually improve as more relevant and fair items enter the ranks.

Fig. 3.3 shows the results of this analysis. Overall, we see that most joint measures are not very sensitive to changes in Rel and Fair scores, i.e., they may vary, but negligibly. This verifies the scale mismatch between most joint measures and the single-aspect measures observed in §3.4.1 & §3.4.3. While the overall change is negligible for most measures, a common observation between the joint measures is that their scores become (slightly) better as Rel and Fair scores improve. An exception to this is IFD. This is because IFD measures fairness based on the pairwise difference in the combined value of exposure and relevance. Thus, when the relevant items start to be moved to the top k, the gap between the exposure weight of relevant items in and outside the top k increases, and so does unfairness. Among joint measures that (slightly) improve with more insertion, there are also differences. IBO/IWO improve linearly; as both measures are percentages of items, the change is proportional to the amount of inserted items. HD also improves, but its improvement fluctuates due to randomness introduced by the unstable sort in the computation, as per the original implementation in [103]. ↓IAA/IFD$_\times$/MME/II-F/AI-F improve non-linearly. However, their scores are extremely close to 0, i.e., on the scale of 10^{-3} or less. The lower bound of the measure is 0, hence these small scores indicate that the recommendation is close to the fairest, even at the start of the process where the Rel and Fair scores are the worst in the entire progression. These joint measures are also rather insensitive to changes in Rel and Fair scores. Here, their score range is (0, 0.0015), while the range of Rel, Fair, and IBO/IWO scores is [0,1].

3.5 Related Work

Fairness evaluation in RSs. Among prior work on fairness evaluation in RSs [9, 175, 177, 228, 252], our study is close to Amigó et al. [9], who study RS relevance and fairness for groups/individuals and between items and users. Yet, the focus of our work, i.e., individual item fairness, is not covered in [9]. Raj and Ekstrand [175] overview evaluation measures for item group fairness. Their study includes the IAA measure as a group fairness measure (whereas we focus on individual item fairness).

[18] This is to keep the number of items exactly km.

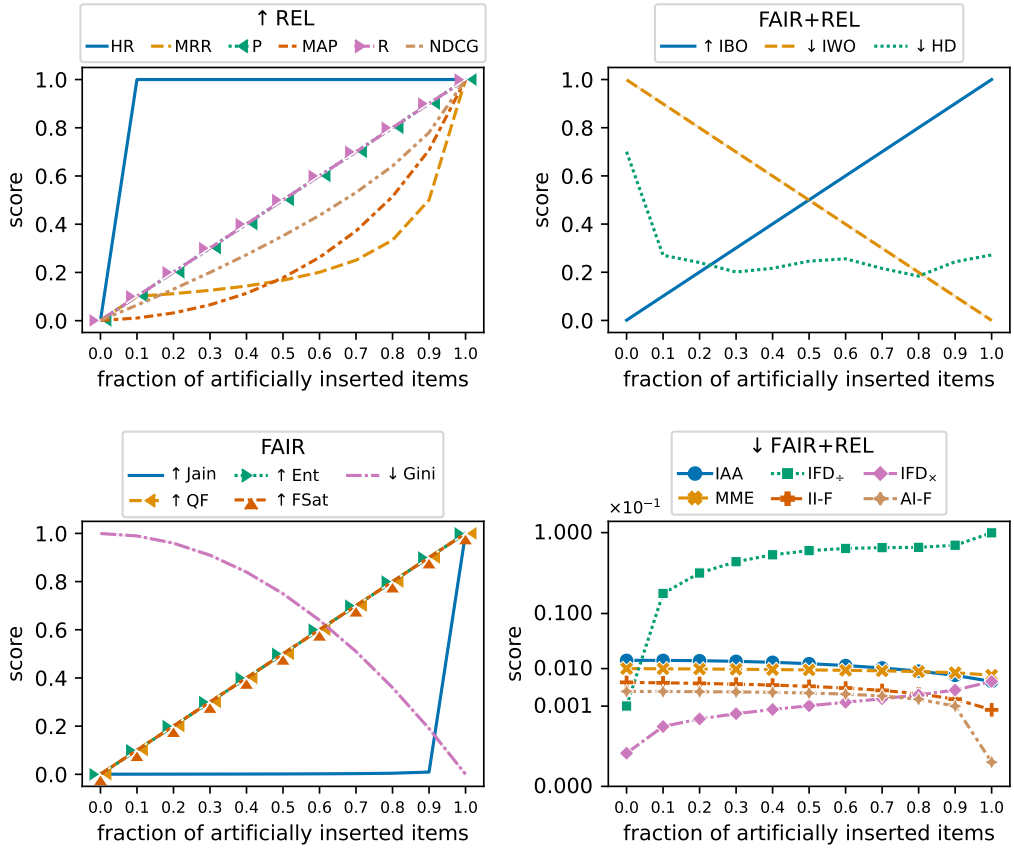


Figure 3.3: Artificial insertion of items with $m = 1000$ (users).

Lastly, Rampisela et al. [177] survey individual item fairness measures that are exclusively linked with fairness and identify the limitations within them, while we focus on measures that jointly account for both fairness and relevance.

Joint measures of relevance and fairness. Outside the strict domain of individual item fairness for RS, there exist other measures that quantify relevance and fairness jointly: Gao et al. [79] present a measure combining KL-divergence and IDCG to jointly quantify relevance and group fairness in IR [79]. In [239], utility and provider fairness in RSs are simultaneously evaluated with a weighted sum between relevance and fairness. Another approach used in [81] to evaluate individual fairness in ranking is to compare item position based on ground truth relevance against its position in system-produced rankings. None of the joint measures in our work is a combination of two single-aspect measures as in [79] or in the form of weighted sum as in [239]. The measure in [81] is similar to HD [103]. However, we do not use it in our work because it was not defined for RS fairness, and considerable modifications and assumptions are required prior to using it to evaluate RS fairness.

3.6 Appropriate Usage of Joint Measures

We find that joint measures of relevance and fairness (1) tend to align differently with single-aspect measures; (2) most of them consistently score almost perfect fairness, even when recommendations are highly irrelevant and unfair based on single-aspect measures; and (3) are rather unresponsive to changes in the recommendation relevance and fairness, especially compared to single-aspect measures. Next, we suggest how to best use these joint measures.

Avoid using similar joint measures. In §3.4.2 & §3.4.3 we find three groups of similar joint measures: (i) IAA/HD/II-F, (ii) IFD/MME/AI-F, and (iii) IBO/IWO. Only one measure per group should be used. Yet, considering that typically recommendations are evaluated with Rel measures, we discourage using measures in (i), as they are highly aligned with Rel measures. Measures in (ii) correlate strongly with Fair measures, and can be viable options, and likewise for measures in (iii) that do not consistently correlate with single-aspect measures. However, we argue that measures in (iii) are more useful than those in (ii). Measures in (ii) can be replaced by Fair measures, which are faster to compute and do not need complete relevance judgements, while still achieving highly similar conclusions.

Be aware of the unintuitive or inconsistent behaviour, insensitivity, and computational complexity of the measures. We recommend that practitioners be aware of the measure limitations in groups (ii) and (iii). Specifically, both IFD

versions worsen with higher percentages of jointly relevant and fair recommendations, while the opposite should happen in a healthy measure. $\text{IFD}_{\div}$ is unaffected by different cut-off k values, as its original formulation only considers full rankings, so it should not be used when different k matters. MME is costly to compute as it is a pairwise measure (∼30 mins for the larger datasets), and the same applies to $\downarrow\text{IFD}_{\times}$, albeit to a lesser extent. Further, $\text{IFD}_{\times}$/MME/AI-F tend to have extremely small scores, which are therefore hard to interpret and discriminate across runs. They are also rather insensitive to changes reflected by single-aspect measures, meaning that their overall expressiveness is limited. IBO/IWO is sensitive in this aspect. Considering the limitations and the redundancy between measures, IBO/IWO seem to be the most viable measure out of the existing ones, but it is the least consistent between all other measures, due to varying alignments for different datasets, so it should be interpreted cautiously.

Avoid score misinterpretation in measures with small empirical scales. Due to the small empirical scales of ↓IAA/IFD/MME/II-F/AI-F, their scores tend not to represent fairness, or relevance and fairness jointly, i.e., scoring very close to 0 even if the recommendation is very irrelevant and unfair based on Rel and Fair measures. Moreover, different systems can have very similar scores which do not translate to similar performance. E.g., two models differing in scores by only 0.001 can be interpreted as performing the same, even though the measure has a small empirical range and is not very sensitive to begin with. This issue can be fixed via apriori/posthoc normalisation based on experimental values of the measures [234].

Measure fairness separately from relevance. As most joint measures (IAA/IFD /MME/II-F/AI-F) are difficult to interpret because their scores tend to be compressed in a very low range, and are also rather insensitive to changes in fairness and relevance, we recommend measuring individual item fairness and relevance separately. Otherwise, the joint scores can be close to the theoretical fairest value even if Rel and Fair scores are low (§3.4.4). Overall, the above joint measures have unreliable scores, are not as sensitive as the Fair measures, and are subject to more under/overestimation of fairness than Fair measures which have more consistent empirical range. The remaining joint measures are not reliable either: IBO/IWO aligns inconsistently to the single-aspect measures, while HD is almost always consistent with Rel measures and thus does not add another dimension of fairness measurement. It is also unstable due to sorting of items with identical relevance level.

Overall, the joint measures cannot be compared easily as they have different scales, and they quantify two aspects that are hard to combine due to mismatching scales. The measures tend to correlate highly with either Rel or Fair measures,

instead of having a good balance between them. As such, optimising for a joint measure directly may not result in a simultaneously optimal recommendation based on Rel and Fair scores. Another obstacle in measuring fairness is the need to consider user-item relevance in the entire dataset (not just the recommended items), which can be an issue with extremely sparse datasets. It is thus inherently difficult to devise a measure that can jointly quantify relevance and fairness.

3.7 Conclusions and Future Work

We presented a novel empirical study on the properties of all evaluation measures that jointly account for individual item fairness and relevance in recommender systems. We found that out of 9 joint measures, 3 align with traditional relevance-only measures, 4 agree more with fairness-only measures, and the rest behave inconsistently. We also found that only a few joint measures are sensitive to a simultaneous decrease in relevance and increased fairness in the recommendation. Most surprisingly, nearly all joint measures are almost unresponsive to increases in relevance and fairness. Even worse, the majority tend to compress scores at the low end of their range, giving the illusion of an extremely fair recommendation, even when the relevance- and fairness-only scores are close to the theoretical worst value. Based on these findings, we formulated recommendations on the appropriate usage of these measures.

Future work includes improving the design of joint measures by addressing or mitigating the limitations of the current measures outlined, to have a single score that reflects recommendation relevance and fairness more accurately and in a more balanced way. The individual fairness and relevance measures can also be optimised jointly with multi-objective approach, to obtain both fair and relevant recommendations. Future work can also investigate whether the findings hold when the models are directly optimised for fairness, or when different family of models are used.

Acknowledgements

Funded by Algorithms, Data & Democracy (Villum & Velux funds).

Chapter 4

Relevance-aware Individual Item Fairness Measures in Recommender Systems: Limitations and Usage Guidelines

Abstract

Recommender Systems (RSs) aim to provide relevant items to users, with a recent emphasis on improving recommendation fairness. Quantifying fairness of the recommended items can be done with two types of evaluation measures: measures that are purely based on item exposure (*exposure-based*) and measures that account for both item exposure and item relevance (*relevance-aware*). While exposure-based measures have been thoroughly analysed, relevance-aware measures have not been examined in such detail yet. We gather all existing relevance-aware individual item fairness measures for RSs and study their theoretical properties. We find that all measures suffer from one or more limitations, which may cause issues in their computation, interpretability, or expressiveness. To address this, we correct the affected measures or explain why a limitation is unresolvable. Further, we empirically investigate the extent of the limitations on the measures and compare the original measures to our reformulations under common and extreme evaluation scenarios across real-world and synthetic data. Our experiments show that our reformulated measures successfully resolve the issues in the original measures. We conclude by providing practical guidelines on how to select measures for a range of use cases.

4.1 Introduction

Fairness has become an increasingly important evaluation aspect in Recommender Systems (RSs). RS fairness can be evaluated on an individual level, where the aim is to ensure that similar individuals receive similar treatments [63], or on a group level, which aims to provide similar treatment to both dominant and protected groups (e.g., gender minority) [54, 65]. In this work, we focus on individual fairness, specifically for items.

Individual item fairness in RSs is based on item exposure, i.e., how often an item is recommended to users. There are two broad ways of defining fairness for individual items. One definition focuses solely only the exposure distribution of items [177], where the more uniform the exposure distribution across all items, the fairer. Another definition considers both item exposure and item relevance to users, where fairness means that items of similar relevance should be recommended a comparable proportion of times [19, 168, 209]. Each fairness notion has its own advantages: the exposure-only definition ensures that each item gets a chance to be seen by users, whereas the definition based on exposure and relevance maintains user utility, which is affected by user-item relevance.

Following these two broad definitions, two types of evaluation measures for RS individual item fairness exist: *exposure-only fairness measures*, which do not consider relevance, and *relevance-aware fairness measures*, which consider item exposure w.r.t. relevance. While exposure-only measures have been thoroughly examined with respect to their limitations [177], to the best of our knowledge, the same has not been done for relevance-aware measures. Given that relevance-aware measures are commonly used in RS fairness evaluation [228], they warrant a closer inspection in terms of their potential strengths and weaknesses (e.g., whether they can be easily interpreted, whether there are cases for which they cannot be used), as well as how they relate to exposure-only fairness measures and to common relevance-only measures, like NDCG.

To this end, in our prior work [178] we empirically studied relevance-aware individual item fairness measures for RSs. In this paper, we extend Rampisela et al. [178] as follows:

1. We identify five theoretical limitations of those measures and analyse the extent of the limitations in various evaluation scenarios. To the best of our knowledge, we are the first to identify these limitations in the measures.

2. We amend the measures to resolve the limitations or justify why some limitations cannot be resolved.

3. We empirically show that our corrected measures overcome the limitations of the original measures in evaluating both real-world and synthetic data.

4. We recommend detailed guidelines on measure usage for various practical evaluation scenarios.

4.2 Evaluation Measures of Fairness for Individual Items

We present our notation (§4.2.1) and review existing individual item fairness measures: exposure-based measures (§4.2.2) and relevance-aware measures (§4.2.3). We also review exposure-based measures as we compare them to relevance-aware measures in our experiments (§4.5).

4.2.1 Notation and Examination Functions

Given a set of n items, $I = \{i_1, i_2, \ldots, i_n\}$ and a set of m users, $U = \{u_1, u_2, \ldots, u_m\}$, for each user $u \in U$ we rank the n items to obtain the full recommendation list L_u^n.[1] We denote the list of user u's top-k recommended items as L_u^k; for brevity, we write this as L_u. If item i is relevant to u, we write $r_{u,i} = 1$, otherwise $r_{u,i} = 0$. The set of all items that are relevant for user u is R_u^*. Item i's rank position in user u's recommendation list is $z(u, i)$. Some fairness measures are defined for multiple recommendation rounds. In this case, we denote the item i's rank position for user u in round w as $z(u, i, w)$ and item i's relevance to u in round w as $r_{u,i,w}$. We summarise our notation in Tab. 4.1.

Item fairness in RSs is centred on the concept of exposure, which refers to the item appearance in a user's top-k recommendation. Item exposure can be quantified (weighed) using various examination functions $e(\cdot)$, which model the probability of a user viewing an item. All examination functions studied in this work, presented in Tab. 4.2, assume that the probability only depends on the rank position p, or the item position for the user ($z(u, i)$ or $z(u, i, w)$). These functions either apply linear, logarithmic, exponential, or inverse discounting. In general, the exposure weight decreases at the highest rate with the inverse discount and decreases the slowest with the penalty based on Discounted Cumulative Gain (DCG).

The linear examination function, e_{li} and its normalised version $\tilde{e}_{\text{li}}$ discount rank positions up to k linearly [23]. Meanwhile, the examination functions e_{DCG} and

[1] For simplicity, we assume that all n items are always ranked.

Table 4.1: Summary of the notation

Notation	Explanation
$U = \{u_1, u_2, \ldots, u_m\}$	The set of users
$I = \{i_1, i_2, \ldots, i_n\}$	The set of items
$\lvert U\rvert = m$	The number of unique users in dataset
$\lvert I\rvert = n$	The number of unique items in dataset
p	The rank position
k	The cut-off threshold
W	The number of recommendation rounds
$r_{u,i} \in \{0, 1\}$	The relevance of item i to user u
$r_{u,i,w} \in \{0, 1\}$	The relevance of item i to user u in round w
R_u^*	The set of relevant items for user u
$z(u, i) \in \{1, 2, \ldots, n\}$	The rank position of item i for user u
$z(u, i, w) \in \{1, 2, \ldots, n\}$	The rank position of item i for user u in round w
$L_u^k = L_u$	The top-k recommendations for user u
$L_{u,w}$	The top-k recommendations for user u in round w
L_u^n	The full recommendation list for user u
$1_{\mathcal{A}}(x) = 1$ if $x \in \mathcal{A}$, else 0	Indicator function

e_{RBP} apply discounts based on DCG [102] and Rank-Biased Precision (RBP) [230] respectively. The function e_{RBP} has a user patience parameter γ, i.e., the probability of a user viewing the next item in the rank. The patience parameter has been set at, e.g., $\gamma \in \{0.8, 0.9\}$ in previous literature [103, 234]. In e_{inv}, the inverse examination function, the discount factor is based on the inverse of the rank position [191].

4.2.2 Exposure-based Fairness Measures

Exposure-based individual item fairness measures use the item exposure distribution, without considering item relevance to users. In the top-k recommendation scenario, where each user receives a list of k unique items, instead of evaluating the recommendation lists across all users for their relevance, e.g., with Precision@k, exposure-based measures evaluate the recommendations for individual item fairness based on item exposure. Generally, the recommendation is deemed fairer when the exposure is more uniformly distributed across items.

Multiple exposure-based measures have been used to quantify individual item fairness in RSs. All of these measures evaluate fairness based on item exposure in

Table 4.2: Examination functions used by measures studied in this work. We denote as $\tilde{e}$ the min-max normalised examination function.

	equation	measure	reference
linear	$e_{\text{li}}(u,i,w) = k+1-z(u,i,w)$ $\tilde{e}_{\text{li}}(u,i,w) = \frac{e_{\text{li}}(u,i,w)-1}{k-1} = \frac{k-z(u,i,w)}{k-1}$	IAA	[23]
DCG	$e_{\text{DCG}}(u,i,w) = 1/\log_2(z(u,i,w)+1)$	IFD	[161, 205]
RBP	$e_{\text{RBP}}(u,i,w) = \gamma^{z(u,i,w)-1}$	HD, II-F, AI-F	[103, 234]
inverse*	$e_{\text{inv}}(p) = 1/p$	MME, IBO/IWO	[191]

*The inverse examination function directly depends on the rank position p.

the top-k recommendation lists across all users and compare the aggregated item exposure to the total number of items in the dataset. Some measures are adapted from other domains, e.g., the Gini Index was originally developed to quantify income inequality [85], while the Jain Index was developed to quantify fairness in computer networks [101]; others were specifically defined for fairness in recommendation, e.g., QF [264]. For a more detailed review of these measures, we refer the reader to Rampisela et al. [177].

4.2.3 Relevance-aware Fairness Measures

We present seven measures that have been used to evaluate fairness considering relevance (Joint measures henceforth). To our knowledge, these are all the Joint measures for RSs published up to 1 July 2024. Each measure employs an examination function (Tab. 4.2), which quantifies item exposure, and therefore the measure evaluates recommendation fairness of item exposure jointly with item relevance.

We present these measures next. We precede a measure name with ↑ (↓) to indicate that a higher (lower) measure score means a fairer recommendation. The subscript $\cdot_{\text{ori}}$ denotes the original measure formulation, as opposed to our reformulation that is presented in §4.4. Note that all Joint measures, except Hellinger Distance (HD), are defined for multiple rounds of recommendations or stochastic rankings, where the evaluation is performed on a distribution over rankings [19].

4.2.3.1 Inequity of Amortized Attention (IAA) [19]

↓IAA[2] quantifies the aggregated difference between item exposure and item relevance to a user, in a stochastically-generated series of rankings [19]. Intuitively, a sequence of rankings is fair based on IAA when the item exposure is proportional to its relevance to the user. The rank position of the item is a proxy of its exposure. To measure fairness in multiple recommendation rounds, IAA was modified in Borges and Stefanidis [23]:

$$\text{IAA}_{\text{ori}} = \frac{1}{m} \sum_{u \in U} \text{IAA}_{\text{ori}}(u) \tag{4.1}$$

$$\text{IAA}_{\text{ori}}(u) = \frac{1}{n} \frac{1}{W} \sum_{i \in I} \left| \sum_{w=1}^{W} 1_{L_{u,w}}(i) \cdot \tilde{e}_{\cdot}(u,i,w) - \tilde{r}(u,i,w) \right| \tag{4.2}$$

where $\tilde{e}_{(\cdot)}(u,i,w)$ is the min-max normalised linear examination function $\tilde{e}_{\text{li}}(u,i,w)$ (see Tab. 4.2) and $\tilde{r}(u,i,w) \in [0,1]$ is the min-max normalised item i relevance to user u in round w, $r_{u,i,w}$ [23].

The min/max relevance values are based on each item's relevance to a user u in a particular round w, i.e., $\min_{i \in I} r_{u,i,w}$. The higher the relevance, the closer the value is to 1. The IAA score is more unfair with a higher difference between item relevance and its exposure. IAA ranges between $[0,1]$.

4.2.3.2 Individual Fairness Disparity (IFD) [161, 205]

↓IFD averages the pairwise difference of the combined item exposure and item merit, where merit can be defined as a function of item relevance.[3] Similar to IAA, IFD is also based on the concept of allocating item exposure considering its relevance. However, while IAA is based on the disparity between the exposure of each item and its relevance, IFD measures the difference in exposure allocation between item pairs. There are two IFD variations, based on how exposure is combined with relevance: $\text{IFD}_{\div}$, which divides item exposure by its relevance [205], and $\text{IFD}_{\times}$, which multiplies them [161]. Further, the two variants slightly differ in how they compute the pairwise difference, how they form the set of item pairs, and how they weigh item exposure.[4] Fairness in ranking has been evaluated using both versions [161, 205, 242, 243]. We

[2]The measure is named IAA in Raj and Ekstrand [175] and L1-norm in Wang et al. [228].

[3]The item relevance value is taken as the item merit, as per Singh and Joachims [205].

[4]While exposure is weighed proportional to e_{DCG} in Singh and Joachims [205], we directly employ e_{DCG} for simplicity.

write $\text{IFD}_{(\cdot)}$ or IFD when referring to the measure in general.

$$\text{IFD}_{(\cdot)} = \frac{1}{m} \sum_{u \in U} \text{IFD}_{(\cdot)}(u) \tag{4.3}$$

$$\text{IFD}_{\div\text{-ori}}(u) = \frac{1}{|H_u|} \sum_{(i,i') \in H_u} \max\{0, J_{\div}(u,i) - J_{\div}(u,i')\} \tag{4.4}$$

$$\text{IFD}_{\times\text{-ori}}(u) = \frac{1}{n(n-1)} \sum_{i \in I} \sum_{i' \in I \setminus i} [J_{\times}(u,i) - J_{\times}(u,i')]^2 \tag{4.5}$$

$$J_{\div\text{-ori}}(u,i) = \frac{1}{W} \sum_{w=1}^{W} \frac{e_{\text{DCG}}(u,i,w)}{r_{u,i,w}} \tag{4.6}$$

$$J_{\times\text{-ori}}(u,i) = \frac{1}{W} \sum_{w=1}^{W} r_{u,i,w} \cdot 1_{L_{u,w}}(i) \cdot e_{\text{DCG}}(u,i,w) \tag{4.7}$$

where the $J_{(\cdot)}(u,i)$ function combines item i's exposure and relevance for user u and $H_u = \{(i,i') \in I \mid r_{u,i} \geq r_{u,i'} > 0\}$. $\text{IFD}_{\div}$ ranges between $[0, \infty)$ and it is 0 when the exposure of each relevant item is exactly proportional to its relevance [205]. Based on empirical results, $\text{IFD}_{\times}$ ranges between $[0, \infty)$ [242].

4.2.3.3 Hellinger Distance (HD) [103]

↓HD has been used to quantify the fairness of individual items for top-k contextual bandits by taking the difference between the relevance- and interaction-distributions of the top-k items with the highest (ground truth) relevance [103]. The interaction probability depends on item exposure, item relevance, and user patience (i.e., the probability of a user viewing the next item in the ranking). HD deems a recommendation fair when an item's interaction probability (Eq. (4.10)) is proportional to its relevance probability (Eq. (4.9)). To obtain the relevance and interaction probabilities, for each user, k items with the highest (ground truth) relevance are selected to construct a reference list used in another step. The interaction probability of each item in the reference list depends on its rank position in another item list, which is based on model predictions. In the reference list, for each item at the same position (but for different users), the relevance probabilities are aggregated across users. We therefore obtain two aggregated values for each rank position: relevance

and interaction. These aggregated values are then inputted to the distance metric (Eq. (4.8)).

$$\text{HD}_{\text{ori}} = \frac{1}{\sqrt{2}}\sqrt{\sum_{p=1}^{k}\left(\sqrt{q'_p} - \sqrt{c'_p}\right)^2} \tag{4.8}$$

$$q'_p = \frac{1}{m}\sum_{u\in U}\sum_{i\in I}\delta\left(z^*(u,i) = p\right)\cdot r'_{u,i} \tag{4.9}$$

$$c'_p = \frac{1}{m}\sum_{u\in U}\frac{c^*_{u,p}}{\sum_{\ell=1}^{k} c^*_{u,\ell}} \tag{4.10}$$

$$c^*_{u,p} = \sum_{i\in I}\delta\left(z^*(u,i) = p\right)\cdot c^{full}_{u,i} \tag{4.11}$$

$$c^{full}_{u,i} = c'_{u,p} \text{ if } \exists p : z(u,i) = p\,, \text{ otherwise } 0 \tag{4.12}$$

$$c_{u,p} = \sum_{i\in L_u}\delta\left(z(u,i) = p\right)\cdot r_{u,i}\cdot\gamma\; e_{\text{RBP}}(u,i)\cdot s_{u,p} \tag{4.13}$$

$$s_{u,p} = \prod_{1\le j<p} 1 - \sum_{i\in L_u}\delta(z(u,i) = j)\cdot r_{u,i} \tag{4.14}$$

where q'_p and c'_p are the normalised relevance and interaction probability of the item at position p respectively, and interaction is based on both relevance and exposure. Item i's position sorted based on ground-truth relevance is $z^*(u,i)$. User u's probability of interacting with an item at position p, $c_{u,p}$, is based on three variables: $s_{u,p}$, the probability that all items before position p are irrelevant for u, and the user patience $\gamma\; e_{\text{RBP}}(u,i)$. The user-wise normalised relevance value of item i to user u is $r'_{u,i} = r_{u,i}/\sum_{i\in I} r_{u,i}$ and the user-wise normalised interaction probability is $c'_{u,p} = c_{u,p}/\sum_{p=1}^{k} c_{u,p}$. The value of $\delta(\cdot) = 1$ if the expression $\cdot$ is true, otherwise it is 0. The range of HD is between 0 and 1 [96].

4.2.3.4 Mean Max Envy (MME) [191]

↓MME is based on the envy-freeness concept, which deems a recommendation fair if no item is disadvantaged by its own exposure allocation, compared to having any other item's exposure allocation. At a high level, MME can be said to quantify unfairness as the disadvantage suffered by an item if its exposure allocation is swapped with that of another item. The disadvantage is measured based on an impact score, which depends on item exposure and relevance: given the full recommendation lists, L^n_u (size n) across all users, each item i is first swapped with another item i'. For each

item, the pre- and post-swap impact scores are computed. If item i's pre-swap score is not less than its post-swap score, we have envy-freeness for item i w.r.t. item i'. MME averages the maximum difference in the impact imposed if item i's exposure allocation is swapped with that of another item i'. MME is defined as follows:

$$\text{MME}_{\text{ori}} = \frac{1}{n} \sum_{i \in I} \left\{ \max_{i' \in I} Imp_i(i') - Imp_i(i) \right\} \tag{4.15}$$

$$Imp_i(i') = \sum_{u \in U} \sum_{p=1}^{k} r_{u,i} \cdot e_{\text{inv}}(p) \cdot X_{u,i',p} \tag{4.16}$$

$$X_{u,i',p} = \frac{1}{W} \frac{1}{m} \sum_{w=1}^{W} 1_{L_{u,w}}(i') \cdot \delta(z(u, i', w) = p) \tag{4.17}$$

where $Imp_i(i')$ is the impact of item i when its exposure allocation is swapped with that of item i', $X_{u,i',p}$ is the probability of recommending item i' to user u at position p across W recommendation rounds, and $e_{\text{inv}}(p)$ is the exposure weight of item at position p, where the weight is discounted with the inverse examination function (see Tab. 4.2). The range of MME is $[0, \infty)$.

4.2.3.5 Item Better-Off (IBO) & Item Worse-Off (IWO) [191]

↑IBO and ↓IWO are based on the principle of *dominance over uniform ranking*, where fairness is achieved when every item has a higher impact (as defined in MME §4.2.3.4) under the current ranking policy, than if it were under the uniform random ranking policy π_{unif}, which randomly samples all possible item permutations. IBO/IWO is the percentage of items for which the current ranking policy improves/worsens impact by at least 10% compared to π_{unif}[5]:

$$\text{IBO}_{\text{ori}} = \frac{100}{n} \sum_{i \in I} \delta \left(\frac{Imp_i(i)}{Imp_i^{unif}} \geq 1.1 \right) \tag{4.18}$$

$$\text{IWO}_{\text{ori}} = \frac{100}{n} \sum_{i \in I} \delta \left(\frac{Imp_i(i)}{Imp_i^{unif}} \leq 0.9 \right) \tag{4.19}$$

$$Imp_i^{unif} = \frac{1}{m} \frac{1}{n} \sum_{p=1}^{k} \frac{1}{p} \cdot \sum_{u \in U} r_{u,i} \tag{4.20}$$

[5]The threshold is 10% in Saito and Joachims [191] and in our work, but this can be a variable.

$Imp_i(i)$ is as per Eq. (4.16) and Imp_i^{unif} is the impact item i has under the ranking policy π_{unif}, with $e_{\text{inv}}(p)$ as the examination function (see Tab. 4.2). The range of IBO/IWO is $[0, 100]$.

4.2.3.6 Individual-user-to-Individual-item Fairness (II-F) [234]

↓II-F was first proposed by Diaz et al. [56] as an unfairness measure in ranking, where unfairness is the difference between system exposure and target exposure of individual queries and individual items. II-F was reformulated for RSs [234] as:

$$\text{II-F}_{\text{ori}} = \frac{1}{m}\frac{1}{n}\sum_{u\in U}\sum_{i\in I}\left(E_{u,i} - E^*_{u,i}\right)^2 \tag{4.21}$$

$$E_{u,i} = \frac{1}{W}\sum_{w=1}^{W} 1_{L_{u,w}}(i)\cdot e_{\text{RBP}}(u,i,w) \tag{4.22}$$

$$E^*_{u,i} = \frac{r_{u,i}}{|R^*_u|}\cdot\frac{1-\gamma^{|R^*_u|}}{1-\gamma} \text{ if } |R^*_u| > 0\,, \text{ otherwise } 0 \tag{4.23}$$

where $E_{u,i}$ is the system exposure, i.e., the expected exposure of i to u under a stochastic ranking policy. $E^*_{u,i}$ is the target exposure, i.e., the expected exposure of i to u under an ideal stochastic ranking policy, where relevant items receive equal expected exposure [56]. As such, II-F deems a recommendation fair when the system-allocated item exposure is equal to the exposure allocated to items by the ideal stochastic ranking policy. The RBP-based examination function (see Tab. 4.2) is used in $E_{u,i}$ and the formulation of $E^*_{u,i}$ also depends on this examination function [234]. $|R^*_u|$ is the number of relevant items for user u. II-F ranges between $[0, 1]$.

4.2.3.7 All-users-to-Individual-item fairness (AI-F) [234]

↓AI-F quantifies how much RSs under/overexpose an item to all users by averaging the deviation of the overall system exposure from a target exposure:

$$\text{AI-F}_{\text{ori}} = \frac{1}{n}\sum_{i\in I}\left(\frac{1}{m}\sum_{u\in U}E_{u,i} - \frac{1}{m}\sum_{u\in U}E^*_{u,i}\right)^2 \tag{4.24}$$

where the system exposure $E_{u,i}$ and the target exposure $E^*_{u,i}$ are given in Eqs. (4.22)–(4.23) (the same as the ones for II-F). As with II-F, AI-F also evaluates fairness based on the disparity between the system exposure and the target exposure. While this difference is computed per user-item pair in II-F, AI-F computes the exposure difference

Table 4.3: Measures of relevance-aware individual item fairness and their theoretical limitations. We have identified two causes of non-realisability, denoted by C in this table. Source refers to where the limitation is first identified. Empty cells mean that the measure does not suffer from the limitation.

Legend ●: we fully resolve the limitation ◐: we partially resolve the limitation for single-round and binary relevance cases ○: the limitation is unresolvable (§4.4.3)	Source	IAA [23]	IFD$_{\div}$ [205]	IFD$_{\times}$ [161]	HD [103]	MME [191]	IBO/IWO [191]	IL-F [234]	AI-F [234]
Non-realisability: cannot reach max/min score (cause number denoted by C)									
C1. Non-realisability due to unknown formulation of max/min score	[177]				○	○	○		○
C2. Non-uniform exposure weight	us	◐	◐	◐				◐	
Non-localisation: requires item relevance information beyond cut-off k	[153]	○	○		○		○	○	○
Undefinedness: cannot be computed (undefined value)	[177]	●					●		
Zero-exposure: gives zero exposure weight to exposed item	us	●							
Top-k-insensitivity: gives nonzero exposure weight to unexposed item k	us		●						

after summing item exposure across users. Consequently, given a larger percentage of unique items in the recommendation (as opposed to recommending the same few items to all users), ↓AI-F would give a fairer score than ↓II-F. AI-F has a range of $[0, 1]$.

4.3 Measure Limitations

The measures in §4.2.3 suffer from theoretical limitations. Following Rampisela et al. [177], the term "limitation" refers to a measure failing to quantify or fulfill criteria that are important in fairness evaluation. Regardless of the reason (e.g., deliberate design choice or unintentional flaw), the limitation restricts the usage of the measure. Several limitations may render the measure completely unusable (§4.3.3, §4.3.5), but others can still be used with an awareness of the limitation and how it affects the measure or its interpretation (§4.3.1, §4.3.2, §4.3.4). Here, we consider measure usage in the case of top-k recommendations, as one of the most popular recommendation scenarios.

Altogether, we identify five limitations affecting the measures, summarised in Tab. 4.3. In each heading below, we put in brackets the names of the measures affected by the limitation.

4.3.1 Limitation 1: Non-realisability

All JOINT measures are affected by this limitation, which is defined as the measure theoretically failing to reach the max/min score, given all possible item rankings across users [177]. The non-realisability limitation causes difficulty in interpreting the fairness score, as the best or worst possible fairness score changes depending on the dataset distribution, (e.g., how many relevant items a user has), or experimental choices (e.g., the cut-off k). Thus, how far the fairness score is from the max or min score (based on its range) cannot be known. For example, if a lower-is-fairer measure has a [0,1]-range and a model scores 0.2, one might think that the model is not perfectly fair and that fairness can still be improved. Yet, if a score of 0.2 is already the minimum that can be achieved given any combination of values, it would mean that the recommendation cannot be made fairer; none of the JOINT measures scores can show this and their scores may therefore be misleading. In other words, this means that the achievable measure range may not actually be [0,1] as per the theoretical range, but only a subrange of it.

This limitation is first defined in Rampisela et al. [177] in relation to the number of recommendation slots and the number of items in the dataset. Here, we show that the same limitation occurs in relation to the item relevance labels and exposure weights. Four causes of non-realisability were previously identified in Rampisela et al. [177]. We also identify a new cause of non-realisability in the JOINT measures. Next, we explain the two causes of non-realisability, one that has been identified in Rampisela et al. [177] and the new cause that we identified.

4.3.1.1 Cause 1 (↓HD, ↓MME, ↑IBO/↓IWO, ↓AI-F)

The first cause of non-realisability is due to the exact formulation of the max/min achievable score being unknown [177]: given a set of user-item relevance labels, how do we order items in the recommendation list to produce the maximum or minimum fairness score? In other words, the (un)fairest recommendation cannot be determined solely through theoretical analysis due to the complexity of the measure (e.g., score aggregation per item, various permutations of user-item relevance values and their combination with item exposure weight). While one can always enumerate all permutations to get the achievable max/min scores, it is a highly costly process; thus, in the examples below, we use toy examples with a small number of users and items. Even if our hypothetical examples here are only with small values, we argue that a fairness measure should still work properly given any number of users or items. Further, in real-life settings, there are recommendation use cases with only a few items, e.g., insurance product recommendation [22]. Evaluating fairness in

such cases would need a fairness measure that works properly for a small number of items. Note that all measures affected by this limitation except HD are defined for multiple rounds of recommendations, but since MME, IBO, and IWO have only been computed under cases equivalent to a single round of recommendation (i.e., a deterministic ranking policy), we only provide examples for a single recommendation round for these measures.

Example (HD). In principle, ↓HD ranges between 0 and 1, but for $k = m = 2$ and $n = 3$, its maximum score is only 0.707, which is achieved when all top-k items are irrelevant. This maximum score is obtained after iterating through all permutations of the recommendation list and item relevance values while assuming binary relevance $r_{u,i} \in \{0, 1\}$. Given the score of 0.707, when comparing it to the theoretical range of [0,1], one may think that the recommendation is rather unfair, but not severely unfair, even if this is the most unfair score that is empirically achievable given the experimental choice and dataset composition. Hence, the non-realisability may hinder an accurate representation of the fairness level.

Example (MME, IBO, IWO). The most fair ↓MME, MME = 0, means that recommendations are envy-free, which we explain in §4.2.3.4. For $k > 1$, an envy-free recommendation is not always guaranteed [191], so MME = 0 may be unreachable. For example, when $k = n = 3$, $m = 2$, and when each item is relevant to all users, the minimum achievable ↓MME is $\frac{1}{18}$ under a deterministic ranking policy.

While the upper bound (fairest score) of ↓MME is unclear, the fairest scores for ↑IBO (IBO = 1) and ↓IWO (IWO = 0), are not always achievable. For example, under a deterministic ranking policy, when $k = n = 3$, $m = 2$, and for any combinations of relevance values for all users and items, the maximum achievable ↓MME for any permutations of recommendation lists across all users is $\frac{7}{18}$, the fairest ↑IBO is $\frac{2}{3}$, and the fairest ↓IWO is $\frac{1}{3}$.

Both ↑IBO/↓IWO compare the impact, based on exposure and relevance, produced by a recommendation list to the thresholded impact, based on the impact produced by recommendation lists with uniform exposure. The thresholded impact depends on a multiplying constant which is set arbitrarily and a priori (1.1 for ↑IBO and 0.9 for ↓IWO in Eq. (4.18)–(4.19)). It is thus possible that the optimal distribution of impact in our recommendation list is sometimes less than the thresholded impact (for ↑IBO) or more than the thresholded impact (for ↓IWO) due to the dataset composition. Intuitively, not all items may have a better impact than the thresholded impact, which causes ↑IBO < 1, and not all items would have a worse impact than the thresholded impact, which causes ↓IWO > 0. So, comparing impacts as above may cause the fairest score to be unreachable.

Example (AI-F). ↓AI-F may also not reach the unfairest score of 1. For example, given $k = m = 2, n = 3$, for all permutations of items in a possible recommendation list and user-item binary relevance values, its maximum score is only 0.88. This score is obtained when the same k items are recommended to all users, and both items are irrelevant to each user. Even if the setup is extended to multi-round settings, the maximum AI-F would remain 0.88, obtained from recommending the same k items to all users in all rounds. This is because the target exposure $E^*_{u,i}$ (Eq. (4.23)) is independent of the number of rounds, while the system exposure $E_{u,i}$ producing the unfairest AI-F in multi-round settings is the same as that of the single-round, as it produces the maximum aggregated difference with the target exposure.

4.3.1.2 Cause 2 (↓IAA, ↓IFD$_{\div}$, ↓IFD$_{\times}$, ↓II-F)

We identify a second cause of non-realisability, which is due to the items having non-uniform exposure. All measures affected by this limitation compare and combine item exposure and relevance through some functions (e.g., subtraction, division, multiplication). When the exposure weight and relevance do not change at the same rate down the ranks, aggregating these scores at the user/item level, e.g., through computing their differences, will not be exactly zero. As such, the resulting fairness score will also have some residuals due to this discrepancy. As a result, a lower-is-fairer score that should have been 0 (the fairest possible) may not be achieved, leading to the misinterpretation that the recommendation is not completely fair, even if it is already the fairest possible, given the dataset composition.

Example (IAA). For ↓IAA, in an ideal effectiveness-based ranking, relevance and exposure weight may decrease at different rates down the rank. This causes the fairest score to be unreachable, as the minimum score cannot be achieved if the normalised relevance value does not decrease at the same rate as the examination function down the ranks. For example, for $k = 2, n = 4$, assume a user's full (untruncated) recommendation list with normalised relevance values of $[1, 0.8, 0, 0]$, and the list of normalised exposure weights given by the linear examination function $[1, 0, 0, 0]$. The two exposed items are with the highest relevance among the 4 items, but $\text{IAA}_{\text{ori}}(u) = \frac{0.8}{4} = 0.2 > 0$ even though the recommendations are made as fair as possible. The issue persists when extending this to a multi-round scenario ($W > 1$). In the previous example, the fairest $\text{IAA}(u)$ remains 0.2 for any W, as the recommendation list is already ordered by decreasing relevance, assuming the relevance $r_{u,i,w}$ for a specific user-item pair stays constant throughout different rounds. Similarly, the most unfair score cannot be achieved if the normalised relevance value does not decrease at the

same rate as the exposure weight. In this case, the maximum IAA is only 0.7 instead of 1.[6]

Example (IFD). $\downarrow$IFD$_{\div}$ and $\downarrow$IFD$_{\times}$ may neither reach the fairest score of 0 nor the unfairest score of 1. Under a binary relevance scenario, for $k = n = 10$, considering all possible combinations of recommendation list and item relevance values, the fairest $\downarrow$IFD$_{\div}(u)$ is 0.003, while the unfairest is 0.158.[7] Similarly, the fairest $\downarrow$IFD$_{\times}(u)$ can only be 0.0167 and the unfairest is at most 0.265. For non-binary relevance, suppose $r_{u,i} \in \{0, 0.5, 1\}$, $\downarrow$IFD$_{\div}(u)$ can only be 0 when there are exactly two items with nonzero relevance. Otherwise, its minimum is 0.006, and its maximum is 0.428. Under the same setting, $\downarrow$IFD$_{\times}(u)$ is bounded between 0.004 and 0.265. For similar reasons, these issues would also extend to multi-round scenarios, even though the lower/upper bounds may change (i.e., it may be possible to obtain an (un)fairer score than with single rounds).

Example (II-F and AI-F). The unfairest and fairest $\downarrow$II-F score may not be achievable either. For instance, given $k = m = 2, n = 3$, for all possible recommendation lists and item relevance values (considering only binary relevance), the maximum $\downarrow$II-F is only 0.88 (obtained through recommending all relevant items at the top-k), while its minimum is 0.007 (obtained through recommending all irrelevant items at the top-k). We also enumerate all possibilities for two and three recommendation rounds. The minimum II-F can be 0 across two rounds, but it is still not exactly 0, (i.e., 0.0007) if there are three rounds. However, the maximum II-F will still be 0.88 in either case, for the same reason as we explain for AI-F in §4.3.1.1. Hence, even in multi-round scenarios, the theoretical max/min II-F score may not be achievable.

4.3.2 Limitation 2: Non-localisation ($\downarrow$IAA, $\downarrow$IFD$_{\div}$, $\downarrow$HD, $\uparrow$IBO/$\downarrow$IWO, $\downarrow$II-F, $\downarrow$AI-F)

The non-localisation limitation is based on the localisation property of effectiveness measures in Information Retrieval (IR) [153]. The localisation property is defined as "A score at depth k can be computed based solely on knowledge of the documents that appear in the top-k" [153]. We redefine the property as a measure limitation in RSs, where the measure requires item relevance information (e.g., whether the item is relevant to the user, the total number of relevant items) beyond the top-k. While

[6]Obtained by having a recommendation list with non-decreasing normalised relevance values, i.e, $[0, 0, 0.8, 1]$.

[7]Excluding the trivial case where the user only has one relevant item out of all items in the dataset.

this may be a deliberate design choice, it is still a limitation, as having additional relevant items beyond the top-k may cause the measure score to appear fairer than if there are no additional relevant items. In practice, this may lead to overestimation of fairness, and instability in the evaluation results. Since most RS datasets are highly sparse, with only a few relevant items per user, their fairness evaluation can also be affected by this limitation.

We explain the limitation through an example. Suppose that given $k = 2, n = 5$, a user's full recommendation list L_u^n has the normalised relevance values of $[1, 1, ?, ?, ?]$, where a '?' represents unknown relevance, and the list of normalised exposure weights given by the linear examination function is $[1, 0, 0, 0, 0]$. The relevance of an item may be unknown due to, for example, the item not being shown to the user or the item being shown to the user but not rated/interacted with. We refer to items with unknown relevance as unobserved items henceforth. Following the convention in IR evaluation [154], unobserved items are usually regarded conservatively as irrelevant. In this case, $\downarrow\text{IAA}(u) = 0.2$. However, if all unobserved items are actually relevant instead, $\downarrow\text{IAA}(u)$ would be 0.8, and thus the difference from the actual score is 0.6.[8] Considering that IAA ranges in $[0, 1]$, the large difference is worrying, as a score of 0.8 could be interpreted as highly unfair, while 0.2 is rather fair. Even if there is only one unobserved item, the difference in IAA based on the two ways of treating the unobserved items would be 0.2, a non-negligible difference.

Using the same example as the one used for IAA, we show that $\text{IFD}_{\div}(u)$ also suffers from the same limitation. When all unobserved items are regarded as irrelevant, $\downarrow\text{IFD}_{\div}(u) = 0.092$. The opposite treatment will result in a higher $\downarrow\text{IFD}_{\div}(u)$, i.e., 0.114. Worse, having only one additional relevant item can cause an even higher difference in the $\text{IFD}_{\div}(u)$ scores. If the additional relevant item is at position n, e.g., $[1, 1, 0, 0, 1]$, $\downarrow\text{IFD}_{\div}(u) = 0.136$.

IBO, IWO, II-F, AI-F, and HD also suffer from non-localisation. Item relevance information of all user-item pairs is required to compute Eq. (4.20) for IBO/IWO and to compute Eq. (4.23) for II-F/AI-F. On the other hand, HD technically requires only the top-k items with the highest ground truth relevance per user (Eq. (4.10)). However, obtaining this list of items requires (partially) sorting the ground truth relevance of all items, regardless of their position in the predicted ranking.

4.3.3 Limitation 3: Undefinedness (IAA, IBO/IWO)

The undefinedness limitation is defined as the measure being incomputable, e.g., the measure giving an undefined value when encountering a specific case (but not

[8] This difference is the same as 'residual' in Moffat and Zobel [154].

necessarily an edge case) [177]. In both IAA and IBO/IWO, this happens due to division by zero. Consequently, fairness cannot be evaluated as the measure does not yield any scores.

For IAA, when $k = 1$, the denominator of the linear examination function $\tilde{e}_{\text{li}}(u, i, w)$ is 0. The choice of $k = 1$ can be seen on mobile phone applications, where there is limited screen space and thus only the top recommendation is presented to users [19, 41] or in conversational RSs [141]. For IBO/IWO, a division by zero occurs when $Imp_i^{unif} = 0$. This happens when item i is not relevant for any users (see Eq. (4.20)).

Note that IFD$_\div$ has been stated to have the zero denominator problem when encountering irrelevant items with $r_{u,i} = 0$ (Eq. (4.6)) [161, 242], but this is actually false. The original measure [205] only computes fairness based on relevant items. Hence, IFD$_\div$ does not suffer from the undefinedness limitation.

We exclude the trivial, unrealistic cases where there are no users ($m = 0$) or no items ($n = 0$) for this limitation, as evaluating fairness (or any other dimensions) under these scenarios does not make sense. In addition, we identify the following other cases that cause a measure to be undefined. IFD$_\times$ is undefined when $n < 2$, as it is a pairwise measure. IFD$_\div$ is undefined if there is no relevant item for a user (based on the ground truth) as the divisor in Eq. (4.4) will be zero.[9] Similarly, HD requires at least one relevant item per user, otherwise, the divisor in Eq. (4.10) will be zero. However, we do not consider these measures to suffer from the undefinedness limitation for the following reasons. IFD$_\times$ does not suffer from undefinedness for $n < 2$ as evaluating item fairness when there is only one item does not make sense as only one item can be recommended to any users. IFD$_\div$ and HD also do not suffer from undefinedness for users with zero relevant items, as such users are generally not included in the evaluation, which is also the case when evaluating effectiveness for RSs.

4.3.4 Limitation 4: Zero-exposure (↓IAA)

We define the zero-exposure limitation as the measure's examination function giving zero exposure weight to an item that is recommended at the top-k. While this may arguably be an intentional design choice, it is unintuitive and confusing as there is no difference between the exposure weight assigned to an exposed item and an unexposed item.

For ↓IAA, in Eq. (4.1), $\tilde{e}_{\text{li}}(u, i, w) = 0$ for an (exposed) item at position k. The zero value is due to min-max normalisation that re-scales all exposure weights of

[9] This is related to the completeness property [153].

the top-k recommended items within $[0, 1]$. However, an item at k is still exposed; weighing the item with a nonzero exposure value makes more sense, to distinguish its weight from the unexposed item at position $k+1$. In practice, assigning zero exposure weight for an item at position k may result in an underestimation of fairness if the item is relevant and vice versa. For example, if an item at $k = 10$ has a relevance value of 1, having a positive exposure weight (e.g., 0.1) at that position would result in a 0.9 difference between the item exposure weight and the item relevance. If the exposure weight is 0 instead, the difference would be larger (i.e., 1). In this case, the ↓IAA score is higher (more unfair) than it should be.

4.3.5 Limitation 5: Top-k-insensitivity (↓IFD$_{\div}$)

We define the limitation of top-k-insensitivity as the measure's inability to respond to the change in cut-off k. This means that given a set of recommendations across users, the fairness score would remain the same given any values of k. A measure that cannot reflect a change in fairness level given a different k is not ideal, as recommendations are commonly evaluated for both effectiveness and fairness at different cut-offs. This limitation can also be seen as the inverse of the zero-exposure limitation, in the sense that a nonzero exposure weight is assigned to an unexposed item.

IFD$_{\div}$ suffers from this limitation as Eq. (4.6) does not incorporate the function $1_{L_{u,w}}(i)$, which indicates whether item i is recommended at top-k. Consequently, items at position $k + 1$ and beyond are assigned nonzero exposure weights, which is unintuitive. For example, given $n = 3$, suppose a user u's full recommendation list, L_u^n has the following relevance values $[0, 1, 1]$. The exposure weights for the three items based on e_{DCG} are $[1, 0.631, 0.500]$. In this case, given any $k \in \{1, 2, 3\}$, ↓IFD$_{\div} = 0.033$. Yet, we may obtain a different IFD$_{\div}$ score if we assign zero exposure weights to the unexposed items. Suppose $k = 2$, the exposure weights should have been $[1, 0.631, \underline{0}]$ instead and ↓IFD$_{\div} = 0.158$, which is higher than the first score. With $k = 1$, ↓IFD$_{\div} = 0$, which is lower than the initial score. Hence, this limitation may lead to the under/overestimation of fairness, as the ↓IFD$_{\div}$ score is lower/higher than it should be.

4.4 Resolving Limitations

We show how to resolve some of the measure limitations, either fully or partially, and explain why some remain unresolvable (Tab. 4.5). To distinguish between the original version of an evaluation measure M and our modification of it, we use M_{ori}

and M_{our}, respectively. When referring to the measure in general, we omit the $\cdot_{\text{ori}}$ or $\cdot_{\text{our}}$ subscripts.

4.4.1 Resolving Non-realisability (Limitation 1) and Top-k-insensitivity (Limitation 5)

We resolve the **top-k-insensitivity** limitation for the only measure that suffers from it, IFD$_{\div}$, and we partially resolve **non-realisability** due to Cause 2 for all measures that suffer from the limitation under Cause 2 (i.e., IAA, IFD, and II-F). We resolve top-k-insensitivity by employing the top-k indicator function in the measure. We partially resolve non-realisability by normalising the measure scores based on their theoretically achievable minimum and maximum values. Specifically, we min-max normalise the score per user, using the ideal and worst recommendation list (in terms of fairness), similar to the idea of normalising DCG [102].[10] As a result of the normalisation, the scores 0 and 1 refer to the fairest and unfairest possible recommendations, respectively, and the (un)fairest recommendation is unique for each measure. We provide the intuition on obtaining the (un)fairest recommendation list in this subsection, and the mathematical workings in the appendix. In some cases, it is difficult to identify the (un)fairest recommendation directly from the measure equations, so we simulate all possible rankings for various data distributions (with a small number of items) and experimental choices (e.g., k). We then compute the measure and identify the rankings that produce the (un)fairest scores. Tab. 4.4 summarises the strategies to achieve the (un)fairest recommendations per measure.

We resolve the non-realisability limitation (Cause 2) only for single-round cases (or multiple rounds with a deterministic ranking policy) and for binary relevance scenarios, as the aggregation across rounds and non-binary relevance introduce extra complexity in formulating the (un)fairest recommendation list. While this may seem like a constrained setting, both scenarios are popular in RS evaluation [31, 228]. Furthermore, we also exclude the highly unlikely scenario where the relevance value of all n items are the same for a user u ($r_{u,1} = r_{u,2} = \cdots = r_{u,n}$).[11]

[10] As the normalisation is done per user, prior to averaging across all users, correcting through posthoc normalisation on the final score of the original measures is not possible.

[11] In this case, the particular user u may either be excluded from the computation or the user score may be set to 0 for lower-is-fairer measures and to 1 for higher-is-better measures.

Table 4.4: Summary of ranking strategies to achieve the (un)fairest scores of the JOINT measures.

Measure	Fairest	Unfairest
$\text{IAA}_{\text{our}}(u)$	As many relevant items as possible at the top-k	As few relevant items as possible at the top-k
$\text{II-F}_{\text{our}}(u)$	As many relevant items as possible at the top-k	As few relevant items as possible at the top-k
$\text{IFD}_{\div\text{-our}}(u)$	As few relevant items as possible at the top-k	Half of the relevant items at the top; the other half unexposed
$\text{IFD}_{\times\text{-our}}(u)$	As few relevant items as possible at the top-k	*As many relevant items as possible at the top-k

*Works for most cases, but not all. See §4.4.1.3.

4.4.1.1 Inequity of Amortized Attention (IAA)

To resolve the non-realisability limitation in IAA, we compute $\text{IAA}_{\text{min}}(u)$, i.e., the fairest possible ↓IAA for user u, and $\text{IAA}_{\text{max}}(u)$, the unfairest possible ↓IAA for u. The fairest $\text{IAA}(u)$ is obtained from a recommendation list that is based on monotonically decreasing item relevance for the user, while the unfairest $\text{IAA}(u)$ is based on a list with monotonically increasing item relevance. To minimise $\text{IAA}(u)$, we have to minimise the absolute difference between each item's relevance score and its exposure weight. Intuitively, all relevant items ($r_{u,i} = 1$, for binary relevance case) should be placed at the top of the list, as item exposure weights at the top-k are nonzero, and the irrelevant items ($r_{u,i} = 0$) should be recommended beyond the top-k, where the exposure weights are zero. In contrast, maximising $\text{IAA}(u)$ requires doing the opposite, i.e., placing the irrelevant items at the top and the relevant items at the bottom.

We reformulate IAA as follows:

$$\text{IAA}_{\text{our}} = \frac{1}{m} \sum_{u \in U} \text{IAA}_{\text{our}}(u) \tag{4.25}$$

$$\text{IAA}_{\text{our}}(u) = \frac{\text{IAA}_{\text{ori}}(u) - \text{IAA}_{\text{min}}(u)}{\text{IAA}_{\text{max}}(u) - \text{IAA}_{\text{min}}(u)} \tag{4.26}$$

where all $\text{IAA}_{(.)}(u)$ in Eq. (4.26) are computed with the corrected examination function, which we explain later in §4.4.2.

4.4.1.2 Individual Fairness Disparity with Division ($\text{IFD}_{\div}$)

$\text{IFD}_{\div\text{-ori}}$ suffers from **top-k-insensitivity** as it does not employ the function that indicates whether an item i is in the top-k recommendations for user u in round w, $1_{L_{u,w}}(i)$. As such, we resolve this limitation by incorporating the indicator function in the $J_{\div\text{-ori}}(u,i)$ function, which combines item exposure and relevance. We reformulate $J_{\div\text{-ori}}(u,i)$ as follows:

$$J_{\div\text{-our}}(u,i) = \frac{1}{W}\sum_{w=1}^{W}\frac{1_{L_{u,w}}(i)\cdot e_{\text{DCG}}(u,i,w)}{r_{u,i,w}} \tag{4.27}$$

The above modification zeros out the combined item exposure and item relevance values for items that are ranked beyond the top-k. Hence, given a different cut-off k, $\text{IFD}_{\div\text{-our}}$ (Eq. (4.28)) can more accurately quantify the disparity between the combined exposure-relevance scores of items in and out of the top-k.

As $\text{IFD}_{\div\text{-ori}}$ also suffers from non-realisability, we explain how to normalise its score per user. To obtain the fairest $\downarrow\text{IFD}_{\div}(u)$, i.e., $\text{IFD}_{\div\text{-min}}(u)$, we place all relevant items at the bottom of the list, as $\text{IFD}_{\div}$ takes the pairwise difference of the exposure weight of only the relevant items (after resolving the top-k-insensitivity limitation). Thus, when the relevant items are at the bottom of the list, their exposure weights are (closer to) 0, resulting in the least $\downarrow\text{IFD}_{\div}(u)$ score.

Obtaining the recommendation list with the unfairest $\downarrow\text{IFD}_{\div}(u)$, i.e., $\text{IFD}_{\div\text{-max}}(u)$, is not as straightforward. For the base case where there are two relevant items for a user ($|R_u^*| = 2$), the maximum difference can be obtained when placing one relevant item at the top, and the other one at the bottom.[12] Generalising to an even number of $|R_u^*|$, following the same idea, placing half of the relevant items at the top and the rest at the bottom should achieve the maximum $\text{IFD}_{\div}(u)$. For example, if the dataset has $n = 8$ items, and 4 are relevant to u, the items should be placed such that the item relevance is ordered in the following way: $[1,1,0,0,0,0,1,1]$.

If $|R_u^*|$ is odd, $|R_u^*|$-1 of the items can be recommended with the same strategy for an even number of $|R_u^*|$. The positioning of the remaining item can be closer to either the top or the bottom, e.g., $[1,1,\underline{1},0,0,0,1,1]$ or $[1,1,0,0,0,\underline{1},1,1]$. Through a simulation with various numbers of items, $n \in \{10,15,20\}$, cut-off $k \in \{1,3,5,10\}$, and an odd number of relevant items, $|R_u^*| \in \{3,5,\ldots,k-1\}$, we find that placing the item either way always produces the unfairest $\text{IFD}_{\div}(u)$.

[12] There may be multiple possible recommendation lists that result in the same maximum $\text{IFD}_{\div}(u)$ score, this is just one possibility.

We reformulate $\text{IFD}_{\div}(u)$ as follows:

$$\text{IFD}_{(\cdot)\text{-our}} = \frac{1}{m} \sum_{u \in U} \text{IFD}_{(\cdot)\text{-our}}(u) \tag{4.28}$$

$$\text{IFD}_{\div\text{-our}}(u) = \frac{\text{IFD}_{\div\text{-ori}}(u) - \text{IFD}_{\div\text{-min}}(u)}{\text{IFD}_{\div\text{-max}}(u) - \text{IFD}_{\div\text{-min}}(u)} \tag{4.29}$$

where all $\text{IFD}_{\div\text{-}(\cdot)}(u)$ in Eq. (4.29) are computed with our reformulation of the function combining item exposure and relevance, $J_{\div\text{-our}}$ (Eq. (4.27)). Note that this normalisation only applies to users with at least two relevant items ($|R_u^*| \geq 2$). As $\text{IFD}_{\div}(u)$ takes the pairwise difference between the combined exposure-relevance value of a user's relevant items, $\text{IFD}_{\div\text{-ori}}(u)$ is always zero when a user only has one relevant item. One way to treat such users is to assign $\text{IFD}_{\div\text{-our}}(u) = 0$ (the fairest score). Alternatively, they can also be excluded from the evaluation. As their scores would always be zero, these users would pull the score of $\downarrow\text{IFD}_{\div\text{-our}}$ down, making the score seem fairer than it should be.

4.4.1.3 Individual Fairness Disparity with Multiplication ($\text{IFD}_{\times}$)

For $\text{IFD}_{\times\text{-ori}}$, we obtain the fairest $\downarrow\text{IFD}_{\times\text{-min}}(u)$ through the same recommendation list that produces $\text{IFD}_{\div\text{-min}}(u)$, i.e., all relevant items are at the bottom. As items at the bottom have less exposure weight than the ones at the top, this ranking minimises the pairwise difference of the item scores, which are the item exposure weight times its relevance.

Formulating the item ranking with the unfairest $\text{IFD}_{\times\text{-ori}}$, i.e., $\downarrow\text{IFD}_{\times\text{-max}}(u)$ is not feasible through analysing the measure equation. To find the ranking that produces $\text{IFD}_{\times\text{-max}}(u)$, we simulate rankings with various numbers of items, $n \in \{10, 15, 20\}$, cut-off $k \in \{1, 3, 5, 10\}$, and relevant items, $|R_u^*| \in \{2, 3, \ldots, k-1\}$. For each combination of the experimental choices, we permute all possible rankings, compute $\text{IFD}_{\times\text{-ori}}$ for each ranking, and identify which rankings produce the unfairest score. We find that the items can be ranked in at least one of these two ways: (i) all relevant items at the top; or (ii) a items at the top, and $|R_u^*| - a$ items at the bottom. Our simulation shows that strategy (i) produces the unfairest score for small cut-offs $k \in \{1, 3\}$. As $|R_u^*|$ increases, we observe that strategy (i) fails and strategy (ii) succeeds in achieving the unfairest score; this happens at different amounts of $|R_u^*|$, given a different number of items n.[13] For example, given $n = 20$ and $k = 10$, strategy (i) works for $|R_u^*| \in \{2, 3, \ldots, 6\}$, and for $|R_u^*| \in \{7, \ldots, n-k\}$ onwards,

[13] We find that with a larger n, strategy (i) starts failing at a larger $|R_u^*|$.

the unfairest recommendation list is in the form of (ii), i.e., placing 6 relevant items at the top, and $|R_u^*| - 6$ items at the bottom. For $|R_u^*| \in \{n-k+1, \ldots, n-1\}$, i.e., the general form is $|R_u^*| - (n-k) = |R_u^*| - 10$ items at the top, and the rest at the bottom. There are some exceptions, where this increasing pattern does not hold. For example, given $k = n = 10$, for $|R_u^*| \in \{6, 7\}$, strategy (ii) works, i.e., placing 5 items at the top and 1 item at the bottom for $|R_u^*| = 6$; we denote this as 5–1. For $|R_u^*| = 7$, the list is in the form of 5–2, but for $|R_u^*| = 8$, the list is in the form of 4–4 instead, which deviates from the pattern. However, this deviation is unexpected as the evaluation scenario is highly unlikely (i.e., the cut-off is the same as the number of items), and in any case, the scores of the 4–4 and 5–3 lists only differ by ∼0.0006.

To find $\text{IFD}_{\times\text{-max}}(u)$ (or a good estimate of it), we precompute $\text{IFD}_{\times\text{-ori}}(u)$ with the two strategies for a given k, m, and $|R_u^*|$, and take the maximum as $\text{IFD}_{\times\text{-max}}(u)$. We then compute $\text{IFD}_{\times\text{-our}}$, which resolves the non-realisability limitation in the measure, by averaging $\text{IFD}_{\times\text{-our}}(u)$ across all users:

$$\text{IFD}_{\times\text{-our}}(u) = \frac{\text{IFD}_{\times\text{-ori}}(u) - \text{IFD}_{\times\text{-min}}(u)}{\text{IFD}_{\times\text{-max}}(u) - \text{IFD}_{\times\text{-min}}(u)} \tag{4.30}$$

4.4.1.4 Individual-user-to-Individual-item Fairness (II-F)

To resolve the non-realisability in II-F, we first rewrite II-F (Eq. (4.21)) as follows, to separate the II-F score per user:

$$\text{II-F}_{(\cdot)} = \frac{1}{m} \sum_{u \in U} \text{II-F}_{(\cdot)}(u) \tag{4.31}$$

$$\text{II-F}_{\text{ori}}(u) = \frac{1}{n} \sum_{i \in I} \left(E_{u,i} - E_{u,i}^*\right)^2 \tag{4.32}$$

We then min-max normalise II-F(u):

$$\text{II-F}_{\text{our}}(u) = \frac{\text{II-F}_{\text{ori}}(u) - \text{II-F}_{\text{min}}(u)}{\text{II-F}_{\text{max}}(u) - \text{II-F}_{\text{min}}(u)} \tag{4.33}$$

To obtain $\text{II-F}_{\text{min}}(u)$, i.e., the fairest ↓II-F(u), all relevant items should be ranked at the top, while to obtain $\text{II-F}_{\text{max}}(u)$, i.e., the unfairest ↓II-F(u), all relevant items should be ranked at the bottom. This is because minimising the sum of squares in Eq. (4.32) requires $E_{u,i}$, i.e., the exposure weight of item i in user u's recommendation list to be as close as possible to $E_{u,i}^*$, which is computed based on item relevance. Therefore, relevant items (with $r_{u,i} = 1$) should be placed at the top, so that they are

subtracted with nonzero exposure weights, and irrelevant items ($r_{u,i} = 0$) should be ranked below, such that the difference between exposure weight and $E^*_{u,i}$ is minimised. Following a similar line of reasoning, the opposite should be done to maximise the sum of squares. We also verify through simulations with $\gamma = 0.8$, $n \in \{10, 15, 20\}$, cut-off $k \in \{1, 3, 5, 10\}$, and relevant items, $|R^*_u| \in \{2, 3, \ldots, k-1\}$ that these ranking strategies always produce the min/max II-F(u), or close to that (i.e., with a negligible difference of $<10^{-3}$). II-F$_{\text{our}}$, which is computed with the min-max normalised II-F(u) (Eq. (4.33)), resolves the non-realisability limitation in II-F$_{\text{ori}}$.

4.4.2 Resolving Undefinedness (Limitation 2) and Zero-exposure (Limitation 4)

We resolve **undefinedness** for all measures that suffer from it (i.e., IAA, IBO, and IWO) and we resolve **zero-exposure** for IAA, which is the only measure that has it.

For IAA$_{\text{ori}}$, the cause of **undefinedness** and **zero-exposure** is the erroneous min-max normalisation in the linear examination function (see Tab. 4.2). We correct the normalisation by substituting k with $k+1$ in the linear examination function (Tab. 4.2), which results in:

$$\tilde{e}_{\text{our-li}}(u, i, w) = \frac{e_{\text{li}}(u, i, w) - 1}{k} = \frac{k + 1 - z(u, i, w)}{k} \tag{4.34}$$

This modification ensures that items at rank $\leq k$ receive nonzero exposure and items at $k+1$ onwards receive zero exposure. Accordingly, the IAA corrected for the non-realisability limitation (IAA$_{\text{our}}$, Eq. (4.25)) needs to be computed with this corrected examination function.

In the following example, we show how using the corrected examination function may result in a different score than the original examination function. For $k = 2, n = 4$, assume a user's recommendation list with normalised relevance values of $[0.8, 1, 0, 0]$. The normalised exposure weights given by the original examination functions are $[1, \underline{0}, 0, 0]$, while the ones given by the corrected linear examination function $\tilde{e}_{\text{our-li}}(u, i, w)$ are $[1, \underline{0.5}, 0, 0]$. The score of IAA$_{\text{our}}(u) = 0.2$ with the former, whereas IAA$_{\text{our}}(u) = 0.133$ with the latter.

For IBO$_{\text{ori}}$/IWO$_{\text{ori}}$, to avoid **undefinedness** due to division by zero, we move the divisor Imp_i^{unif} in the left-hand side of the inequalities in Eq. (4.18) & Eq. (4.19) to the right-hand side. We reformulate Eq. (4.18) & (4.19) as follows:

$$\text{IBO}_{\text{our}} = \frac{100}{|I^-|} \sum_{i \in I^-} \delta \left(Imp_i(i) \geq 1.1 \cdot Imp_i^{unif} \right) \tag{4.35}$$

Table 4.5: Summary of how we fully or partially resolve the limitations. Dash '–' means that the limitation does not apply.

Measure	Non-realisability	Undefinedness	Zero-exposure	Top-k-insensitivity
IAA	Min-max normalising score per user	Redefining the linear examination function e_{li}	Redefining the linear examination function e_{li}	–
IFD$_{\div}$	*idem*	–	–	Adding top-k indicator function
IFD$_{\times}$	*idem*	–	–	–
II-F	*idem*	–	–	–
IBO/IWO	–	Excluding items that no users will find relevant	–	–

$$\text{IWO}_{\text{our}} = \frac{100}{|I^-|} \sum_{i \in I^-} \delta \left(Imp_i(i) \leq 0.9 \cdot Imp_i^{unif} \right) \tag{4.36}$$

where I^- is the set of items with at least one user that finds the item relevant. This ensures that the set of items that cause $\delta(\cdot) = 1$ in IBO_{our} is disjoint from that in IWO_{our}. We exclude items which no users would find relevant, as for these items $Imp_i(i) = Imp_i^{unif} = 0$, causing the same items to be considered 'better-off' and 'worse-off' at the same time. Otherwise, this results in problems in the interpretability of the scores as some items may be double counted in IBO_{ori} and IWO_{ori}.

4.4.3 Unresolvable Limitations

We have resolved some measure limitations and summarised these in Tab. 4.5. We explain why the rest are unresolvable.

Non-realisability (Cause 1) cannot be resolved as it is unknown what kind of recommendation list produces the (un)fairest score [177]. Without this formulation of the min/max score, the measures cannot be normalised to be between 0 and 1. The measures that suffer from non-realisability under Cause 1 are more complex than the ones under Cause 2; the former aggregate scores across items and tend to have more variables than the latter. As such, no explicit solution or mathematical formulation

of the measures' upper and/or lower bounds has been determined. Furthermore, computationally solving for the bounds (e.g., through constrained optimization) is also impractical, considering the large size of a typical RS dataset. Technically, if the theoretically achievable best and worst fairness scores are known, the non-realisability in HD, MME, IBO/IWO, and AI-F can be resolved in a similar manner to the measures in §4.4.1.

Non-localisation cannot be resolved as the measures require item relevance information beyond the top-k by their definitions. In previous work, to avoid having unknown item relevance information beyond the top-k, the model-predicted relevance is used instead of the relevance based on the test set (e.g., IAA [19], HD [103]). Consequently, the fairness score depends on how well the model estimates relevance, and will also be incomparable to that produced with other models that may have different estimations of relevance. When comparing multiple models, even if the estimated relevance from one model is used to compute the fairness score for all models, this raises the question of which model should be used as a reference. Alternatively, one can also report the possible measure range, in a similar way to RBP residuals [154]. However, computing the Joint measures range may be more complex than the RBP range. Unlike the RBP range, which narrows when given more document relevance labels, the Joint measure range may widen with fewer missing labels beyond the top-k, due to, e.g., pairwise comparison (§4.3.2). Thus, reporting the range of the Joint measures can be challenging.

For $\text{IFD}_{\div}$, II-F, and AI-F, non-localisation can technically be resolved by modifying the measure partially. For example, if we have a fusion of $\text{IFD}_{\div}$ and $\text{IFD}_{\times}$, where the exposure-relevance aggregation is still done through the division operation, but for all item-pairs, instead of only pairs of relevant items, then this modification would result in a measure that does not suffer from non-localisation. For II-F and AI-F, the 'target exposure' (Eq. (4.23)) can be changed to another target that does not involve the number of relevant items for a user. However, all these approaches are substantial changes beyond simple corrections, which may fundamentally alter the behaviour or the concept behind the measure.

4.5 Experimental Setup

We study the original Joint measures (§4.2.3) and compare them to our corrected versions (§4.4), which resolve some limitations of the original measures (see Tab. 4.3 for the summary of limitations and Tab. 4.5 for the summary of the resolutions). We also investigate the robustness of the original measures by analysing the extent

Table 4.6: Statistics of the preprocessed datasets.

dataset	#users (m)	#items (n)	#interactions	sparsity (%)
Lastfm [32]	1,859	2,823	71,355	98.64
Amazon-lb [156]	1,054	791	12,397	98.51
QK-video [247]	4,656	6,423	51,777	99.83
ML-10M [94]	49,378	9,821	5,362,685	98.89

of the limitations that they suffer from. Here, we provide our general experimental setup, and we further describe our experiments in §4.6.[14]

Datasets

We use four real-world datasets with varying sizes from several domains: music (Lastfm) [32], e-commerce (Amazon Luxury Beauty, i.e., Amazon-lb) [156], videos (QK-video) [247], and movies (ML-10M) [94]. All of the datasets are as provided by Zhao et al. [260], except for QK-video, which is obtained from [247]. QK-video has several interaction types; we only use the 'sharing' interactions.

Preprocessing

We retain users and items with at least 5 interactions (5-core filtering). To handle duplicate interactions, we keep the most recent one. For Amazon-lb and ML-10M, we convert ratings equal/above 3 to 1 and discard the rest, as their ratings range between $[1, 5]$ and $[0.5, 5]$, respectively. Lastfm and QK-video are not converted as they only have implicit feedback. Tab. 4.6 presents statistics of the preprocessed datasets.

Data Splits

To form the train/val/test sets, we apply global temporal splits with a ratio of 6:2:2 on the preprocessed Amazon-lb and ML-10M [151]. As Lastfm and QK-video have no timestamps, we use global random splits with the same ratio. From all splits, we remove users with less than 5 interactions in the train set.

[14]Our code: https://github.com/theresiavr/relevance-aware-item-fairness-measures-recsys.

Recommenders

We rank items with four well-known top-k recommenders: item-based K-Nearest Neighbour (ItemKNN) [55], Bayesian Personalised Ranking (BPR) [185], Variational Autoencoder with multinomial likelihood (MultiVAE) [135], and Neighbourhood-enriched Contrastive Learning (NCL) [139]. BPR, MultiVAE, and NCL are trained for 300 epochs with early stopping using RecBole [260]. The final model is the model with the best Normalised Discounted Cumulative Gain at 10 (NDCG@10) on the validation split.[15] To generate recommendations during testing, we exclude each user's train/val items, and rank the rest of the items.

Fair Re-rankers

The recommenders above are not trained to optimise for fairness, so the items are re-ranked to make the recommendations fairer. We re-rank the top-k' items to increase the exposure of items that are originally ranked beyond the top-k. We set k' to be greater than the cut-off $k = 10$. However, to maintain recommendation relevance, k' should not be overly big (e.g., 100) as a typical RS dataset has very few relevant items per user. For all datasets and recommenders, we set $k' = 25$. We re-rank items per user with COMBMNZ (CM) [125], a robust rank fusion method.[16] CM creates a new ranking for each user by fusing two lists of scores: one based on relevance and another based on exposure. We min-max normalise the predicted relevance score to get the relevance-based score. We create the fairness-based scores from the coverage of each top-k' item, i.e., based on how many times the item appears in the top-k. We then compute 1 minus the normalised coverage to assign a higher score for less exposed items, thus increasing fairness. After the two lists of scores are fused, they are sorted to generate the final ranking that considers both item relevance and item fairness.

Measures

We evaluate the models with all relevance-aware fairness measures (Joint measures) presented in §4.2.3 and their corrected versions if they exist (§4.4). For IAA, we use the ground truth relevance as item relevance. For $\text{IFD}_{\div\text{-our}}$, we define $\text{IFD}_{\div\text{-our}}(u) = 0$ if u only has 1 relevant item in the test split. We set $\gamma = 0.9$ for HD as per [103] while $\gamma = 0.8$ for II-F and AI-F, following [234]. We rescale IBO/IWO into [0,1] to

[15] We provide the hyperparameter search space and the best values in the code repository.

[16] Other re-rankers exist but do not fit our setup, e.g., item similarity computation is needed [227], but true similarity is hard to obtain [63, 216].

be consistent with other measures' scale. For comparison to the JOINT measures, we evaluate recommendation effectiveness (EFF) with: Hit Rate (HR), Mean Reciprocal Rank (MRR), Precision (P), Recall (R), Mean Average Precision (MAP), and NDCG. We also evaluate fairness (FAIR) with measures that consider only item exposure (without factoring in item relevance): Jain Index (Jain) [101, 264], Qualification Fairness (QF) [264], Entropy (Ent) [167, 200], Fraction of Satisfied Items (FSat) [167], and Gini Index (Gini) [85, 147]. We compute the corrected FAIR measures as per Rampisela et al. [177]. We compute all measures at $k = 10$, unless stated otherwise.

4.6 Empirical Analysis

We evaluate recommendation performance with all EFF, FAIR, and JOINT measures (§4.6.1) and study their correlations (§4.6.2). We compare the achievable scores of the original JOINT measures and our corrections (§4.6.3). We analyse their sensitivity to various cut-offs (§4.6.4), varying number of missing relevance labels (§4.6.5), and different recommendation effectiveness and exposure distribution (§4.6.6).

4.6.1 Evaluation Results of All Measures

We compute all EFF and FAIR measures (i.e., single-aspect measures), as well as both the original and corrected JOINT measures. We present all scores per dataset and model in Tab. 4.7. We denote higher-is-better measures with ↑ and vice versa with ↓.

Often, the purpose of evaluation is to identify the best model in terms of effectiveness, fairness, or other aspects, so that one can choose which model to deploy. However, these aspects can be measured by one or more measures, and they may or may not reach the same conclusion regarding the best model. Here, we study if the measures agree on the best model and we further analyse the measure agreement when they rank recommender models in §4.6.2. We also analyse the score range as an indicator of the measure expressiveness.

Best model agreement

Do the measures agree on the same best model? We answer this question by comparing the agreement on the best model: (i) between EFF and FAIR measures; (ii) among the JOINT measures; as well as (iii) between the original and corrected JOINT measures.

Table 4.7: Effectiveness (EFF), fairness (FAIR), and JOINT scores at $k = 10$ without and with re-ranking the top 25 items using COMBMNZ (CM). Bold indicates the most effective/fair score per measure. The score 0.000 does not mean that the scores are exactly 0; this is because the scores are small ($< 10^{-3}$) and rounded to 3 d.p. The term 'nan' stands for "not a number".

		model	ItemKNN		BPR		MultiVAE		NCL	
		re-ranker	-	CM	-	CM	-	CM	-	CM
Lastfm	EFF	↑ HR	0.765	0.581	0.773	0.587	0.778	0.523	**0.793**	0.571
		↑ MRR	0.484	0.270	0.492	0.280	0.476	0.232	**0.503**	0.260
		↑ P	0.172	0.089	0.178	0.092	0.176	0.076	**0.184**	0.087
		↑ MAP	0.137	0.053	0.141	0.058	0.138	0.045	**0.148**	0.050
		↑ R	0.218	0.114	0.224	0.119	0.224	0.098	**0.234**	0.110
		↑ NDCG	0.245	0.119	0.252	0.126	0.247	0.102	**0.261**	0.115
	FAIR	↑ Jain	0.042	0.094	0.058	0.140	0.097	**0.222**	0.082	0.215
		↑ QF	0.474	**0.679**	0.362	0.528	0.517	0.678	0.453	0.657
		↑ Ent	0.589	0.735	0.610	0.740	0.707	**0.826**	0.671	0.810
		↑ FSat	0.129	0.216	0.147	0.228	0.202	**0.321**	0.178	0.286
		↓ Gini	0.904	0.790	0.910	0.818	0.839	**0.696**	0.872	0.728
	JOINT	↑ IBO_{ori}	nan	nan	nan	nan	nan	nan	nan	nan
		↑ IBO_{our}	0.209	0.256	0.208	0.253	0.261	0.278	0.242	**0.292**
		↓ IWO_{ori}	nan	nan	nan	nan	nan	nan	nan	nan
		↓ IWO_{our}	0.791	0.744	0.792	0.747	0.739	0.722	0.758	**0.708**
		↓ IAA_{ori}	**0.004**	**0.004**	**0.004**	**0.004**	**0.004**	**0.004**	**0.004**	**0.004**
		↓ IAA_{our}	0.764	0.887	0.757	0.881	0.762	0.903	**0.749**	0.894
		↓ $IFD_{\div\text{-ori}}$	0.074	0.053	0.075	0.054	0.073	**0.049**	0.076	0.052
		↓ $IFD_{\div\text{-our}}$	0.442	0.241	0.453	0.250	0.445	**0.205**	0.464	0.232
		↓ $IFD_{\times\text{-ori}}$	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
		↓ $IFD_{\times\text{-our}}$	0.275	0.126	0.282	0.134	0.273	**0.107**	0.290	0.122
		↓ HD_{ori}	0.099	0.177	0.104	0.174	0.095	0.203	**0.092**	0.177
		↓ MME_{ori}	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
		↓ $II\text{-}F_{ori}$	**0.001**	0.002	**0.001**	0.002	**0.001**	0.002	**0.001**	0.002
		↓ $II\text{-}F_{our}$	0.755	0.884	0.748	0.878	0.753	0.901	**0.739**	0.890
		↓ $AI\text{-}F_{ori}$	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Amazon-lb	EFF	↑ HR	**0.046**	0.016	0.011	0.021	0.039	0.014	0.034	0.011
		↑ MRR	0.020	0.011	0.003	0.007	**0.023**	0.004	0.022	0.003
		↑ P	**0.005**	0.002	0.001	0.002	0.004	0.002	0.004	0.001
		↑ MAP	**0.006**	0.004	0.002	0.004	**0.006**	0.003	**0.006**	0.001
		↑ R	**0.013**	0.005	0.005	0.010	0.010	0.008	0.012	0.003
		↑ NDCG	**0.011**	0.005	0.003	0.006	0.010	0.004	**0.011**	0.002
	FAIR	↑ Jain	0.271	**0.431**	0.223	0.359	0.035	0.097	0.026	0.080
		↑ QF	**0.650**	0.612	0.549	0.594	0.222	0.286	0.229	0.310
		↑ Ent	0.802	**0.839**	0.747	0.809	0.418	0.558	0.371	0.534
		↑ FSat	0.370	**0.438**	0.314	0.376	0.114	0.152	0.091	0.138
		↓ Gini	0.665	**0.598**	0.747	0.660	0.949	0.899	0.959	0.910
	JOINT	↑ IBO_{ori}	nan	nan	nan	nan	nan	nan	nan	nan
		↑ IBO_{our}	**0.062**	0.029	0.019	0.038	0.029	0.029	0.038	0.024
		↓ IWO_{ori}	nan	nan	nan	nan	nan	nan	nan	nan
		↓ IWO_{our}	**0.938**	0.971	0.981	0.962	0.971	0.971	0.962	0.976
		↓ IAA_{ori}	**0.011**	**0.011**	**0.011**	**0.011**	**0.011**	**0.011**	**0.011**	**0.011**
		↓ IAA_{our}	**0.988**	0.995	0.996	0.994	0.990	0.994	0.989	0.998
		↓ $IFD_{\div\text{-ori}}$	0.005	0.003	0.003	**0.002**	0.005	**0.002**	0.005	0.003
		↓ $IFD_{\div\text{-our}}$	0.016	0.006	**0.001**	0.005	0.018	**0.001**	0.016	0.004
		↓ $IFD_{\times\text{-ori}}$	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
		↓ $IFD_{\times\text{-our}}$	0.010	0.006	0.002	0.005	0.012	0.003	0.012	**0.001**
		↓ HD_{ori}	**0.580**	0.630	0.661	0.626	0.597	0.653	0.598	0.667
		↓ MME_{ori}	**0.001**	**0.001**	**0.001**	**0.001**	0.003	**0.001**	0.004	**0.001**
		↓ $II\text{-}F_{ori}$	**0.006**	**0.006**	**0.006**	**0.006**	**0.006**	**0.006**	**0.006**	**0.006**
		↓ $II\text{-}F_{our}$	**0.989**	0.995	0.997	0.995	0.990	0.995	0.990	0.998
		↓ $AI\text{-}F_{ori}$	**0.000**	**0.000**	**0.000**	**0.000**	0.001	**0.000**	0.002	**0.000**

Table 4.7 (continued): Effectiveness (Eff), fairness (Fair), and Joint scores at $k = 10$ without and with re-ranking the top 25 items using COMBMNZ (CM). Bold indicates the most effective/fair score per measure. The score 0.000 does not mean that the scores are exactly 0; this is because the scores are small ($< 10^{-3}$) and rounded to 3 d.p. The term 'nan' stands for "not a number".

		model	ItemKNN		BPR		MultiVAE		NCL	
		re-ranker	-	CM	-	CM	-	CM	-	CM
QK-video	Eff	↑ HR	0.040	0.047	0.099	0.045	0.109	0.061	**0.130**	0.077
		↑ MRR	0.013	0.013	0.039	0.015	0.039	0.021	**0.048**	0.024
		↑ P	0.004	0.005	0.011	0.005	0.012	0.006	**0.014**	0.008
		↑ MAP	0.005	0.005	0.017	0.006	0.018	0.009	**0.022**	0.010
		↑ R	0.014	0.019	0.043	0.019	0.051	0.027	**0.061**	0.033
		↑ NDCG	0.009	0.010	0.029	0.011	0.031	0.016	**0.038**	0.019
	Fair	↑ Jain	0.483	**0.589**	0.081	0.379	0.012	0.032	0.020	0.071
		↑ QF	**0.901**	0.790	0.625	0.823	0.100	0.163	0.201	0.365
		↑ Ent	0.933	**0.937**	0.755	0.903	0.420	0.547	0.507	0.674
		↑ FSat	0.443	**0.547**	0.212	0.382	0.052	0.090	0.077	0.150
		↓ Gini	0.472	**0.442**	0.807	0.570	0.982	0.959	0.966	0.902
	Joint	↑ IBO_{ori}	nan	nan	nan	nan	nan	nan	nan	nan
		↑ IBO_{our}	0.033	0.038	**0.054**	0.036	0.031	0.036	0.043	**0.054**
		↓ IWO_{ori}	nan	nan	nan	nan	nan	nan	nan	nan
		↓ IWO_{our}	0.967	0.962	**0.946**	0.964	0.969	0.964	0.957	**0.946**
		↓ IAA_{ori}	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
		↓ IAA_{our}	0.991	0.989	0.970	0.989	0.966	0.983	**0.959**	0.980
		↓ $IFD_{\div\text{-ori}}$	0.009	**0.007**	0.014	0.008	0.014	0.009	0.015	0.010
		↓ $IFD_{\div\text{-our}}$	**0.015**	**0.015**	0.040	0.016	0.040	0.022	0.049	0.026
		↓ $IFD_{\times\text{-ori}}$	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
		↓ $IFD_{\times\text{-our}}$	**0.007**	**0.007**	0.023	0.008	0.023	0.012	0.028	0.014
		↓ HD_{ori}	0.576	0.560	0.490	0.565	0.478	0.535	**0.457**	0.519
		↓ MME_{ori}	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
		↓ $II\text{-}F_{ori}$	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**	**0.001**
		↓ $II\text{-}F_{our}$	0.992	0.990	0.973	0.990	0.970	0.985	**0.964**	0.983
		↓ $AI\text{-}F_{ori}$	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
ML-10M	Eff	↑ HR	0.487	0.443	0.512	0.386	0.417	0.387	**0.521**	0.402
		↑ MRR	0.282	0.225	0.299	0.185	0.237	0.191	**0.302**	0.203
		↑ P	0.137	0.105	0.146	0.088	0.107	0.096	**0.154**	0.094
		↑ MAP	0.089	0.060	0.095	0.047	0.067	0.054	**0.101**	0.052
		↑ R	0.022	0.018	0.025	0.012	0.020	0.016	**0.026**	0.013
		↑ NDCG	0.150	0.113	0.160	0.092	0.119	0.100	**0.167**	0.100
	Fair	↑ Jain	0.011	0.027	0.037	**0.115**	0.003	0.006	0.024	0.069
		↑ QF*	0.044	0.068	0.145	**0.216**	0.014	0.025	0.086	0.132
		↑ Ent	0.407	0.514	0.596	**0.716**	0.238	0.324	0.519	0.638
		↑ FSat*	0.044	0.068	0.145	**0.216**	0.014	0.025	0.086	0.132
		↓ Gini	0.987	0.971	0.945	**0.879**	0.997	0.993	0.969	0.930
	Joint	↑ IBO_{ori}	nan	nan	nan	nan	nan	nan	nan	nan
		↑ IBO_{our}	0.031	0.046	0.069	**0.091**	0.012	0.018	0.054	0.074
		↓ IWO_{ori}	nan	nan	nan	nan	nan	nan	nan	nan
		↓ IWO_{our}	0.969	0.954	0.931	**0.909**	0.988	0.982	0.946	0.926
		↓ IAA_{ori}	**0.008**	0.009	**0.008**	0.009	0.009	0.009	**0.008**	0.009
		↓ IAA_{our}	0.850	0.887	0.840	0.910	0.880	0.899	**0.830**	0.902
		↓ $IFD_{\div\text{-ori}}$	0.018	0.012	0.019	0.011	0.016	**0.010**	0.020	0.012
		↓ $IFD_{\div\text{-our}}$	0.164	0.126	0.176	**0.102**	0.132	0.110	0.185	0.109
		↓ $IFD_{\times\text{-ori}}$	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
		↓ $IFD_{\times\text{-our}}$	0.164	0.121	0.175	**0.097**	0.131	0.105	0.181	0.107
		↓ HD_{ori}	0.221	0.255	0.226	0.262	0.265	0.273	**0.218**	0.257
		↓ MME_{ori}	**0.001**	**0.001**	**0.001**	**0.001**	0.003	**0.001**	**0.001**	**0.001**
		↓ $II\text{-}F_{ori}$	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
		↓ $II\text{-}F_{our}$	0.846	0.885	0.836	0.908	0.878	0.898	**0.828**	0.900
		↓ $AI\text{-}F_{ori}$	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**

*QF ≡ FSat for ML-10M, as QF is computed based on the percentage of recommended items from all items, which in this case is equivalent to FSat.

Generally, all EFF measures agree on the best model across all four datasets, and likewise for the FAIR measures. In contrast, the JOINT measures tend to disagree on the best model more often than the single-aspect measures.

For each dataset, we observe the following. First, the best model according to EFF measures for each dataset always differs from the one given by FAIR measures, except for QF in Amazon-lb. This means that the most effective model is generally not the fairest. Second, even though the best model per dataset is always the same for all EFF measures (except MRR and MAP for Amazon-lb) and likewise for FAIR measures (except QF), this is not the case for the JOINT measures. Some JOINT measures always have the same best model per dataset (e.g., IBO_{our} with IWO_{our}, IAA with HD and II-F, or MME with AI-F), but the trend is less consistent for IFD, which only agrees with MME/AI-F occasionally for some datasets/measure versions. The JOINT measures that have the same best models tend to have similar formulations: both IBO/IWO are the percentage of items with an impact score higher/lower than a threshold; MME/AI-F aggregate exposure across users prior to taking the difference in exposure, while IAA/HD/II-F do not; and MME/IFD are pairwise measures. Third, the original IAA/II-F/$\text{IFD}_{\times}$ always agree with their corrected versions.[17] Yet, the original $\text{IFD}_{\div}$ mostly disagrees with its correction. This disagreement may be because $\text{IFD}_{\div\text{-ori}}$ assigns nonzero exposure weight for items beyond top-k, while $\text{IFD}_{\div\text{-our}}$ does not. In IAA/II-F/$\text{IFD}_{\times}$ the exposure weight assignment does not differ significantly between the original and corrected version.

In summary, we can cluster all JOINT measures into three groups based on their general alignment to single-aspect measures: (i) EFF-aligned measures, IAA/HD/II-F (ii) FAIR-aligned measures, IFD/MME/AI-F; and (iii) measures without consistent alignment, IBO/IWO. The JOINT measure alignment to single-aspect measure can somewhat be explained through their formulation. For example, the fairest ranking based on IAA and II-F is the one that places all relevant items at the top, which aligns with EFF measures. As for AI-F, the measure scores fairer when there are more unique items exposed to the users, which aligns with FAIR measures. We also find that most original JOINT measures have the same best models as the corrected versions, except for $\text{IFD}_{\div}$. Therefore, computing $\text{IFD}_{\div\text{-ori}}$ may lead to a different conclusion from $\text{IFD}_{\div\text{-our}}$. We revisit the measure agreement in §4.6.2.

[17] As this cannot be seen from the rounded scores in Tab. 4.7, we verify this from the pre-rounded scores.

Score range

We find three issues with the range of the original JOINT measure scores: (i) consistently small scores for several JOINT measures; (ii) scale mismatch between single-aspect measures and JOINT measures; and (iii) scale mismatch between the original JOINT measures.

Regarding (i), the extremely low scores of the original ↓JOINT measure scores ($\leq 10^{-3}$) for all datasets and models, make it hard to distinguish the performance of models for a specific dataset. For instance, $\text{IFD}_{\times\text{-ori}}$ always scores near 0 for all datasets, as Eq. (4.7) is often 0 due to the low number of relevant items per user.[18] For the same reason, MME_{ori} and II-F_{ori}/AI-F_{ori}, Eq. (4.16) and Eq. (4.23) often result in 0. These low scores may also be due to the **non-realisability** limitation, as the measures cannot reach the unfairest score of 1. Our versions of $\text{IFD}_{\times}$ and II-F do not suffer from this issue, as they have been normalised based on the achievable max/min score. Therefore, distinguishing the $\text{IFD}_{\times\text{-our}}$ and II-F_{our} scores across models for a dataset is easier than doing so for the original measures.

Regarding (ii), i.e., the scale mismatch between single-aspect measures and JOINT measures, while the above JOINT scores differ in the fourth decimal point or later, the EFF and FAIR scores differ in the second decimal point or before. For instance, the NDCG (EFF score) of MultiVAE-CM and NCL for Lastfm differs by ∼0.16 and their Jain (FAIR score) differs by ∼0.14. These differences are non-negligible, but the joint scores of IAA_{ori}/$\text{IFD}_{\times\text{-ori}}$/$\text{MME}_{\text{ori}}$/$\text{II-F}_{\text{ori}}$/$\text{AI-F}_{\text{ori}}$ differ only by $\leq 10^{-3}$, which may seem negligible.[19] As the difference of magnitude is inconsistent between the single-aspect and JOINT measures, the scores are harder to interpret. On the other hand, the differences in the scores of IAA_{our}/$\text{IFD}_{\times\text{-our}}$/$\text{II-F}_{\text{our}}$ per dataset can either be seen in the first or second decimal point, except for $\text{IFD}_{\times\text{-our}}$, which differs in the third decimal point for Amazon-lb and QK-video. Hence, our version of IAA/IFD/II-F is more expressive than the original version, as it better matches the change in the single-aspect measures.

Regarding (iii), i.e., the scale mismatch between the original JOINT measures, we observe considerable gaps in the range of all JOINT measure scores, e.g., between IWO_{our}, HD_{ori}, and AI-F_{ori}, despite all three being lower-is-better measures. For instance, in ML-10M, $\downarrow\text{IWO}_{\text{our}} \approx 1$ (very unfair) based on its theoretical $[0, 1]$-range, $\downarrow\text{HD}_{\text{ori}}$ is roughly a quarter of the $\downarrow\text{IWO}_{\text{our}}$ score (somewhat fair), and $\downarrow\text{AI-F}_{\text{ori}} \approx 0$ (extremely fair). This discrepancy causes confusion in score interpretation. Between

[18] The median number of relevant items per user in a dataset is at most 46, across all four datasets

[19] NDCG and Jain are used in the comparison as they tend to be stricter* than some other single-aspect measures used in this work.

some of the original and the reformulated measures, we also find huge differences. For example, across all models for ML-10M: $\downarrow$IAA$_{\text{ori}}$ $\approx$ 0.010 (very fair), while $\downarrow$IAA$_{\text{our}}$ $\approx$ 0.9 (very unfair). Similarly, $\downarrow$II-F$_{\text{ori}}$ $\approx$ 0.000 (very fair), while $\downarrow$II-F$_{\text{our}}$ $\approx$ 0.9 (very unfair). Similar gaps can also be observed for the same measure pairs in the other datasets. Thus, evaluation with IAA$_{\text{ori}}$ or II-F$_{\text{ori}}$ scores may result in a misleading conclusion of having extremely fair models, even if the models are actually very unfair. In contrast, evaluating with IAA$_{\text{our}}$ or II-F$_{\text{our}}$ leads to a more accurate conclusion.

Undefined scores

The scores of IBO$_{\text{ori}}$/IWO$_{\text{ori}}$ are undefined for all datasets and all models. Due to division by zero, the computation results in 'nan', which stands for 'not a number'. The measures are incomputable as some items are not relevant to any users, which causes Eq. (4.20) to be 0, and this is a manifestation of the **undefinedness** limitation (§4.3.3). Our reformulation of both measures (§4.4.2) resolves this issue.

4.6.2 Measure Correlations

When comparing evaluation measures, it is important to understand the extent to which they agree in ranking recommender models based on the measure scores. To this end, we compute Kendall's τ correlation as a measure of agreement between the orderings of the recommenders based on the measure scores computed in §4.6.1. We also report the statistical significance of the correlations ($\alpha = 0.05$) and apply Bonferroni's correction to account for multiple testing errors within a dataset. We present the result in Fig. 4.1. We study three types of agreement: (i) between Joint measures and single-aspect measures; (ii) among Joint measures in general; and (iii) between each original Joint measure and its corrected version.

(i) Agreement of joint and single-aspect measures

Eff measures are known to not correlate strongly with Fair measures [177], so we do not expect the Joint measures to correlate strongly with either Eff or Fair measures, as they should account for both aspects. Generally, we find that Joint measures do not correlate consistently with Eff measures. IBO/IWO's correlations largely vary ($\tau \in [-0.64, 0.77]$); IAA, HD, and II-F have positive correlations that are either moderate or strong ($\tau \in [0.57, 1]$); IFD and MME have negative correlations that range from weak to strong ($\tau \in [-1, -0.29]$ for IFD and $\tau \in [-0.79, -0.14]$ for MME); and AI-F has non-positive correlations ($\tau \in [-0.64, 0]$).

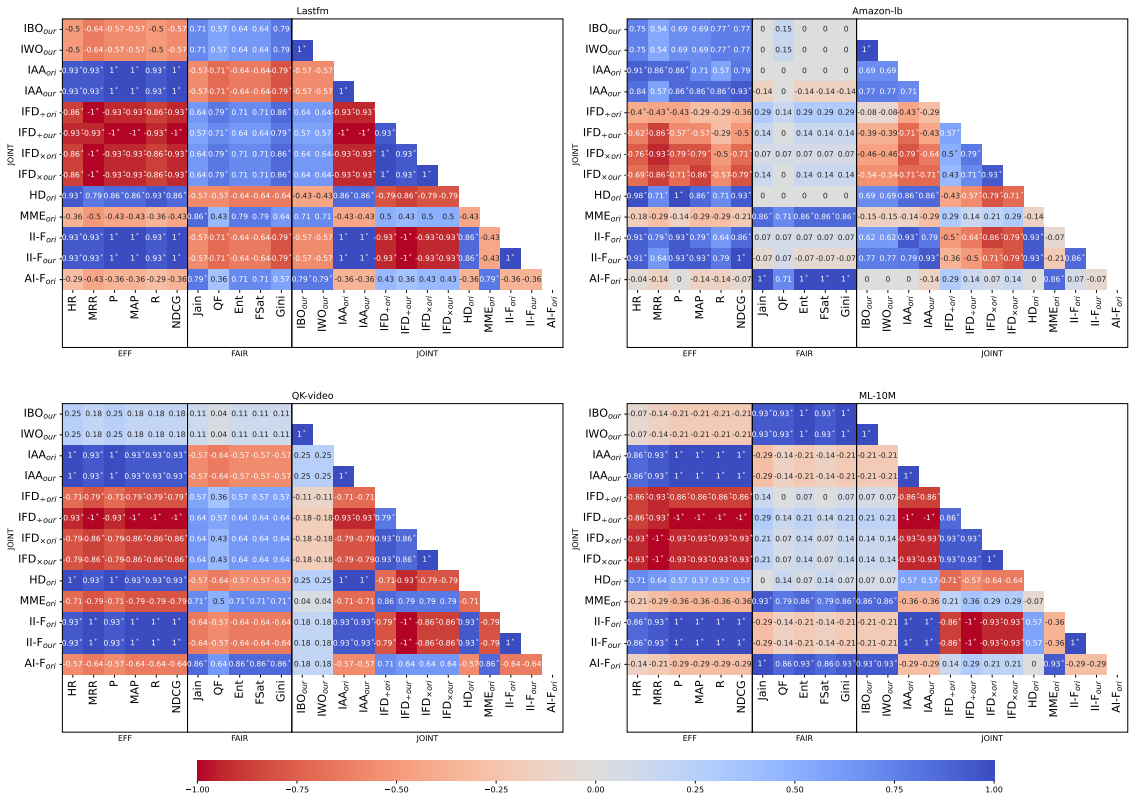


Figure 4.1: Kendall's τ correlation between Joint, Eff, and Fair measures. Asterisk (*) denotes a statistically significant correlation ($\alpha = 0.05$) after applying Bonferroni's correction.

We also find that the JOINT measures do not correlate consistently with FAIR measures either. Again, the correlations of IBO/IWO have large variance, though less than with EFF measures. IAA/HD/II-F shows two distinct trends across datasets: they have negative correlations that are moderate to strong ($\tau \in [-0.79, -0.57]$) for Lastfm and QK-video, but correlate weakly for Amazon-lb and ML-10M ($\tau \in [-0.29, 0.14]$). The correlations of IFD also vary across datasets. On one hand, IFD strongly correlates with FAIR measures for Lastfm and QK-video (except with QF for QK-video), $\tau \in [0.57, 0.86]$, and on the other hand, IFD weakly correlates to FAIR measures for the other datasets ($\tau \in [0, 0.29]$). For all datasets, MME and AI-F strongly correlate with FAIR measures, except for QF in Lastfm ($\tau \in [0.5, 1]$).

In summary, even though we expect the JOINT measures to have weak correlations with both EFF and FAIR measures as the JOINT measures consider both item relevance and exposure, this is only true for IBO/IWO. Note that JOINT measures with strong agreement to EFF measures do not always strongly disagree with FAIR measures, and vice versa. For instance, IAA/HD/II-F strongly agree with EFF measures for Amazon-lb, but their correlations with FAIR measures are weak.

(ii) Agreement among Joint measures.

We expect the agreement among JOINT measures to follow the groupings found in §4.6.1. Overall, the three JOINT measure groups identified in §4.6.1 indeed have strong positive correlations between measures of the same group and strong negative correlations between measures from different groups. For instance, IBO consistently correlates perfectly with IWO, as their formulation is similar. IAA, HD, and II-F strongly agree with each other, $\tau \in [0.57, 1]$. $\text{IFD}_{\div}$ correlates moderately or highly with $\text{IFD}_{\times}$, $\tau \in [0.43, 1]$ due to their similar formulations. MME and AI-F always correlate highly, $\tau \in [0.79, 0.93]$. IFD occasionally has moderate-to-strong correlations with MME and AI-F, $\tau \in [0.36, 0.86]$ for Lastfm and QK-video, but the correlations are weaker for Amazon-lb and ML-10M, $\tau \in [0.07, 0.36]$. Conversely, IAA/HD/II-F have moderate-to-strong disagreement with IFD, $\tau \in [-1, -0.36]$ except for $\text{IFD}_{\div}$ in Amazon-lb ($\tau = -0.29$).

From the findings of (i) and (ii), we conclude that: IBO/IWO has inconsistent relationships with single-aspect and JOINT measures; IAA/HD/II-F strongly disagrees with FAIR measures; and IFD/MME/AI-F strongly disagrees with EFF measures (even if IFD occasionally disagrees with FAIR measures as well). Among the JOINT measures, IBO/IWO has weak correlations with the single-aspect measures for QK-video, and similarly with $\text{IFD}_{\div}$ for Amazon-lb, but the trend is inconsistent. Thus,

we argue that no JOINT measures reliably account for both recommendation effectiveness and fairness, where fairness is quantified purely based on exposure.

(iii) Agreement between original and corrected Joint measures

We expect each original JOINT measure to strongly correlate with its respective corrected version, as most (but not all) corrections are linear transformations. As the corrected measures min-max normalises the original JOINT score per user prior to averaging across users, the original measures may not perfectly correlate to their corrections. In general, the original JOINT measures strongly agree with the corrected ones ($\tau \in [0.57, 1]$), and the correlations tend to be weaker in Amazon-lb. We cannot compare $\text{IBO}_{\text{ori}}/\text{IWO}_{\text{ori}}$ with $\text{IBO}_{\text{our}}/\text{IWO}_{\text{our}}$ as the original version returns undefined scores for all datasets and models (§4.6.1). IAA_{ori}, $\text{IFD}_{\times\text{-ori}}$, and II-F_{ori} perfectly correlate with their respective corrected versions for all datasets other than Amazon-lb, where $\tau \in [0.71, 0.93]$. $\text{IFD}_{\div\text{-ori}}$ have weaker (but still strong) correlations to $\text{IFD}_{\div\text{-our}}$ ($\tau \in [0.57, 0.93]$). We posit that this is because correcting the top-k-insensitivity of $\text{IFD}_{\div\text{-ori}}$ results in a non-linear transformation of the combined exposure-relevance score in Eq. (4.6).

In conclusion, while the original measures strongly agree with the corrected ones, they do not always give identical rank orderings of the models. This means that in some cases, by computing the original measure, we may reach different conclusions on the relative performances of the model as opposed to computing the corrected version.

4.6.3 Achievable Fairness Scores (Related to Non-realisability)

Given the (un)fairest possible ranking, to what extent can the JOINT measures achieve their theoretical fairest or unfairest score (0 or 1) for different datasets and different cut-off k? Here, we investigate: (i) the severity of the **non-realisability** limitation (*Cause 4*) in the original JOINT measures; and (ii) the extent to which the corrected measures resolve the limitation. As such, we experiment only with the JOINT measures for which we have resolved the limitation, i.e., IAA, $\text{IFD}_{\div}$, $\text{IFD}_{\times}$, and II-F, and evaluate recommendations at various cut-offs, $k \in \{1, 3, 5, 10\}$. To generate the recommendations, we employ three strategies: `top` (rank all relevant items at the top), `bottom` (rank all relevant items at the bottom), `half-half` (rank half of the relevant items at the top and rank the rest at the bottom). Each strategy corresponds to either the fairest or unfairest recommendation as per Tab. 4.4,

which differs per measure. Specifically, `top` corresponds to the fairest ranking for IAA and II-F and the unfairest for $\text{IFD}_{\times}$,[20] `bottom` corresponds to the fairest ranking for IFD and the unfairest for IAA and II-F, and `half-half` corresponds to the unfairest ranking for $\text{IFD}_{\div\text{-our}}$. For every measure, we employ two strategies, that each correspond to the measure's fairest and unfairest ranking. We show the results in Figs. 4.2 and 4.3.

Fairest achievable score (Fig. 4.2)

All of our corrected versions of the measures in this experiment always successfully reach the theoretical fairest score (i.e., 0) for all datasets and choices of k. On the other hand, the original Joint measures of IAA, $\text{IFD}_{\div}$, and II-F always fail to reach the fairest score (i.e., 0) due to the non-realisability limitation. While the scores may seem to be close to 0, they may cause confusion as they are not exactly 0, even if the recommendation cannot be fairer than that. Only $\text{IFD}_{\times\text{-ori}}$ manages to achieve 0, and this is because the number of relevant items per user ($|R_u^*|$) is much lower than $n-k$, which is the number of unexposed items per user. If $|R_u^*| > n-k$ instead, like the example in §4.3.1, $\text{IFD}_{\times\text{-ori}}$ may fail to reach 0, as an item that is both relevant and recommended at the top-k would contribute a nonzero score. Among IAA, $\text{IFD}_{\div}$, and II-F, the measure with the highest difference between the original and the corrected scores is IAA, followed by II-F and $\text{IFD}_{\div}$. For IAA and II-F, this difference changes across various k, but there is no consistent trend across datasets. Consequently, the scores are more unpredictable and harder to interpret as the lower bound changes with different k. $\text{IFD}_{\div\text{-ori}}$ remains constant for different k as it suffers from the **top-k-insensitivity** limitation, though in this case, achieving the same score for various k is arguably the more desired case as the fairest achievable score should be 0, no matter the k. Further, IAA_{ori} is incomputable for $k=1$ due to the **undefinedness** limitation, as setting $k=1$ causes division by zero in the original, normalised linear examination function (Tab. 4.2). IAA_{our} does not suffer from this issue.

Unfairest achievable score (Fig. 4.3)

All original Joint measures in this experiment never reach the theoretical unfairest score (i.e., 1) for all combinations of dataset and k. Even more worrying, the measures' achievable unfairest scores are all less than 0.25, which is very far from 1, and

[20] For $\text{IFD}_{\times}$, we experimented with the two strategies described in §4.4.1.3 and found that the `top` strategy results in an unfairer score than the other.

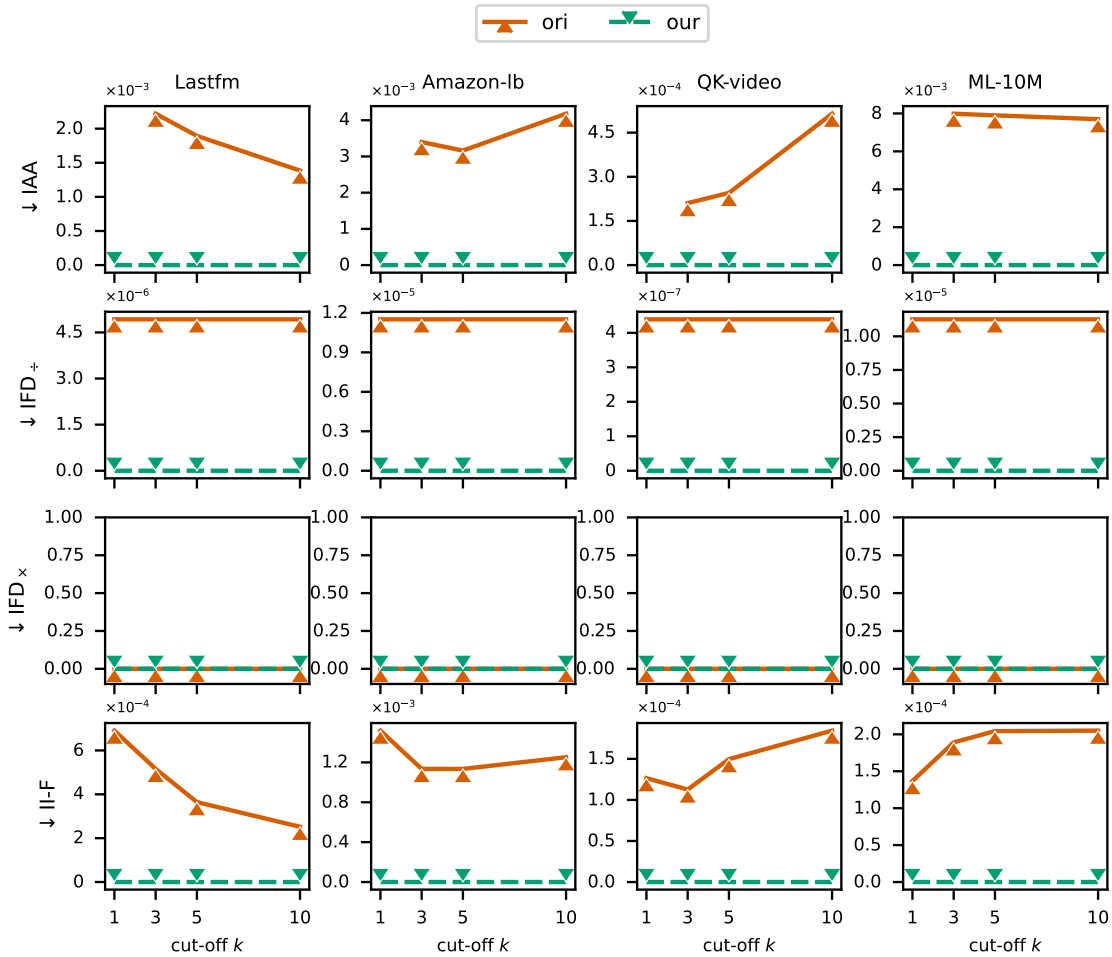


Figure 4.2: The fairest achievable ↓JOINT measure scores for varying k, both the original (ori) and corrected (our) versions. IAA_{ori} is incomputable at $k = 1$ due to the undefinedness limitation. Both $\text{IFD}_{\times}$ versions overlap.

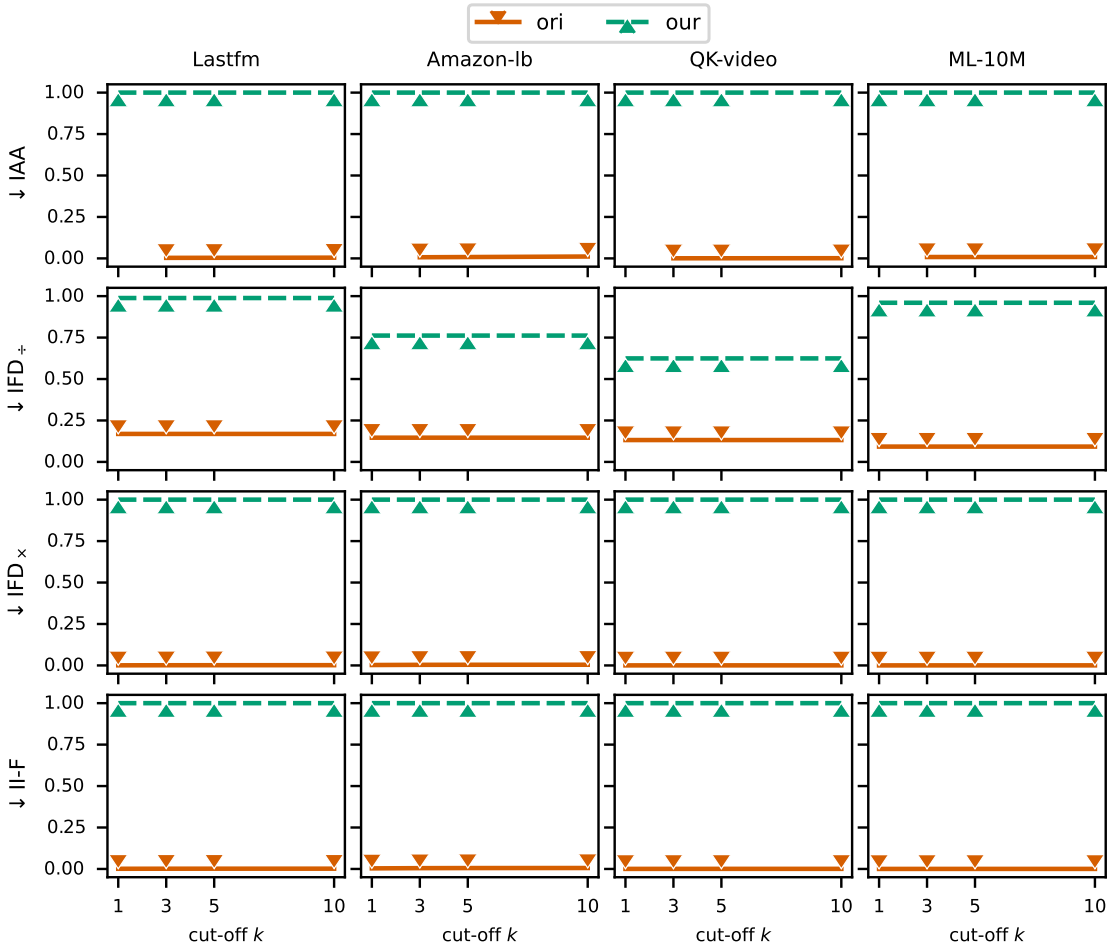


Figure 4.3: The unfairest achievable ↓JOINT measure scores for varying k, both the original (ori) and corrected (our) versions. IAA_{ori} is incomputable at $k = 1$ due to the undefinedness limitation.

much closer to the theoretical fairest score (i.e., 0) instead. In contrast, all of our corrected measures always achieve 1, except for $\text{IFD}_{\div}$. $\downarrow\text{IFD}_{\div\text{-our}}$ cannot reach 1 because the datasets have some users with only one relevant item in the test set. In our experiment setup, we set $\downarrow\text{IFD}_{\div\text{-our}}(u) = 0$ for all users with only one relevant item, as this measure computes the pairwise difference of the combined exposure-relevance score from the relevant items. If these users are excluded from the evaluation, the score of $\text{IFD}_{\div\text{-our}}$ would be 1. For different k values, the scores of both the original and our versions of the measures do not vary, which is a desirable property, as the unfairest score should be constant for different cut-offs.[21] Other than issues related to the non-realisability of the original measures, we also observe that IAA_{ori} cannot be computed at $k = 1$ due to the **undefinedness** limitation, but this issue is resolved in IAA_{our}.

Based on the above results, we find that the empirical ranges of the original version of IAA, IFD, and II-F are constrained to a small part of their theoretical range. As their unfairest achievable scores are extremely close to the theoretical fairest score, any scores given by the measures would be deceptively fair, leading to challenges in interpreting the scores. We show that all of our corrected measures resolve the compressed range issue. The scores of the corrected measures, except for $\text{IFD}_{\div\text{-our}}$, would also be easier to interpret, as 0 is mapped to the fairest possible recommendation, and 1 is mapped to the unfairest possible recommendation, unlike IAA_{ori}, $\text{IFD}_{\times\text{-ori}}$, and II-F_{ori}, which has scores other than 0 or 1 mapped to the (un)fairest scores. Therefore, the corrected measures can successfully reach the theoretical max/min score, except for $\text{IFD}_{\div\text{-our}}$, which fails to achieve the theoretical max score unless all users with only one relevant item are excluded from the evaluation.

4.6.4 Sensitivity to Cut-off k (Related to Top-k-insensitivity)

This experiment aims: (i) to show how $\text{IFD}_{\div\text{-ori}}$, which suffers from the **top-k-insensitivity** limitation, is unaffected by the choice of k; and (ii) to investigate whether the corrected version, $\text{IFD}_{\div\text{-our}}$ resolves this issue. In this experiment, we evaluate the runs from the NCL model at $k \in \{1, 3, 5, 10, 20\}$. We choose the NCL runs as they generally perform the best in terms of effectiveness. We compute the following measures: $\text{IFD}_{\div\text{-ori}}$, $\text{IFD}_{\div\text{-our}}$, and $\text{IFD}_{\div\text{-ori}} + J_{\text{our}}$, i.e., $\text{IFD}_{\div\text{-ori}}$ computed with $J_{\div\text{-our}}$ (Eq. (4.27)). The latter is $\text{IFD}_{\div\text{-ori}}$ that resolves only top-k-insensitivity, as opposed to $\text{IFD}_{\div\text{-our}}$, which resolves both top-k-insensitivity and non-realisability.

We present the results in Fig. 4.4. For all datasets, $\text{IFD}_{\div\text{-ori}}$ is invariant to the choice of cut-off, while the two corrected versions vary with different k. The score

[21] We verify this from the actual measure scores, and not just through the visualisation in Fig. 4.3.

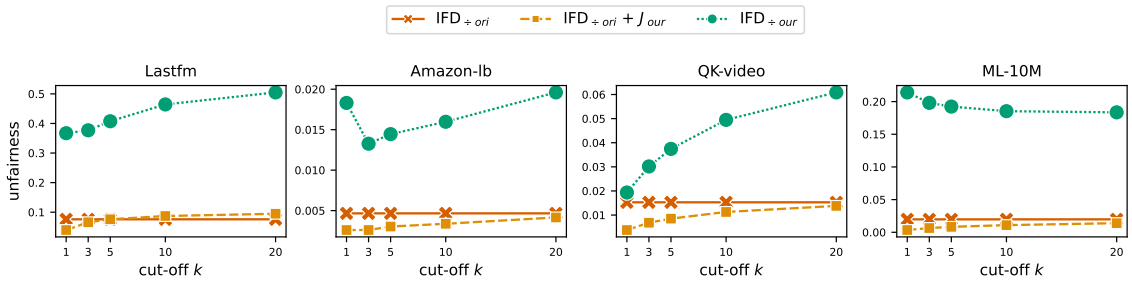


Figure 4.4: The original and corrected versions of IFD$_{\div}$ across different cut-offs on the runs from the NCL model.

changes are even visible in the non-rescaled corrected version (IFD$_{\div\text{-ori}} + J_{\text{our}}$), albeit to a smaller extent compared to those in IFD$_{\div\text{-our}}$. Among the corrected versions, there are differences in both the score magnitude and the score change, given different k. The scores of of IFD$_{\div\text{-our}}$ is always higher, which means that evaluating with IFD$_{\div\text{-ori}}+J_{\text{our}}$ would lead to an underestimation of the recommendation fairness. For IFD$_{\div\text{-ori}}+J_{\text{our}}$, unfairness generally increases with a larger k. The same trend applies to IFD$_{\div\text{-our}}$ as well, for Lastfm and QK-video. However, as k increases, ↓IFD$_{\div\text{-our}}$ decreases for ML-10M. For QK-video, ↓IFD$_{\div\text{-our}}$ first decreases, then increases. This means that for Amazon-lb and ML-10M, the two corrected IFD$_{\div}$ versions show different behaviour with increasing k, e.g., recommending the top-20 items is fairer than only recommending the top-1 item for IFD$_{\div\text{-our}}$, but the opposite applies for IFD$_{\div\text{-ori}}$ + J_{our}.

We posit that the findings for ↓IFD$_{\div\text{-our}}$ differ across datasets due to the disparity in the effectiveness scores across datasets (Tab. 4.7). The effectiveness disparity across datasets means that there are more relevant items at the top-k in one dataset than the other, and the difference in the rank positions of the relevant items affects the ↓IFD$_{\div\text{-our}}$ scores. In addition, while IFD$_{\div\text{-ori}}$ + J_{our} behaves monotonically, IFD$_{\div\text{-our}}$ does not. This is because IFD$_{\div\text{-our}}$ averages the min-max normalised IFD$_{\div\text{-ori}}$ + J_{our} score per user instead of directly min-max normalising the final IFD$_{\div\text{-ori}}$ + J_{our}. Therefore, IFD$_{\div\text{-our}}$ is not expected to show the same monotonic trend as IFD$_{\div\text{-ori}}$ + J_{our}.

To conclude, we find that IFD$_{\div\text{-ori}}$ is insensitive to changes in the cut-off, and the corrected measure resolves this issue. Our results also show that while simply reformulating the exposure-relevance combination function $J_{\div}$ is sufficient to resolve the top-k-insensitivity limitation, leaving the non-realisability limitation unresolved would lead to differing conclusions that may affect the perceived recommendation fairness (based on the score magnitude). The difference in the conclusion also con-

cerns the choice of k that would result in the fairest recommendation, which may be important in deciding how many items to recommend to each user.

4.6.5 Effect of Relevant Items Beyond the Top-k (Related to Non-localisation)

In RS datasets, there are typically a lot of items whose relevance to a specific user is unknown, as the item has not been shown to the user, or the user did not interact with it when it was shown. We refer to such items as 'unobserved items'. During evaluation, these items are usually considered irrelevant, while they can actually be relevant. Additionally, some items could also be misrated, e.g., they are rated lower than intended due to misclick. In practice, items that are rated lower than a threshold may be discarded during preprocessing. Consequently, these items are regarded as irrelevant, even though the opposite could be true. Here, we investigate whether treating unobserved items or potentially misrated items as either relevant or irrelevant affects the Joint measure scores.

IAA, IFD$_{\div}$, HD, IBO, IWO, II-F, and AI-F suffer from the **non-localisation** limitation (Tab. 4.3), i.e., they require item relevance information beyond the top-k. While we do not resolve this limitation, we study how additional relevant items beyond the top-k affect the Joint measure scores. Quantifying the effect of additional relevant items is important to understand the Joint measure robustness against changing the number of relevant items for a user, even if these extra items are not recommended at the top-k. For example, if there are actually more relevant items, how much does the initial score underestimate or overestimate the recommendation fairness?

To answer the above questions, we use the recommendations generated by the NCL model, which overall has the best Eff scores across all datasets. We simulate the additional relevant items by treating an unobserved item or a low-rated item as 'relevant', one item per user at a time. We do this iteratively until each user has ten additional relevant items. Only the items ranked beyond the top-k are considered, i.e., items at position $k+1, k+2, \ldots, n-1, n$. We employ two strategies to choose the 10 items: the first 10 irrelevant items, starting from position $k+1, k+2, \ldots$ (`top`) and the last 10 irrelevant items, starting from position $n, n-1, \ldots$ (`bottom`). We compute at $k = 10$ the measures that suffer from non-localisation: IWO, IAA, IFD$_{\div}$, HD, II-F, and AI-F. IBO also suffers from non-localisation, but we do not compute it, as it correlates perfectly with IWO (§4.6.2). We compute both the original measures and their reformulations, if the latter exists. Even if the reformulated measures do not aim to resolve non-localisation, the compressed empirical range of the original

measures (§4.6.3) may mislead the result interpretations. For example, if the original measure scores only vary a little given additional relevant items, it would seem that the additional relevant items do not considerably affect the scores.

We present the experiment results in Fig. 4.5 and compare the effect of additional relevant items across datasets, measures, and item selection strategy. IWO_{ori} scores are unavailable for all datasets due to the **undefinedness** limitation.

Effect across datasets

The measure scores for ML-10M are generally more stable when adding the number of relevant items, compared to the scores for other datasets. This may be due to ML-10M having more relevant items per user (with a median of 46) than the other datasets, which have a median of 8 or less. As most RS datasets are sparse, each user may not have many relevant items. Thus, the measure scores would be more susceptible to the effect of additional relevant items.

Effect across measures

The measures respond differently to additional relevant items: (i) II-F_{ori} and AI-F_{ori} seem to be mostly stable (their maximum change is less than 10^{-3}); (ii) IWO_{our}, IAA_{ori}, IAA_{our}, and II-F_{our} overall become more unfair; and (iii) $\text{IFD}_{\div\text{-ori}}$, $\text{IFD}_{\div\text{-our}}$, and HD_{ori} generally become fairer.

About (i), the apparent stability of II-F_{ori} and AI-F_{ori} may be due to their compressed empirical range, as II-F_{our}, which has a more reasonable empirical range, exhibits a different trend. For all datasets, the range of II-F_{ori} across varying the number of additional relevant items is one or two magnitudes smaller than that of II-F_{our}.

About (ii), we explain why IWO_{our}, IAA, and II-F_{our} become more unfair. Each relevant item beyond the top-k has a zero exposure weight (as it is unexposed) and a relevance score of 1 (as it is relevant). For ↓IAA, adding more relevant items beyond the top-k results in a higher difference between the item exposure weight and item relevance value. For ↓II-F, changing the item relevance from 0 to 1 would increase the item target exposure (Eq. (4.23)). As a result, the difference between the item exposure and target exposure increases. For ↓IWO, when there are more users for whom an item i is relevant, Imp_i^{unif} (Eq. (4.20)) increases. Imp_i^{unif} is the upper bound of the inequality in Eq. (4.36). When this upper bound increases, the inequality can be satisfied by more items, resulting in a more unfair IWO.

About (iii), $\downarrow\text{IFD}_{\div}$ becomes fairer as having a relevant item beyond the top-k lowers the average pairwise difference between the combined item exposure-relevance

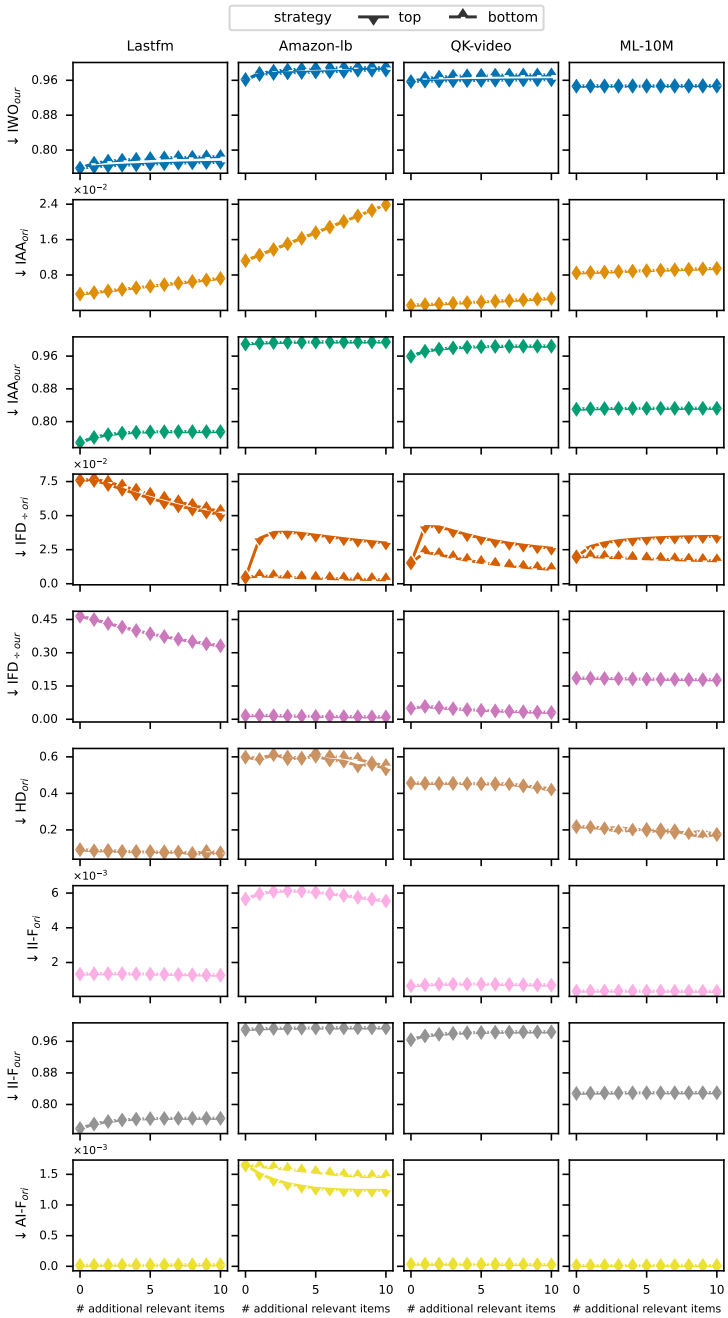


Figure 4.5: JOINT measure scores for different numbers of artificially added relevant items per user. We treat unobserved items or low-rated items as 'relevant', starting from position $k+1$ downwards (`top`) or from position n upwards (`bottom`).

score per user. Next, we explain why $\downarrow$HD$_{\text{ori}}$ also decreases. HD$_{\text{ori}}$ becomes fairer when the difference between the normalised item relevance at p (q'_p, Eq. (4.9)) and normalised item interaction probability at p (c'_p, Eq. (4.10)) becomes smaller, for $p \leq k$. As such, the value of c'_p is unaffected by additional relevant items. However, q'_p decreases, as the user-wise normalised relevance, $r'_{u,i} = r_{u,i} / \sum_{i \in I} r_{u,i}$, decreases given additional relevant items per user. Consequently, $\downarrow$HD becomes fairer.

More interestingly, for some measures, we note that having fewer additional relevant items causes higher fluctuations than having a greater amount of additional relevant items. For example, having 1–3 additional relevant items per user causes a higher score deviation from the score for no additional items, compared to the deviation caused by 8–10 additional items per user. This is evident for IFD$_{\div\text{-ori}}$, II-F$_{\text{ori}}$, and HD$_{\text{ori}}$, particularly for Amazon-lb. For $\downarrow$IFD$_{\div\text{-ori}}$, this is because the measure takes the mean pairwise difference of the combined exposure-relevance value and the exposure weight decreases down the rank positions. Thus, the mean pairwise difference decreases with a greater number of additional relevant items beyond the top-k. For $\downarrow$HD$_{\text{ori}}$, this may be due to both the treatment of items with identical relevance by the unstable sorting method and the same reason for increased fairness explained in the previous paragraph (i.e., the normalised item relevance decreases, while the normalised item interaction probability remains constant). For $\downarrow$II-F$_{\text{ori}}$, given additional relevant items, the target exposure (Eq. (4.23)) decreases and the system exposure (Eq. (4.22)) remains constant as it is unaffected by item relevance. Hence, the difference between the target and system exposure becomes smaller, resulting in a lower final score.

Given the high likelihood of having many unobserved items in RS datasets, the knowledge of which measures under/overestimate fairness and how the measure is affected by the number of additional relevant items may be useful in choosing which Joint evaluation measure to use. Measures that underestimate fairness (IFD$_{\div\text{-ori}}$, IFD$_{\div\text{-our}}$, and HD$_{\text{ori}}$) may be more desirable to avoid overclaiming, for example. In our experiments, we also find that these measures tend to become fairer with a greater amount of relevant items, and thus can serve as an 'upper bound'. Meanwhile, measures that overestimate fairness (IWO$_{\text{our}}$, IAA, and II-F$_{\text{our}}$) tend to amplify the overestimation with a greater amount of relevant items.

Effect across item selection strategy

How the additional relevant items are selected with the `top` and `bottom` strategy do not matter for IAA, IFD$_{\div\text{-our}}$, and II-F, but matters for IWO$_{\text{our}}$, IFD$_{\div\text{-ori}}$, HD$_{\text{ori}}$, and AI-F$_{\text{ori}}$. Between the affected measures, there are also differences: the `bottom`

strategy overestimates fairness more than the `top` strategy for IWO_{our}, whereas the opposite applies for $\text{IFD}_{\div\text{-ori}}$. The trends of HD_{ori} follow that of $\text{IFD}_{\div\text{-ori}}$, albeit to a lesser extent. As additional relevant items may exist close to the top-k, close to the bottom, or somewhere in between, measures that do not have large differences between the `top` and `bottom` item selection strategies are more robust than the other measures, and thus more desirable.

All in all, the trends across measures, datasets, and the placement of additional relevant items vary. First, the Joint measure scores are generally less stable against additional relevant items, when there are fewer relevant items per user in the original dataset. Therefore, the non-localisation limitation may be concerning for datasets with few relevant items per user. Second, the Joint measures respond in three different ways to a greater amount of relevant items: (i) II-F_{ori}/AI-F_{ori} are deceptively stable due to their constricted empirical range; (ii) IWO_{our}/IAA/II-F_{our} become more unfair; and (iii) $\text{IFD}_{\div}$/HD_{ori} become fairer. We note that this measure grouping differs from the one in Sections 4.6.1–4.6.2; the alignment of Joint measures to Eff and Fair measures has little to do with how the measures are affected by additional relevance items. For $\text{IFD}_{\div\text{-ori}}$/$\text{HD}_{\text{ori}}$/$\text{II-F}_{\text{ori}}$, fewer additional relevant items also mean worse over/underestimation of fairness. This second set of results could be useful in choosing suitable measures, e.g., if many relevant items are expected to be missing, measures that become fairer with a greater number of additional relevant items would be more appropriate. Third, IAA/$\text{IFD}_{\div\text{-our}}$/II-F are stable against the placement of the additional relevant items, and the rest are not. Considering the second and third sets of results, $\text{IFD}_{\div\text{-our}}$ is the most robust against the non-localisation limitation, while IWO is the least robust.

4.6.6 Sensitivity to Varying Recommendation Effectiveness and Item Exposure Distribution

How sensitive are the Joint measures towards varying proportions of relevant items and item exposure distribution in the recommendation? We assess this sensitivity as it affects score interpretation. If significant changes in item relevance and item exposure distribution do not affect a Joint measure, its score may not be an accurate indicator of recommendation fairness, specifically fairness notions that consider both item relevance and item exposure. We study the following: (i) the sensitivity of the original Joint measures compared to our corrected versions and (ii) how the inserted items affect Joint measures.

To investigate the above, we start with a ranking with the worst Eff and Fair scores, and replace a fraction of the ranking with items that are both more relevant

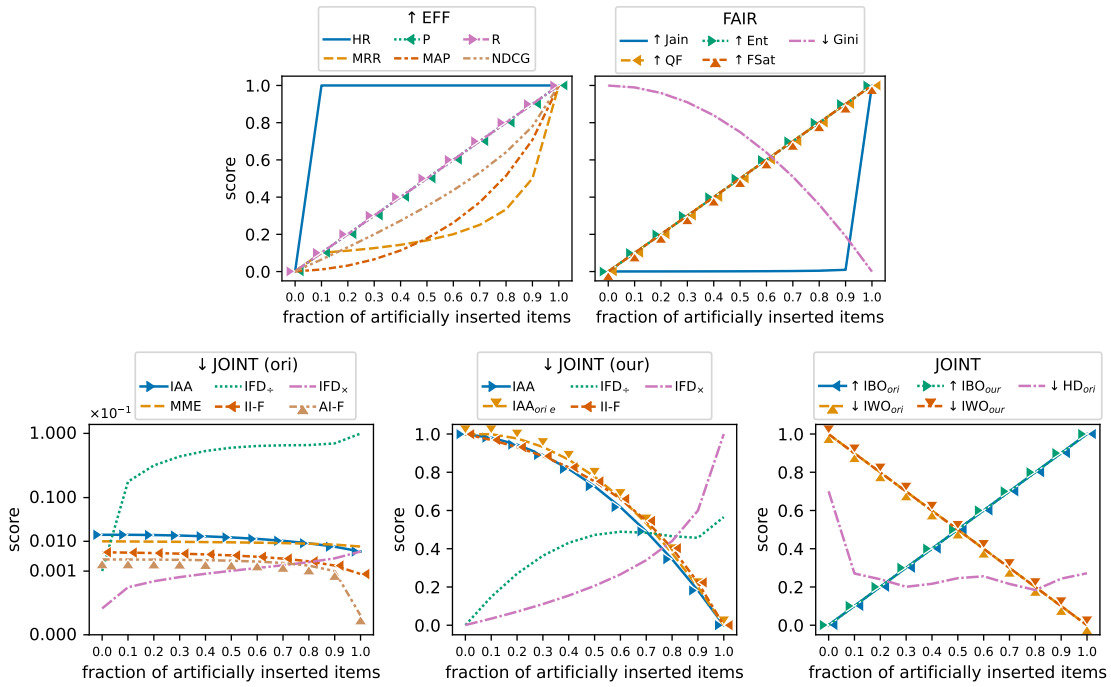


Figure 4.6: Artificial insertion of items with $m = 1000$ (users). $\text{IAA}_{\text{ori e}}$ is IAA_{our} computed with the original examination function e_{li}.

and less exposed, one item per user at a time. The real-life datasets do not suit this analysis, so we generate a synthetic dataset with $m = 1000$ users and $n = 10{,}000$ items. We then generate artificial rankings per user, following [177]. The first ranking has the same $k = 10$ items for all users. These items are irrelevant to all users, except for one.[22] For each user, we replace the item at position k with a relevant item that has less exposure, which improves both Eff and Fair. As IAA and IFD require the full recommendation list of each user, the remaining relevant items are ranked at the bottom of the full ranking (e.g., starting from position n backwards). We do this iteratively for positions $k-1, k-2, \ldots, 1$. Hence, the final rankings across all users altogether have km unique items, where each item is relevant only to the user who has it at the top-k.

We compute all Eff, Fair, and Joint measures at $k = 10$. As a result of the replacements, all Eff and Fair measures are expected to score (close to) the worst possible at the start and improve gradually as more items are inserted into the original rankings. Based on the strategies to obtain the most (un)fair scores for IAA, IFD, and II-F (Tab. 4.4), IAA and II-F are expected to start from the unfairest score and end with the fairest score, while both IFD should start from the fairest score, and $\text{IFD}_{\times}$ should end with the unfairest score. We show the analysis results in Fig. 4.6.

Sensitivity of the original Joint measures vs. ours

Generally, most original Joint measures (↓IAA/$\text{IFD}_{\times}$/MME/II-F/AI-F) are largely insensitive to changes in Eff and Fair scores, i.e., their scores vary only a little. Further, the scores are always very close to 0, i.e., the scores range between (0, 0.0015). Since all of these measures are bounded below by 0, these small scores suggest that the recommendation is almost perfectly fair, even when both the Eff and Fair scores are the worst possible. This finding is consistent with the compressed empirical range observed for real-life datasets (§4.6.1 & §4.6.3). In contrast, the empirical ranges of IBO/IWO/HD and all of our corrected Joint measures are closer to the theoretical range of $[0, 1]$. The corrected version of ↓IAA/IFD/II-F changes as expected, i.e., starting from 0 (or 1), and gradually reaching 1 (or 0). ↓$\text{IFD}_{\div\text{-our}}$ does not reach 1 as the final ranking does not correspond to the most unfair ranking for $\text{IFD}_{\div\text{-our}}$. Hence, our corrected measures are more responsive to changes in the Eff and Fair scores, compared to the original ones.

[22] This is to keep the number of items exactly km [177].

The trends of Joint measures

All Joint measures, both the originals and ours, improve (slightly) as the Eff and Fair scores become better, except for IFD. This is because IFD evaluates fairness by taking the pairwise difference of the items' exposure-relevance combined values; thus, moving the relevant items from the bottom to the top-k increases the disparity between the relevant items' exposure at and outside the top-k. The trends also differ among the Joint measures that become (slightly) fairer with more inserted items: IBO/IWO improves proportionally to the amount of inserted items, as both are the fractions of items with better/worse impact scores than a threshold; HD improves in an irregular manner as the unstable sorting[23] of items based on their ground truth relevance introduces randomness. IAA/IFD/MME/II-F/AI-F improves non-linearly. The trend of IAA and II-F closely resembles that of Gini, while the trend of $\text{IFD}_{\times\text{-our}}$ is close to 1-Gini.

Here, we also see a manifestation of the **zero-exposure** limitation in IAA. We compare IAA_{our} to $\text{IAA}_{\text{ori e}}$, i.e., IAA_{our} computed with the uncorrected linear examination function e_{li}. The uncorrected version has the same value for 0.0 and 0.1 fractions of artificially inserted items in the recommendation. The item at position 10 (corresponding to 0.1 fraction of inserted items in the recommendation) is regarded as unexposed by $\text{IAA}_{\text{ori e}}$, so replacing it with a relevant item does not make a difference. In this case, we cannot directly compare IAA_{our} to IAA_{ori} in Fig. 4.6, as the IAA_{ori} scores are very small and presented in an exponential axis, which makes it hard to see the score difference between 0.0 and 0.1 fractions of artificially inserted items.

In conclusion, the original versions of IAA/IFD/MME/II-F/AI-F are insensitive to increasing Eff and Fair scores, while our corrections to IAA/IFD/II-F reflect the change reasonably. For some original measures (MME/AI-F), the compressed empirical range issue persists, but this issue is not present for HD_{ori}/IBO/IWO. These findings are also helpful in better understanding the characteristics of MME, AI-F, HD, and IBO/IWO, which suffer from the **non-realisability** limitation but have not been corrected for it, i.e., it is not possible to derive a general formulation of the most (un)fair ranking for these measures. This analysis sheds light on the kind of ranking that produces a relatively fairer score than other rankings and thus can be used to work towards a general formulation of the most (un)fair ranking.

[23] as per the original implementation in Jeunen and Goethals [103]

4.7 Guidelines for Measure Usage

Based on the theoretical analysis (§4.3) and the empirical findings (§4.6), we formulate detailed guidelines on the appropriate usage of the relevance-aware measures (summarised in Tab. 4.8).

Table 4.8: Factors related to the measure Alignment (A), Computability (C), Interpretability (I), Expressiveness (X), Stability (S), and Efficiency (E) to consider when selecting relevance-aware item fairness measures. Statements related to (A) are neutral; the rest are desirable.

Legend ✓: the statement applies to the measure in general ⚠: the statement applies to the measure, for binary relevance and single-round recommendation ✗: the statement does not apply to the measure	IAA_{ori}	IAA_{our}	$IFD_{\div\text{-}ori}$	$IFD_{\div\text{-}our}$	$IFD_{\times\text{-}ori}$	$IFD_{\times\text{-}our}$	HD_{ori}	MME_{ori}	IBO/IWO_{ori}	IBO/IWO_{our}	$II\text{-}F_{ori}$	$II\text{-}F_{our}$	$AI\text{-}F_{ori}$
A1. The measure tends to agree with effectiveness measures	✓	✓	✗	✗	✗	✗	✓	✗	✗	✗	✓	✓	✗
A2. The measure tends to agree with fairness measures or disagree with effectiveness measures	✗	✗	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✓
C1. The measure can be computed even if the full ranking of items (not just the top-k) is unavailable	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
C2. The measure can be computed even if item relevance beyond top-k is unknown	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✓	✓	✓
C3. The measure can be computed even if the total # of relevant items per user is unknown	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗	✗
C4. The measure can be computed even if there is an item that is not relevant to any user	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓
C5. The measure can be computed for any k	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
I1. The measure maps the unfairest and the fairest recommendation to the endpoints of its theoretical range	✗	⚠	✗	⚠	✗	⚠	✗	✗	✗	✗	✗	⚠	✗
I2. The measure distinguishes between an item exposed at k and an unexposed item at $k+1$	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
X1. The measure scores are not compressed in a small interval	✗	✓	✗	✓	✗	✓	✓	✗	✓	✓	✗	✓	✗
X2. The measure responds to changes in k	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
S1. The measure is unaffected by item relevance beyond the top-k	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗	✗
S2. The measure is stable or becomes fairer if there are more relevant items beyond the top-k	✗	✗	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗
S3. The measure is unaffected by ties in item relevance	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓
S4. The measure is unaffected by the number of users with only one relevant item	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
E1. The measure can be computed separately per user	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓	✓	✗
E2. The measure can be computed separately per item	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓
E3. The measure is efficient to compute (<60 s) for large datasets (e.g., > 1K users and > 10K items)	✓	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✓	✓

Alignment

Alignment refers to the measure's tendency to agree with another measure, e.g., in having the same best model or ranking of models. We find that some relevance-aware fairness measures align more with effectiveness measures, some agree more with exposure-based measures (§4.6.2), and the rest have inconsistent alignment to the single-aspect measures. Based on their alignment, the measures can be grouped into three: (i) Eff-aligned measures (IAA/HD/II-F); (ii) Fair-aligned measures (IFD/MME/AI-F), which either strongly agree with Fair measures or disagree with Eff measures; and (iii) IBO/IWO.[24] For some measures, their alignment with the single-aspect measures can be explained based on what kind of ranking results in the most (un)fair scores (Tab. 4.4). For example, IAA and II-F align with Eff measures as placing all relevant items at the top will result in the most effective recommendations and the fairest IAA/II-F. In contrast, IFD disagrees with Eff measures, as ranking all relevant items at the bottom produces the fairest IFD, but the worst recommendation effectiveness.

Our suggestions are as follows:

- Avoid using relevance-aware measures that are highly similar to each other in their alignment to single-aspect measures. Instead, compute one measure per group to represent a set of highly similar measures; other measures in the same group do not need to be computed.
- Prioritise evaluating with non-Eff-aligned measures (i.e., IFD/MME/AI-F/IBO/IWO). Typically, recommendations are first evaluated with Eff measures, so quantifying fairness with Eff-aligned measures is not expected to add a new perspective, as the conclusion would mostly align with the Eff measures.

Computability

Computability is the extent to which the measure can be computed, considering that some information required for the computation may not be available, or that some dataset composition or experimental choices would render the measure incomputable. Given five computability-related conditions (C1–C5), only $\text{IFD}_\times$ and MME can be computed for all cases. The second 'best' measures in terms of computability are II-F and AI-F, which can be computed for 4 out of 5 computability-related cases. They

[24] Some measures in group (ii) agree with the Fair measures for only 2/4 datasets. Other measures in this group disagree with the Eff measures for only 2/4 datasets.

cannot be computed if the total number of relevant items per user is unknown. The most restricted measure in terms of computability is IAA_{ori}, which can be computed only if 4/5 conditions are fulfilled. As such, we advise the following:

- The most flexible choices are $\text{IFD}_{\times}$/MME, especially if the total number of relevant items per user is unknown. This may often be the case, e.g., in online evaluation. Otherwise, II-F/AI-F can also be computed.
- Avoid using the original version of IAA as there are too many requirements.
- Compute the corrected version of IAA instead, as it can be computed in more cases than the original.

Interpretability

Measure interpretability refers to how easily we can understand how fair the recommendation is, simply by looking at the measure scores. When a measure has a theoretical range of [0,1], normally we would expect that one endpoint is mapped to the unfairest recommendation, and another to the fairest recommendation. Unfortunately, this is not the case for any of the original measures, as we have analysed in §4.3. When we do not know the empirical achievable range of the measures, it is harder to understand the fairness level, and thus measure interpretability is lower. We have corrected IAA, $\text{IFD}_{\div}$, $\text{IFD}_{\times}$, and II-F, such that a score of 0 means the fairest possible recommendation, and 1 means the unfairest possible recommendation in binary relevance and single-round recommendation scenarios.

Another factor that affects measure interpretability, albeit to a smaller extent, is how item exposure is weighed. IAA_{ori} treats the item at position k as unexposed, which can confuse the interpretation of the scores, e.g., when the item relevance at position k is changed. Our version of IAA does not suffer from the same issue.

Based on the above, we suggest the following:

- For binary relevance and single-round recommendations, prioritise computing IAA_{our}, IFD_{our}, or II-F_{our}, as they are more interpretable than the rest.
- Always compute IAA_{our} over IAA_{ori}, as the latter quantifies exposure in a faulty way.
- The rest of the original measures can technically still be used to compare which recommendation is fairer, but not for quantifying how close the recommendation is to the (un)fairest possible recommendation.

Expressiveness

Expressiveness refers to how much the measure scores can change in quantifying fairness for various distributions of item exposure and item relevance. An expressive measure is desirable, as its score can be easily distinguishable for different rankings. We find that most original relevance-aware measures (IAA/IFD/MME/II-F/AI-F), especially $\text{IFD}_{\times}$/MME/AI-F are not expressive enough, as their scores are almost always close to the fairest even when the items are mostly irrelevant and unfair based on the single-aspect measures of Eff and Fair. In other words, their scores tend to be compressed to the low end of their range, misleading us into thinking that the recommendation is extremely fair, even when both Eff and Fair scores are almost the worst possible.

Further, they are less responsive to changes in the item exposure and relevance distribution than the single-aspect measures. Consequently, their scores are harder to distinguish across runs and need to be computed at higher precision (i.e., more decimal places) so that they can be distinguished. The scores can also be easily misinterpreted. For example, two models with a difference of only 0.001 in the scores can be interpreted to have similar performance, even if the measure has a small empirical range and is not very sensitive to begin with. Correcting IAA/IFD/II-F resolves the issues in the original measures. For other measures, the issue can be resolved by normalising the scores based on observed values of the measures [234].

We also look into the measure expressiveness in terms of the choice of the cut-off k. $\text{IFD}_{\div\text{-ori}}$ is the only measure that is insensitive to the change in k due to the **top-k-insensitivity** limitation, while other measures, including $\text{IFD}_{\div\text{-our}}$, do not have this issue.

Thus, concerning the measure expressiveness, we encourage to:

- Compute HD_{ori}/IBO/IWO or IAA_{our}/IFD_{our}/II-F_{our}, as they are more expressive than the others.
- Avoid computing the original $\text{IFD}_{\times}$/MME/AI-F, as they are extremely inexpressive.
- Be aware of potential misinterpretation in measures with small empirical scales ($\downarrow\text{IAA}_{\text{ori}}$/$\text{IFD}_{\text{ori}}$/$\text{MME}_{\text{ori}}$/$\text{II-F}_{\text{ori}}$/$\text{AI-F}_{\text{ori}}$), as they often score very close to 0.
- Refrain from computing $\text{IFD}_{\div\text{-ori}}$, unless when evaluating on the full/untruncated rankings.

Stability

Measure stability refers to the measure being unaffected by certain conditions that should not change the measure score. The conditions are: having additional relevant items beyond the top-k (S1–S2), having ties in the items' ground truth relevance (S3), and having users with only one relevant item (S4). For S1–S2, in common RS datasets, most users only have a few relevant items each. Hence, RS fairness evaluation may suffer from fluctuations if there are more relevant items than expected, even if these relevant items are not in the top-k. For S2, given more relevant items beyond the top-k, we desire measures that remain constant or become fairer to avoid overestimating fairness. In practice, we may not know the number of additional relevant items. If the measure becomes unfairer with more relevant items beyond the top-k, fairness is overestimated based on the existing item relevance. Thus, it is better to have a score that can guarantee that the score will not become more unfair in case of additional relevant items beyond the top-k. For S3, sorting items based on their ground truth relevance to compute HD introduces some noise depending on the sorting algorithm, especially when there are many items with the same relevance. For S4, $\text{IFD}_{\div}$ automatically gives the fairest score (0) when a user only has 1 relevant item, pulling the score down when there is a large percentage of such users. Out of all stability-related conditions, S4 is the easiest to handle, as users with only 1 relevant item can be excluded from the computation, as $\text{IFD}_{\div}$ is a pairwise measure. Overall, only $\text{IFD}_{\times}$ and MME are stable against the four stability-related conditions.

Considering the above, we make the following suggestions:

- Compute $\text{IFD}_{\times}$ and MME, as they are the most stable and the remaining measures are only stable against 2/4 stability-related conditions.
- HD should ideally not be used for cases where the items' ground truth relevance values include a lot of ties.
- When computing $\text{IFD}_{\div}$, calculate the number of users with a single relevant item. If too many such users exist, we suggest excluding them from evaluation and reporting the number of users with a single relevant item.

Efficiency

Is the measure computationally expensive? Can some parts of the computation be potentially parallelised? Here, we discuss the computational efficiency of the measures and the potential to compute them in a more efficient manner. Most relevance-aware measures are relatively cheap to compute (under 60 seconds per

model, per dataset, based on the measure computation time in Rampisela et al. [181]), except for MME and $\text{IFD}_\times$.[25] Both MME and $\text{IFD}_\times$ are computationally expensive as they are pairwise measures. Computing MME can take up to 30 minutes for larger datasets, while $\text{IFD}_\times$ is slightly faster (∼12 mins). However, we note that the computation of both measures can also be parallelised, as the scores can first be computed either per user or item prior to the aggregation. Therefore, in terms of computational efficiency, our suggestions are as follows:

- Avoid computing MME and $\text{IFD}_\times$ for datasets with many items.
- Consider parallelising the measure computation to save time.

Final suggestions

Overall, considering the alignment, computability, interpretability, expressiveness, stability, and computational efficiency of the relevance-aware individual item fairness measures, we provide the following suggestions:

- Among all relevance-aware individual item fairness measures, computing $\text{IFD}_{\times\text{-our}}$ is one of the better choices, as it exhibits the most desirable characteristics in terms of its alignment to single-aspect measures, computability, interpretability, expressiveness, and stability. Even though it is also the second most computationally expensive measure, and can take more than 10 minutes to compute for larger datasets, the computation time can potentially be improved with parallelisation.
- $\text{IBO}_{\text{our}}/\text{IWO}_{\text{our}}$ are good alternatives, as they align with neither Eff nor Fair measures, suggesting that they quantify a different evaluation dimension from the single-aspect measures. $\text{IBO}_{\text{our}}/\text{IWO}_{\text{our}}$ also do not suffer from expressiveness or efficiency issues.
- Compute the corrected measures whenever possible, as they have more desirable properties overall than the original measures.

[25] The computation time of IFD is based on our experiments §4.6.3.

4.8 Related Work

Fairness evaluation in ranking and RSs

There exist many surveys on fairness in ranking or specifically in RSs [1, 9, 54, 133, 173, 209, 228, 237, 252]. Most of them provide a general overview of how fairness is quantified based on various dimensions, e.g., stakeholders and granularity, yet none has provided an in-depth analysis on the existing fairness measures. Aside from the survey papers, there is other prior work that are closer to ours. Cherumanal et al. [40] analysed the agreement between relevance, fairness, and diversity measures in the argument retrieval task. Additionally, previous work has investigated both the theoretical and empirical properties of fairness measures in ranking [175, 177, 198]. Similar to our work, previous work highlights the limitations of existing measures in various practical settings and provides a better understanding of the distinctive ways in which the measures quantify fairness. However, the granularity of the measures differs: Schumacher et al. [198] and Raj and Ekstrand [175] focus on evaluation measures for group fairness, whereas we focus on individual fairness measures. Rampisela et al. [177] also study individual item fairness, but they examine only exposure-based fairness measures that are detached from relevance, while we investigate measures that jointly account for both item exposure and item relevance.

Group fairness and individual fairness

Based on the granularity, RS fairness can either be evaluated at the group or individual level, although most existing work focuses on the former [54]. Group fairness is defined as providing similar treatment to both dominant and protected groups (e.g., gender minority) [54, 65], while individual fairness is defined as ensuring that similar individuals receive similar treatment [63]. Evaluating individual fairness has several benefits over group fairness. Firstly, individual fairness evaluation does not need sensitive demographic information (e.g., gender, race), which is often used for group fairness evaluation [124], but may not be readily available due to privacy or legal issues. Secondly, evaluating individual fairness provides a more comprehensive overview by considering the distribution across all individuals [124], rather than only considering a few groups, which can be too simplistic as the grouping may not accurately represent the individuals' whole identity [54, 65].

Accounting for relevance and exposure fairness jointly

Other than using an evaluation measure *per se*, jointly quantifying item relevance and exposure can also be done through other approaches. Cherumanal et al. [40] compute the harmonic mean of relevance and exposure-based fairness measures to rank the systems based on both aspects, while Rampisela et al. [181] compare the model scores to the Pareto Frontier, which considers the trade-off between relevance and fairness. Beyond individual item fairness, there are other evaluation measures that jointly quantify relevance and fairness. Gao et al. [79] introduce a measure combining KL-divergence and IDCG to jointly evaluate relevance and group fairness in IR. Xu et al. [239] measure utility and provider fairness in RSs jointly using a weighted sum of the two aspects. The relevance-aware measures in our work differ from the two previous measures [79, 239] as none of them combine two single-aspect measures [40, 79, 181] or apply a weighted sum of the two aspects [239]. In contrast, García-Soriano and Bonchi [81] quantifies fairness and relevance for individual fairness in ranking similarly to HD [103], as both compare item positions based on ground truth relevance to their positions based on predicted rankings. However, we do not use the measure in García-Soriano and Bonchi [81] because it requires significant modifications to customise it for RS fairness.

4.9 Conclusion

We have conducted a novel theoretical and empirical investigation of existing relevance-aware individual item fairness measures in Recommender Systems (RSs), which account for both item exposure and item relevance. We identified theoretical measure limitations and studied the extent of the limitations under common and extreme evaluation settings. Further, we addressed these limitations by redefining measures where possible, or explaining why some limitations are unresolvable. Our comprehensive empirical analysis using real-life and synthetic data led to practical guidelines for selecting the appropriate measures for various use cases.

In this work, we focused primarily on deterministic/single rankings of items with binary relevance and, more briefly discussed measure limitations under non-binary relevance and multi-round scenarios at a conceptual level. Future work should conduct a deeper investigation, both theoretically and empirically, into relevance-aware fairness evaluation given a distribution of rankings (i.e., stochastic rankings or multi-round recommendations), to align with distributionally-informed evaluation in RSs [66], and for graded relevance. Other future work could also explore the dynamics be-

tween individual and group fairness evaluation measures, as well as multistakeholder evaluation, which accounts for user and item fairness jointly [29, 229].

Acknowledgements

The work is supported by the Algorithms, Data, and Democracy project (ADD-project), funded by the Villum Foundation and Velux Foundation. Pietro Tropeano contributed to a sketch of the proof.

4.10 Appendix

4.10.1 Mathematical workings for ranking strategies to achieve the (un)fairest Joint scores

We provide the derivation of the fairest ranking for the original IAA(u), IFD$_{\div}$(u) (computed with $J_{\div\text{-our}}$), IFD$_{\times}$(u), and II-F(u). For IAA(u), we also show the workings behind the unfairest ranking.

4.10.1.1 Inequity of Amortized Attention (IAA)

We show that ranking relevant items at the top-k (i.e., ranking items in non-increasing relevance) produces the fairest $\downarrow$IAA(u). Let a_p^{IAA} be the exposure of item at position p, where item exposure in IAA is based on the product of the top-k indicator function and the examination function, i.e., $1_{L_u}(i) \cdot \tilde{e}_{li\text{-our}}(u, i)$; for brevity, we denote this as a_p. Let $r_{u,i} \in \{0, 1\}$ be the relevance of item i to user u; for brevity, we denote this as r_i. Formally,

Theorem 4.1. *Given any two items i and i', where $z(u, i) = p$, $z(u, i') = p'$, $r_i > r_{i'}$, ranking the item with higher relevance (i) closer to the top than the item with lower relevance (i') would result in a fairer (or equal) $\downarrow$IAA(u) than doing vice versa:*

$$r_i > r_{i'} \wedge p < p' \Rightarrow |a_p - r_i| + |a_{p'} - r_{i'}| \leq |a_p - r_{i'}| + |a_{p'} - r_i|$$

Proof. We prove Theorem 4.1 by contradiction. Assume Theorem 4.1 is false, i.e.:

$$\begin{aligned}
&\neg\left(r_i > r_{i'} \wedge p < p' \Rightarrow |a_p - r_i| + |a_{p'} - r_{i'}| \leq |a_p - r_{i'}| + |a_{p'} - r_i|\right) \\
&\neg\left(\neg(r_i > r_{i'} \wedge p < p') \vee (|a_p - r_i| + |a_{p'} - r_{i'}| \leq |a_p - r_{i'}| + |a_{p'} - r_i|)\right) \\
&(r_i > r_{i'} \wedge p < p') \wedge (|a_p - r_i| + |a_{p'} - r_{i'}| > |a_p - r_{i'}| + |a_{p'} - r_i|)
\end{aligned}$$

As $r_i > r_{i'}$ and we deal with binary relevance, $r_i = 1$ and $r_{i'} = 0$. Substituting these values to the LHS of the inequality, we get: $|a_p - r_i| + |a_{p'} - r_{i'}| = |a_p - 1| + |a_{p'}| = |a_p - 1| + a_{p'}$ (as $a_{p'} \geq 0$). Substituting $r_i = 1$ and $r_{i'} = 0$ to the RHS of the inequality, we get: $|a_p - r_{i'}| + |a_{p'} - r_i| = |a_p| + |a_{p'} - 1| = a_p + |a_{p'} - 1|$ (as $a_p \geq 0$).

Putting the LHS and RHS of the inequality together, we have:

$$
\begin{aligned}
|a_p - 1| + a_{p'} &> a_p + |a_{p'} - 1| \\
|a_p - 1| - |a_{p'} - 1| &> a_p - a_{p'} \\
|a_p - 1| - |a_{p'} - 1| &> 0 && (\because p < p' \Leftrightarrow a_p \geq a_{p'} \Leftrightarrow a_p - a_{p'} \geq 0) \\
(1 - a_p) - (1 - a_{p'}) &> 0 && (\because a_p \leq 1 \Leftrightarrow a_p - 1 \leq 0) \\
a_{p'} &> a_p && (\text{Contradiction with } p < p' \Leftrightarrow a_p \geq a_{p'})
\end{aligned}
$$

Hence, Theorem 4.1 holds. □

Next, we show that ranking relevant items at the bottom (i.e., ranking items in non-decreasing relevance) produces the unfairest $\downarrow$IAA(u).

Theorem 4.2. *Given any two items i and i', where $z(u, i) = p$, $z(u, i') = p'$, $r_i < r_{i'}$, ranking the item with lower relevance (i) closer to the top than the item with higher relevance (i') would result in an unfairer (or equal) $\downarrow IAA(u)$ than doing vice versa:*

$$r_i < r_{i'} \wedge p < p' \Rightarrow |a_p - r_i| + |a_{p'} - r_{i'}| \geq |a_{p'} - r_i| + |a_p - r_{i'}|$$

Proof. We prove Theorem 4.2 by contradiction. Assume Theorem 4.2 is false, i.e.:

$$
\begin{aligned}
\neg\left(r_i < r_{i'} \wedge p < p' \Rightarrow |a_p - r_i| + |a_{p'} - r_{i'}| \geq |a_{p'} - r_i| + |a_p r_{i'}|\right) \\
\neg\left(\neg(r_i < r_{i'} \wedge p < p') \vee (|a_p - r_i| + |a_{p'} - r_{i'}| \geq |a_{p'} - r_i| + |a_p r_{i'}|)\right) \\
(r_i < r_{i'} \wedge p < p') \wedge (|a_p - r_i| + |a_{p'} - r_{i'}| < |a_{p'} - r_i| + |a_p - r_{i'}|)
\end{aligned}
$$

As $r_i < r_{i'}$ and we deal with binary relevance, $r_i = 0$ and $r_{i'} = 1$. Substituting these values to the left-hand side (LHS) of the inequality, we get: $|a_p - r_i| + |a_{p'} - r_{i'}| = |a_p| + |a_{p'} - 1| = a_p + |a_{p'} - 1|$ (as $a_p \geq 0$). Substituting $r_i = 0$ and $r_{i'} = 1$ to the right-hand side (RHS) of the inequality, we obtain: $|a_p - r_{i'}| + |a_{p'} - r_i| = |a_p - 1| + |a_{p'}| = |a_p - 1| + a_{p'}$ (as $a_{p'} \geq 0$). Putting the LHS and RHS of the inequality together, we have:

$$
\begin{aligned}
a_p + |a_{p'} - 1| &< |a_p - 1| + a_{p'} \\
a_p - a_{p'} &< |a_p - 1| - |a_{p'} - 1| \\
0 &< |a_p - 1| - |a_{p'} - 1| && (\because p < p' \Leftrightarrow a_p \geq a_{p'} \Leftrightarrow a_p - a_{p'} \geq 0) \\
0 &< (1 - a_p) - (1 - a_{p'}) && (\because a_p \leq 1 \Leftrightarrow a_p - 1 \leq 0) \\
a_p &< a_{p'} && (\text{Contradiction with } p < p' \Leftrightarrow a_p \geq a_{p'})
\end{aligned}
$$

Hence, Theorem 4.2 holds. □

4.10.1.2 Individual Fairness Disparity with Division ($\text{IFD}_{\div}$)

We show that ranking relevant items at the bottom (i.e., ranking items in non-decreasing relevance) produces the fairest $\downarrow\text{IFD}_{\div}(u)$. We denote as $a_p^{\text{IFD}_{\div}}$ the exposure of item at position p, where item exposure in $\text{IFD}_{\div}$ is based on $J_{\div\text{-our}}$ (Eq. (4.27)). For brevity, we denote as a_p the item exposure.

Theorem 4.3. *Given any two relevant items i and i', where $z(u,i) = p$, $z(u,i') = p'$, $r_i = r_{i'} = 1$, ranking the items closer to the bottom (beyond the top-k) would result in a fairer (or equal) $\downarrow IFD_{\div}(u)$ than doing vice versa. Formally,*

$$\max\left\{0, \frac{a_p}{r_i} - \frac{a_{p'}}{r_{i'}}\right\} + \max\left\{0, \frac{a_{p'}}{r_{i'}} - \frac{a_p}{r_i}\right\} \tag{4.37}$$

is: (1) lower for $k < p < p'$ than for $p < k < p'$ (ranking 1 item at the top-k and another one beyond the top-k); and (2) lower for $k < p < p'$ than for $p < p' < k$ (ranking both items at the top-k).

Proof. We prove (1) and (2) separately. To prove (1), we start from the premise $k < p < p'$. Ranking the two relevant items beyond the top-k results in $a_p = a_{p'} = 0$. As such, Eq. (4.37) is 0. Meanwhile, for $p < k < p'$, we have $a_p > 0$ and $a_{p'} = 0$. As $a_p > a_{p'}$, the expression $\frac{a_p}{r_i} - \frac{a_{p'}}{r_i} = a_p > 0$, while $\frac{a_{p'}}{r_{i'}} - \frac{a_p}{r_i} = -a_p < 0$. Eq. (4.37) evaluates to $\max\{0, a_p\} + \max\{0, -a_p\} = a_p > 0$. Therefore, (1) holds.

To prove (2), we use the results of proving (1), i.e., when $k < p < p'$, Eq. (4.37) $= 0$. For $p < p' < k$, we have $a_p > a_{p'} > 0$. The expression $\frac{a_p}{r_i} - \frac{a_{p'}}{r_i} = a_p - a_{p'} > 0$, while $\frac{a_{p'}}{r_{i'}} - \frac{a_p}{r_i} = a_{p'} - a_p < 0$. Eq. (4.37) evaluates to $\max\{0, a_p - a_{p'}\} + \max\{0, a_{p'} - a_p\} = a_p - a_{p'} + 0 > 0$. Thus, (2) holds.

As both (1) and (2) hold, Theorem 4.3 holds. □

4.10.1.3 Individual Fairness Disparity with Multiplication ($\text{IFD}_{\times}$)

We show that ranking relevant items at the bottom (i.e., ranking items in non-decreasing relevance) produces the fairest $\downarrow\text{IFD}_{\times}(u)$. We denote as $a_p^{\text{IFD}_{\times}}$ the exposure of item at position p, where item exposure in $\text{IFD}_{\times}$ is based on $J_{\times\text{-ori}}$ (Eq. (4.7)). For brevity, we denote as a_p the item exposure.

Theorem 4.4. *Given any two items i and i', where $z(u,i) = p$, $z(u,i') = p'$, $r_i < r_{i'}$, ranking the item with lower relevance (i) closer to the top than the item with higher relevance (i') would result in a fairer (or equal) $\downarrow IFD_{\times}(u)$ than doing vice versa:*

$$r_i < r_{i'} \wedge p < p' \Rightarrow (a_p r_i - a_{p'} r_{i'})^2 \leq (a_{p'} r_i - a_p r_i)^2$$

Proof. We prove Theorem 4.4 by contradiction. Assume Theorem 4.4 is false, i.e.:

$$\begin{aligned} \neg \left(r_i < r_{i'} \wedge p < p' \Rightarrow (a_p r_i - a_{p'} r_{i'})^2 \leq (a_{p'} r_i - a_p r_i)^2 \right) \\ \neg \left(\neg \left(r_i < r_{i'} \wedge p < p' \right) \vee \left[(a_p r_i - a_{p'} r_{i'})^2 \leq (a_{p'} r_i - a_p r_i)^2 \right] \right) \\ \left(r_i < r_{i'} \wedge p < p' \right) \wedge \left[(a_p r_i - a_{p'} r_{i'})^2 > (a_{p'} r_i - a_p r_i)^2 \right] \end{aligned}$$

As $r_i < r_{i'}$ and we deal with binary relevance, $r_i = 0$ and $r_{i'} = 1$. Substituting these values to the inequality, we get $a_{p'}^2 > a_p^2$. There are three possibilities of how the rank positions p, p' relate to the cut-off k:

1. Both rank positions are beyond k, $k < p < p'$. In this case, $a_p = a_{p'} = 0$ as items at position p and p' are unexposed. As $a_{p'}^2 = a_p^2$, this contradicts $a_{p'}^2 > a_p^2$.
2. Both rank positions are at the top-k, $p < p' < k$. Here, $a_p > a_{p'} > 0 \Leftrightarrow a_p^2 > a_{p'}^2$, which contradicts $a_{p'}^2 > a_p^2$.
3. One rank position is at the top-k, and another one beyond top-k, $p < k < p'$. Here, $a_p > 0 = a_{p'} \Leftrightarrow a_p^2 > a_{p'}^2$, which contradicts $a_{p'}^2 > a_p^2$.

Therefore, Theorem 4.4 holds via proof by contradiction. □

4.10.1.4 Individual-user-to-Individual-item Fairness (II-F)

We show that ranking relevant items at the top-k (i.e., ranking items in non-increasing relevance) produces the fairest $\downarrow$II-F(u). We denote as $a_p^{\text{II-F}}$ the exposure of item at position p, where item exposure in II-F is $E_{u,i}$ (Eq. (4.22)). For brevity, we denote as a_p the item exposure and as t_i the target exposure of item i, $E^*_{u,i}$ (Eq. (4.23)).

Theorem 4.5. *Given any two items i and i', where $z(u,i) = p$, $z(u,i') = p'$, $r_i > r_{i'}$, ranking the item with higher relevance (i) closer to the top than the item with lower relevance (i') would result in a fairer (or equal) $\downarrow$II-F(u) than doing vice versa:*

$$r_i > r_{i'} \wedge p < p' \Rightarrow (a_p - t_i)^2 + (a_{p'} - t_{i'})^2 \leq (a_{p'} - t_i)^2 + (a_p - t_{i'})^2$$

Proof. We prove Theorem 4.5 by contradiction. Assume Theorem 4.5 is false, i.e.:

$$\neg\left(r_i > r_{i'} \wedge p < p' \Rightarrow (a_p - t_i)^2 + (a_{p'} - t_{i'})^2 \leq (a_{p'} - t_i)^2 + (a_p - t_{i'})^2\right)$$
$$\neg\left(\neg\left(r_i > r_{i'} \wedge p < p'\right) \vee \left[(a_p - t_i)^2 + (a_{p'} - t_{i'})^2 \leq (a_{p'} - t_i)^2 + (a_p - t_{i'})^2\right]\right)$$
$$\left(r_i > r_{i'} \wedge p < p'\right) \wedge \left[(a_p - t_i)^2 + (a_{p'} - t_{i'})^2 > (a_{p'} - t_i)^2 + (a_p - t_{i'})^2\right]$$

As $r_i > r_{i'}$ and we deal with binary relevance, $r_i = 1$ and $r_{i'} = 0$, which means the target exposures of item i and i' are $t_i > t_{i'} = 0$. Substituting these values to the inequality, we get:

$$\begin{aligned}
(a_p - t_i)^2 + a_{p'}^2 &> (a_{p'} - t_i)^2 + a_p^2 \\
(a_p - t_i)^2 - (a_{p'} - t_i)^2 &> a_p^2 - a_{p'}^2 \\
(a_p^2 - 2a_p t_i + t_i^2) - (a_{p'}^2 - 2a_{p'} t_i + t_i^2) &> a_p^2 - a_{p'}^2 \\
(-2a_p t_i) - (-2a_{p'} t_i) &> 0 \\
2a_{p'} t_i &> 2a_p t_i \\
a_{p'} &> a_p
\end{aligned}$$

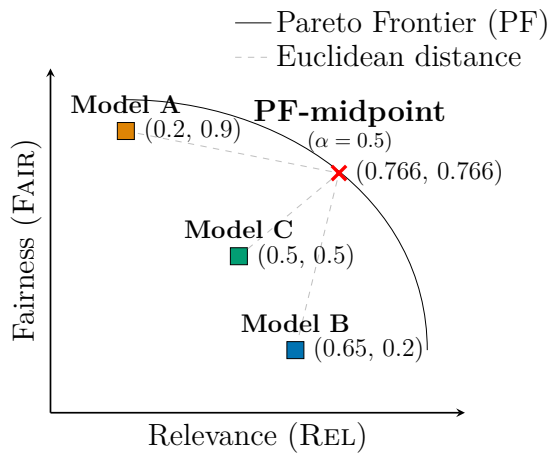
However, as $p < p' \Leftrightarrow a_p \geq a_{p'}$, there is a contradiction. Hence, Theorem 4.5 holds. □

Chapter 5

Joint Evaluation of Fairness and Relevance in Recommender Systems with Pareto Frontier

Abstract

Fairness and relevance are two important aspects of recommender systems (RSs). Typically, they are evaluated either (i) separately by individual measures of fairness and relevance, or (ii) jointly using a single measure that accounts for fairness with respect to relevance. However, approach (i) often does not provide a reliable joint estimate of the goodness of the models, as it has two different best models: one for fairness and another for relevance. Approach (ii) is also problematic because these measures tend to be ad-hoc and do not relate well to traditional relevance measures, like NDCG. Motivated by this, we present a new approach for jointly evaluating fairness and relevance in RSs: Distance to Pareto Frontier (DPFR). Given some user-item interaction data, we compute their Pareto frontier for a pair of existing relevance and fairness measures, and then use the distance from the frontier as a measure of the jointly achievable fairness and relevance. Our approach is modular and intuitive as it can be computed with existing measures. Experiments with 4 RS models, 3 re-ranking strategies, and 6 datasets show that existing metrics have inconsistent associations with our Pareto-optimal solution, making DPFR a more robust and theoretically well-founded joint measure for assessing fairness and relevance. Our code: github.com/theresiavr/DPFR-recsys-evaluation.



Model	↓ Dist. to PF	↑ Avg
A	0.582	**0.55**
B	0.578	0.425
C	**0.376**	0.5

Figure 5.1: (x, y) denotes the pair of relevance and fairness score. Example: Model A is best for fairness, Model B is best for relevance, and Model C is the closest to the Pareto Frontier (PF) midpoint, when relevance and fairness are equally weighted ($\alpha = 0.5$). Averaging relevance and fairness (Avg) leads to falsely concluding that Model A is best for both aspects. Note that distance to PF also beats other existing measures of fairness and relevance (see §5.5.4).

5.1 Introduction

Relevance and fairness are important aspects of recommender systems (RSs). Relevance is typically evaluated using common ranking measures (e.g., NDCG), while various fairness measures for RSs exist [9, 228]. Some fairness measures integrate relevance, so that they evaluate fairness w.r.t. relevance. The problem with these joint measures is that they tend to be ad-hoc, unstable, and they do not account very well for both aspects simultaneously [178]. Another way of evaluating relevance and fairness is to use a different measure for each aspect. However, this does not always provide a reliable joint estimate of the goodness of the models, as it may have two different best models: one for fairness and another for relevance. This can be avoided by aggregating the scores of the two measures into a single score, or by aggregating the resulting model rankings into one using rank fusion. These approaches are also problematic because: (i) the scores of the two measures may have different distributions and different scales, making them hard to combine; (ii) the two measures may not even be computed with the same input, making their combination hard to interpret (relevance scores are computed for individual users and then averaged, while fairness measures for individual items are typically based on individual item

recommendation frequency); and (iii) the resulting scores are less understandable as it is unknown how close the models are to an ideal balance of fairness and relevance, e.g., an acceptable trade-off between fairness and relevance scores.

To address the above limitations, we contribute an approach that builds on the set of all Pareto-optimal solutions [36]. Our approach addresses issue (i) and (ii) above by avoiding direct combination of measures. We directly address (iii) by computing the distance of the model scores to a desired fairness-relevance balance. Our approach uses Pareto-optimality, a popular concept in multi-objective optimization problems across domains, including RSs [187]. A recommendation is Pareto-optimal if there are no other possible recommendations with the same Rel score that achieve better fairness.[1] In other words, given Pareto-optimal solutions, we cannot get other recommendations that empirically perform better, unless relevance is sacrificed. In our approach, we combine existing Fair measures and Rel measures as follows. We build a Pareto Frontier (PF) that first maximises relevance, finds the best fairness achievable under the relevance constraint, and then jointly quantifies fairness and relevance as the distance from an optimal solution, see Fig. 5.1.

Our approach, *Distance to PF of Fairness and Relevance* (DPFR) has several strengths. First, DPFR is *modular*; it can be used with well-known existing measures of relevance and fairness. DPFR is also *tractable* as one can control the weight (α) of fairness w.r.t. relevance. As the resulting score is the distance to the scores of a traditional relevance measure and a well-known fairness measure, DPFR is also *intuitive* in its interpretation. Most importantly, DPFR is a principled way of jointly evaluating relevance and fairness based on an empirical best solution that uses Pareto-optimality. Experiments with different RS models, re-ranking approaches and datasets show that there exists a noticeable gap between using current measures of relevance and fairness and our Pareto-optimal joint evaluation of relevance and fairness. This gap is bigger in larger datasets and when using rank-based relevance measures (i.e., MAP, NDCG), as opposed to set-based relevance measures (i.e., Precision, Recall).

In this work, we focus on **individual item fairness**. This type of fairness is commonly defined as all items having equal exposure, where exposure typically refers to the frequency of item appearance in the recommendation list across all users [147, 167, 177]. Individual item fairness is important in ensuring that each item/product in the system has a chance to be recommended to any user [124].

[1]The opposite is also true, but in RS scenario the Rel score is usually the primary objective, not the Fair score.

5.2 Related Work

Evaluating fairness and relevance together is a type of multi-aspect evaluation. However, none of the existing multi-aspect evaluation methods [140, 143, 164] can be used in this case as these methods require separate labels that are unavailable in RS scenarios. Specifically, it is not possible to label an item as 'fair', because item fairness depends on other recommended items. The same item can be a fair recommendation in one ranking, but unfair in another. In RSs, fairness is typically defined as treating users or items without discrimination [19]. This is often quantified as the opportunity for having equal relevance (for users) or exposure (for items) [19, 227], computed either individually or for groups of items/users [175, 252].

The problem of jointly evaluating RS relevance and fairness is further aggravated by the fact that improved fairness is often achieved at the expense of relevance to users [150]. We posit that this trade-off makes multi-objective optimization a suitable solution. Pareto optimality is a well-known objective for such optimization, and it has been previously used in RS but only to recommend items to users [83, 187, 239, 262]. Because the true PF is often unknown due to the problem complexity [11, 123], prior work has used the model's training loss w.r.t. two different aspects [138] or scores from different models [157, 165] to generate the PF. Our work differs from this in terms of both the purpose of using Pareto-optimal solutions, and the nature of the PF. Specifically, we exploit Pareto-optimality through PF as a robust *evaluation* method, instead of as a recommendation method. In addition, our generated PF is based on the ground truth (i.e., the test set), a common RS evaluation approach, instead of the recommender models' empirical performance, which may not be optimal. Thus, our PF is also model-agnostic, as opposed to the PF in [239]. Our approach differs also from FAIR [79] since the PF considers the empirically achievable optimal solution based on the dataset, while FAIR compares against the desired fairness distribution which might not be achievable. Lastly, [165] selects the optimal solution based on its distance to the utopia point (the theoretical ideal scores), whereas the utopia point may not be realistic due to dataset or measure characteristics [153, 177]. Since our PF is generated based on test data, any of its solutions is empirically achievable.

5.3 Distance to Pareto Frontier (DPFR)

We present definitions (§5.3.1), and then explain DPFR in different steps: given a Fair and a Rel measure, how to generate PF based on the ground truth data in the test set (§5.3.2); how to choose a reference point in the PF based on α (e.g., the

midpoint for $\alpha = 0.5$) (§5.3.3); and how to compute the distance of the FAIR and REL scores to the reference point with a distance measure d (§5.3.3). Additionally, we present a computationally efficient adaptation of DPFR (§5.3.4).

5.3.1 Definitions

We adapt the Pareto-optimality definition [220]: the multi-objective problem is finding the optimum FAIR score s_f, and REL score, s_r from a list of possible recommendations across all users. We define the tuple $s = (s_r, s_f) \in S$, where S is the Cartesian product of all possible REL and FAIR scores. The relation $\geq_A$ ($>_A$) means 'better or equal to' ('better to') according to an aspect $A \in \{\text{REL}, \text{FAIR}\}$.

Definition 5.1 (Pareto Dominance)**.** *A tuple $s = (s_r, s_f)$ dominates $s' = (s'_r, s'_f)$ iff s is partially better than s', i.e., $s_r \geq_{\text{REL}} s'_r$ and $s_f \geq_{\text{FAIR}} s'_f$, in addition to $s_r >_{\text{REL}} s'_r$ or $s_f >_{\text{FAIR}} s'_f$.*

Definition 5.2 (Pareto Optimality)**.** *A solution (recommendation list) that has REL and FAIR scores of $x = (x_r, x_f) \in S$ is Pareto-optimal iff there is no other solution with $x' = (x'_r, x'_f) \in S$ that dominates x.*

Definition 5.3 (Pareto Frontier)**.** *The set of all Pareto-optimal tuples.*

5.3.2 Pareto Frontier generation

Given user-item preference data (e.g., test set), the aim is to explore the empirical, maximum feasible fairness towards individual items, such that the recommendation satisfies Pareto-optimality w.r.t. fairness scores across all items and an average relevance score across users, e.g., MAP@10 = 0.9.[2] This is done to measure how far a model performance is, from these Pareto-optimal solutions. Enumerating all possible recommendations for users and items to find the complete set of Pareto-optimal solutions is computationally infeasible, and there is no analytical solution either. Instead, we contribute an algorithm that iteratively builds upon a maximally relevant initial recommendation list. Our algorithm iteratively finds Pareto-optimal recommendations by prioritising relevance over fairness, as recommendations are usually optimised for relevance (with or without fairness). This prioritisation is known as lexicographic optimization [189]. We call our algorithm ORACLE2FAIR (full technical description in App. 5.7.2). Our algorithm generates the PF of fairness and relevance in two steps: **(1) initialisation** of the recommendations with an *Oracle*

[2]This is how FAIR and REL measures are usually computed.

(App.5.7.2, Algorithm 1). The *Oracle* generates a recommendation with the highest empirical score for relevance, based on user interactions that are part of the test set. This step is followed by **(2) replacements** to make the recommendations as *Fair* as possible; at the end of this algorithm, the Fair scores should reach the empirically fairest score while maintaining as much relevance as possible. Throughout the PF generation, items in a user's train/val split are not recommended to the same user. Henceforth, *relevant items* refers to the items in a user's test split.

(1) Initialisation. The Oracle recommends at most $k = 10$ relevant items, from the n items in the dataset, to each of the m users in the test split, one user at a time. The recommendation begins with users having exactly k items in the test split; only these items can be recommended to those users to gain the maximum relevance. Recommendations to other users are made maximally relevant and fair as follows: if a user has $> k$ relevant items, we pick k items with the least exposure among them. Item exposure is computed based on what has been recommended to other users who already have exactly k items. Note that this process is not trivial (see App. 5.7.2, Algo 1, ll. 2–17). If a user has $< k$ relevant items, we recommend those items at the top (to maximise top-weighted Rel measures) and fill the rest of their recommendation slots with the least exposed items in the dataset (Algo 1, ll. 18–36). This least-exposure prioritisation strategy ensures that the solutions are Pareto-optimal.

(2) Replacements. The algorithm iteratively replaces the recommended items to achieve maximum fairness, such that each replacement results in a fairer recommendation than the previous. We compute the Fair and Rel measures after each replacement as follows. The most popular item, which is recommended most often, is replaced with one of these item types, in decreasing order of priority: an unexposed item, then the least popular item in the recommendation; this increases fairness from the previous recommendations. We do this one item and one user at a time, starting with the users that have the most popular item at the bottom of their recommendation list, to ensure that the decrease in relevance is minimum as the replacement item is mostly not relevant to that user. Nonetheless, the Oracle2Fair prioritises replacing the recommendations of users for whom the replacement item is relevant (if any). As fairness increases and relevance decreases/stays the same from the previous recommendation, the new recommendation is also Pareto-optimal. We continue the replacement until the maximum times any item is recommended is $\lceil km/n \rceil$, i.e., the upper bound of how many times an item can be recommended, if all items in the dataset must appear in the recommendation as uniformly as possible. We explicitly used $\lceil km/n \rceil$ as a stopping condition for Oracle2Fair. To ensure maximum Rel

scores (especially in top-weighted measures), each time a replacement takes place, we rerank the recommendations based on descending relevance.

The resulting (Rel, Fair) scores reflecting Pareto-optimal recommendations from this process make up the PF. If there are duplicates in the Rel value, we keep the best Fair score for a single value of Rel. While it cannot be reasonably verified that the resulting PF matches the theoretical PF, this is a close way to build the full PF, as opposed to building the PF from trained models scores (§5.2).

5.3.3 Distance computation

For each pair of Fair and Rel measures, we find a reference point using a tunable parameter $\alpha \in [0, 1]$; $\alpha = 0$ means only relevance is accounted for, and $\alpha = 1$ means only fairness is accounted for. Next, we explain how to compute the reference point. We first use the following equation to find the length of a subset T of the PF: $lenPF(T) = \sum_{t=1}^{|T|-1} d_E(x^t, x^{t+1})$. Given that P is the set of all Pareto-optimal solutions, $x^t = (x_r^t, x_f^t)$ is the pair of Pareto-optimal solutions (x_r, x_f) with the t-th highest x_r in P, and d_E is the Euclidean distance. The overall PF length is $lenPF(P)$ or simply $lenPF$.

The reference point is $s_\alpha = x^{t'}$, where t' is computed as follows:

$$t' = \underset{j \in [1,\ldots,|P|-1]}{\arg\min} \left| lenPF(T^j) - \ \alpha \cdot lenPF \right|$$

T^j is a subset of P containing the j highest x_r scores. So, the reference point is a point in the PF whose cumulative traversal distance is closest to the α-weighted PF distance travelled from the first point in the PF. The reference point s_α is how far the PF is traversed, from the pair with the best Rel score to the one with the best Fair score, multiplied by α. As the PFs may have different density of points along the frontiers, the reference point is not computed based on a percentile (e.g., median) to avoid bias towards the denser part. Next, the distance between each model's (x_r, x_f) scores and the reference point s_α is computed with a distance measure d that accommodates 2d-vectors. The model with the closest distance is the best model in terms of both relevance and fairness, given the weight α. We call this *Distance to Pareto frontier of Fairness and Relevance* (DPFR).

5.3.4 Efficient computation of Pareto Frontier

Generating the PF as in §5.3.2 is costly. An efficient alternative is to compute a subset of the PF. We pick a fixed amount of Pareto-optimal solutions to compute,

p (e.g., 10). However, to reliably approximate the PF, these solutions should be spread according to the PF distribution, as opposed to e.g., only computing the first p points of the PF. The spread of the points is important, as the reference point in DPFR is computed based on the overall estimated PF. In the estimated PF, the first point corresponds to the measure scores of the initial recommendation given by the Oracle, and the rest are spread evenly throughout the PF generation. To select at which point of the Oracle2Fair algorithm the measures should be computed, we first estimate the total number of replacements needed by examining the distribution of recommended items frequency. This is done by getting the individual frequency count of all items in the recommendation, and subtracting the ideal upper bound of item count $\lceil km/n \rceil$ (§5.3.2) from each count. The number of expected replacements is the sum of the difference between the item frequency count and the ideal upper bound of item count in §5.3.2. Items with recommendation frequency counts less than the upper bound are excluded. With the estimated total number of replacement $numRep$, we compute the measures every $numRep$ $\text{div}(p-1)$ replacements done by Oracle2Fair, such that the measures are computed a total of $p-1$ times during the replacement process + 1 time before the replacement starts. These p points are spread evenly in terms of distance in the PF, which is important as DPFR is a distance-based measurement.

5.4 Experimental Setup

We study how our joint evaluation approach, DPFR, compares to existing single- and multi-aspect evaluation measures of relevance and fairness. Next, we present our experimental setup. The experiments are run in a cluster of CPUs and GPUs (e.g., Intel(R) Xeon(R) Silver 4214R CPU @ 2.40GHz, AMD EPYC 7413 and 7443, Titan X/Xp/V, Titan RTX, Quadro RTX 6000, A40, A100, and H100).

Datasets. We use six real-world datasets of various sizes and domains: e-commerce (Amazon Luxury Beauty, i.e., Amazon-lb [156]), movies (ML-10M and ML-20M [94]), music (Lastfm [32], videos (QK-video [247]), and jokes (Jester [87]). The datasets are as provided by [260], except for QK-video, which we obtain from [247]. Statistics of the datasets are in Tab. 5.1 and extended statistics are in App. 5.7.1.

Preprocessing. We remove duplicate interactions (we keep the most recent). We keep only users and items with ≥ 5 interactions. We convert ratings equal/above the following threshold to 1 and discard the rest: for Amazon-lb and ML-*, the threshold is 3, as their ratings range between $[1, 5]$ and $[0.5, 5]$ resp.; the threshold for Jester is 0, as the ratings range in $[-10, 10]$. Lastfm and QK-video have no ratings. QK-video

Table 5.1: Statistics of the preprocessed datasets.

dataset	**#users** (m)	**#items** (n)	**#interactions**	**sparsity (%)**
Lastfm [32]	1,859	2,823	71,355	98.64%
Amazon-lb [156]	1,054	791	12,397	98.51%
QK-video [247]	4,656	6,423	51,777	99.83%
Jester [87]	63,724	100	2,150,060	66.26%
ML-10M [94]	49,378	9,821	5,362,685	98.89%
ML-20M [94]	89,917	16,404	10,588,141	99.28%

has several interaction types; we use the 'sharing' interactions. For Jester, we remove users with > 80 interactions, to provide a large enough number of item candidates for recommendation during testing.[3]

Data splits. To obtain the train/val/test sets for Amazon-lb and ML-*, we use global temporal splits [151] with a ratio of 6:2:2 on the preprocessed datasets. Global random splits with the same ratio are used for the other datasets that have no timestamps. From all splits, we remove users with < 5 interactions in the train set.

Measures. We measure relevance (Rel) with Hit Rate (HR), MRR, Precision (P), Recall (R), MAP, and NDCG. We measure individual item fairness (Fair), as per [177], with Jain Index (Jain) [101, 264], Qualification Fairness (QF) [264], Gini Index (Gini) [85, 147], Fraction of Satisfied Items (FSat) [167], and Entropy (Ent) [167, 200].[4] We also use joint measures (Fair+Rel) as per [178]: Item Better-Off (IBO) [191],[5] Mean Max Envy (MME) [191], Inequity of Amortized Attention (IAA) [19, 23], Individual-user-to-individual-item fairness (II-F) [56, 234], and All-users-to-individual-item fairness (AI-F) [56]. We denote by ↑/↓ measures where higher/lower is better. DPFR is computed with Euclidean distance and $\alpha = 0.5$ (PF midpoint) for all datasets, so the midpoint is based on the empirically achievable scores, per dataset and measure pairs. For all runs, we use $k = 10$.

Recommenders. We use 4 common collaborative filtering-based recommenders: ItemKNN [55], BPR [185], MultiVAE [135], and NCL [139], with RecBole [260] and tune their hyperparameters. We train for 300 epochs with early stopping, and keep

[3]Some users in Jester have interacted with almost all 100 items. If a user has 80 items in the train/val set, there would only be 20 candidate items to recommend during test, which makes it easier to achieve higher relevance.

[4]These are set-based measures, but we do not expect the conclusions to differ for rank-based measures (App. 5.7.5).

[5]The measure Item Worse-Off is not used as its formulation is highly similar to IBO.

the configuration with the best NDCG@10 during validation. Each user's train/val items are excluded from their recommendations during testing.

Fair Re-rankers. To have fairer recommendations, we reorder the top k' items that are pre-optimised for relevance. Ideally $k' > k$ to allow exposing items that are not in the top k. As there are few relevant items per user in RS data,[6] k' should not be too big (e.g., 100). So, we re-rank the top $k' = 25$ items per dataset and model using three methods: GS, CM, and BC (explained below). We re-rank separately per user for CM and BC, or altogether for GS, for all $k'm$ recommended items, where m is the number of users. Other fair ranking methods exist but cannot be used as they apply to group or two-sided fairness only (e.g., [167, 249, 250]), or to stochastic rankings only (e.g., [161, 234]), or do not scale to larger datasets (e.g., [19, 191]).

1. Greedy Substitution (GS) [227] is a re-ranker for individual item fairness. We modify the GS algorithm, to replace the most popular items with the least popular ones, both considering how many times an item is at the top k' recommendations for all users (App. 5.7.3). As such, items can be swapped across users. To determine which items are most popular (i.e., to be replaced) and least popular (replacement items), the parameter $\beta = 0.05$ is used. We pair these two item types, and for each pair, we calculate the loss of (predicted) relevance if the items are swapped. We then replace up to 25% of the initial recommendations, starting from item pairs with the least loss.

2. COMBMNZ (CM) [125] is a common rank fusion method. Two rankings are fused for each user: one based on the (min-max) normalised predicted relevance score and another based on the coverage of each top k' item (to approximate fairness). We calculate item coverage only based on their appearance in the top k across all users and min-max normalise the score across all users. As favouring items with higher coverage would boost unfairness, we generate the ranking using 1 minus the normalised coverage. CM uses a multiplier based on the item appearance count in the two rankings above; this count is also only based on the top k. The resulting ranking is a fused ranking of fairness and relevance.

3. Borda Count (BC) is a common rank fusion method. For each user, we combine the original recommendation list and the rankings based on increasing item coverage, as in CM. Unlike CM, BC uses points. Higher points are given to items placed at the top. The result is a fused ranking of fairness and relevance.

[6] The median number of relevant items per user across all datasets is 2–53, see App. 5.7.1.

5.5 Experimental Results

We now present the evaluation scores of 16 runs (4 recommenders × 3 re-rankers, including no reranking) (§5.5.1). The relevance and fairness scores of these runs are the input to our DPFR approach. Not all combinations of evaluation measures are suited for PF. We explain this in §5.5.2. We present the generated PF (§5.5.3) and compare existing measures to DPFR (§5.5.4). We compare the results of efficient DPFR to other joint evaluation approaches (§5.5.5).

5.5.1 Groundwork Runs

The scores of Rel, Fair, and Fair+Rel measures for our 16 runs are shown in the appendix (Tab. 5.6–5.7). Two main findings emerge from Tab. 5.6–5.7. First, for all six datasets, **none of the best models according to Rel are also the best according to Fair measures**. This is similar to our toy example (Fig. 5.1), where one model ranks highest for fairness and another for relevance. Second, **the five Fair+Rel measures have no unanimous agreement on the best model**. IBO has a different best model from the others in 4/6 times, but sometimes agrees with one or more Fair measures. MME and AI-F agree on the best model 5/6 times, and sometimes agree on the best model with Fair measures. The best model according to IAA and II-F is always the same, and 4/6 times the same as the best model based on the Rel measures. The overall picture is inconclusive, with some Fair+Rel measures aligning more with Fair measures, and others aligning more with Rel measures.

5.5.2 Measure Compatibility with DPFR

Which pairs of Rel and Fair measures are suitable to generate the PF? We answer this based on the PF slope. The slope is calculated using the two endpoints of the PF, i.e., the start and end of the Oracle2Fair algorithm. A slope of zero means the Rel scores of the PF vary, but the Fair scores do not. As we compute the PF for multiple measures simultaneously, we expect a zero gradient for cases where the initial recommendation according to a Fair measure is already the fairest, even if other Fair scores are not. An undefined gradient value occurs when the initial recommendation is already the fairest and the most relevant according to a pair of Fair and Rel measures. Thus, we posit that a PF with a gradient value other than zero or undefined makes the corresponding pair of measures fit for PF generation (it allows for trade-offs in both aspects). The Rel-Fair measure pairs that are fit

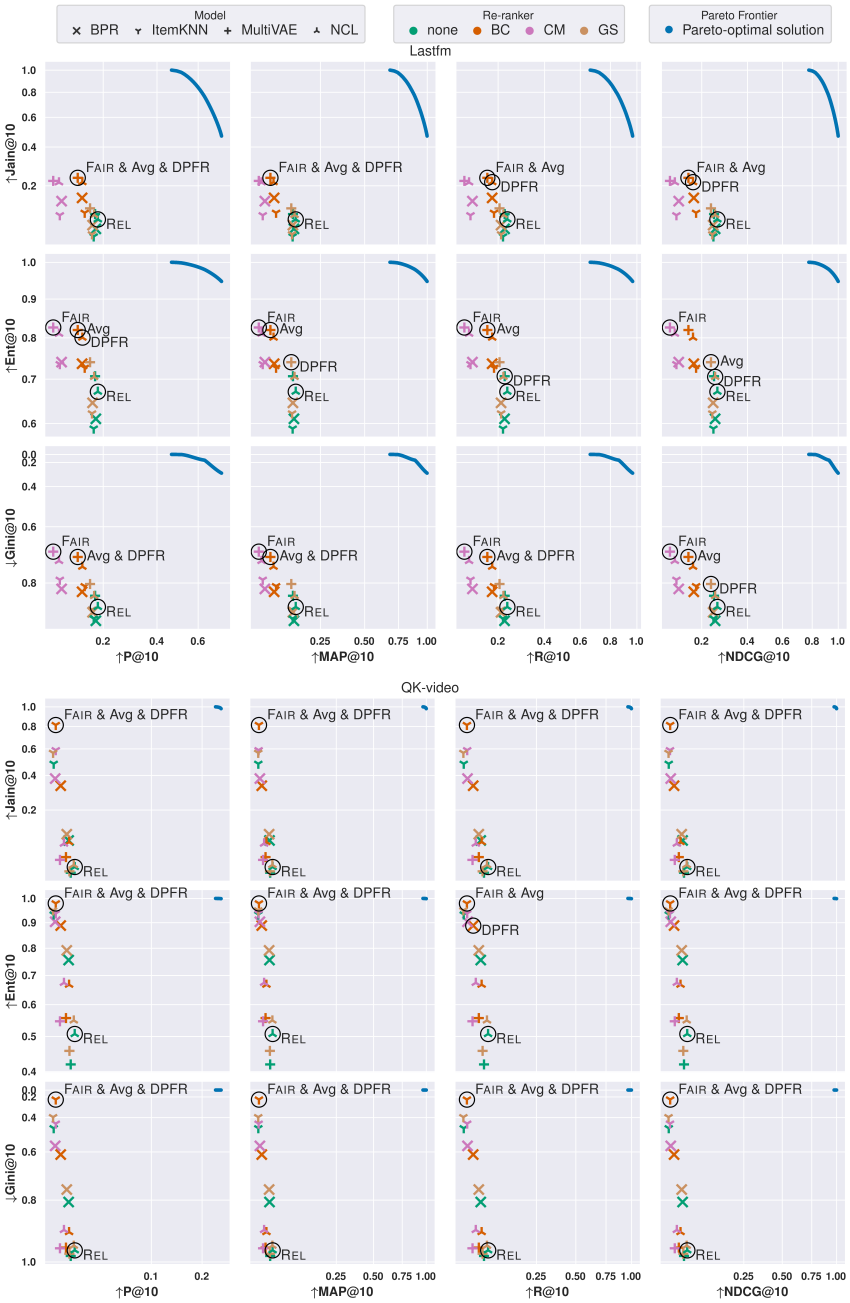


Figure 5.2: Pareto Frontier of fairness and relevance (in blue) and recommender scores for Lastfm and QK-video on exponential-like scales. REL, FAIR, Avg (mean of REL, FAIR), and DPFR are the best model per evaluation approach.

for DPFR based on their gradient are: {P, MAP, R, NDCG} × {Jain, Ent, Gini} (App. 5.7.4, Tab. 5.8). Only results from these pairs are shown henceforth. Next, we explain what causes an undefined or zero gradient.

Causes of zero/undefined gradient. Generating the PF requires a ranking of items. Any score that is based on a single relevant item, e.g., HR and MRR, is by design not suitable. Out of the Fair measures, QF and FSat sometimes behave inconsistently depending on the dataset properties, as follows. A dataset with relatively few relevant items can already be made maximally fair at the start of the PF generation, as QF quantifies fairness with ignorance to frequency of exposure; the score does not change as long as the same set of items appears in the top k recommendations of all users, no matter how many times each. When all items in the dataset already occur in the initial recommendations of our Oracle, nothing can be done to improve QF. For FSat, in few cases, the score is already maximum at the start of the PF generation. A maximum FSat score is achieved when all items in the dataset have at least the maximum possible exposure, if the available recommendation slots are shared equally across all items.[7] In principle, QF and FSat can still be used for DPFR when the initial recommendation by Oracle is not the maximum yet. Otherwise, the interpretation would be less meaningful in joint evaluation, as there is no trade-off between different aspects.

5.5.3 The Generated PF

Fig. 5.2 shows the PF plots of the pairs of Fair and Rel measures that are suited for DPFR, only for Lastfm and QK-video, which are representative of the overall trends in all our datasets (see Fig. 5.4 in the Appendix). The scores plotted are those computed in §5.3.2. The corresponding scores of our recommendation models are in App. 5.7.4. We see that, as the recommendations are made fairer, the generated PF for all datasets is a series of monotonic scores of Fair, specifically monotonic increasing Fair scores (except ↓Gini), and the remaining measures are monotonic decreasing. The monotonic property theoretically and empirically holds for the Fair measures, as we replace an item with the most exposure by another item with the least exposure, thereby making the recommendation fairer. Note that some users do not have exactly k items in the test set, so the perfect relevance score cannot

[7] $\lfloor km/n \rfloor$ times (the total number of recommendation slots across users divided by the number of items).

be reached for Precision@k and Recall@k [153]. NDCG and MAP are implemented with normalisation[8] so that they can still achieve a score of 1 in this situation.

The datasets which were randomly split as they have no timestamps (QK-video, Jester) have relatively short, compact PF. This happens because the random split results in a uniform distribution of items in each split, which means that items in the test split are quite diverse (64–100% of all unique items in the dataset). Considering that the Oracle2Fair algorithm starts by recommending items in the test split and stops when the recommendation reaches the fairest, there is not much room for change in Fair scores, as the initial recommendation is already rather fair. Additionally, there are not many relevant items per user in these datasets (i.e., the median for both datasets is 6 or less); random non-relevant items were chosen to make up for the remaining recommendations.[9] Thus, the PF generation decreases relevance only marginally in 2/6 cases. Correspondingly, we find that in QK-video and Jester, there exist Pareto-optimal recommendations, that are close to maximally fair and maximally relevant, with the exception of P@10. These can be seen in the measure pairs of {MAP, R, NDCG} $\times$ {Jain, Ent, Gini}, where the PF is close to the coordinates of $(1, 1)$, or $(1, 0)$ for the measure pairs with Gini. Thus, in theory, a fair recommendation does not necessarily have to sacrifice relevance.

5.5.4 Agreement between Measures

We study the agreement between DPFR and other evaluation approaches in ranking our 16 runs from best to worst. Low agreement means that the other approaches have few ties to the Pareto-optimal solutions that DPFR uses, and vice versa. We compare DPFR to (a) existing Rel and Fair measures, (b) existing joint Fair+Rel measures (§5.4), and also (c) the average (arithmetic mean) of Fair and Rel scores from the selected measure pairs that are used to generate the PFs. To compute the average for a measure where lower values are better (i.e., Gini), we compute 1−the Gini score instead.

5.5.4.1 Comparison of existing measures to DPFR

We find that for all datasets and all measure pairs, **the best model as per DPFR is always different from the best model as per Rel measures**. Moreover, **half the time, the best model as per DPFR is different from the best**

[8]Only the first $\min(|R_u^*|, k)$ items in a user u's recommendations are considered, where R_u^* is the set of relevant items for user u.

[9]The randomly-split Lastfm does not have a short PF because on average it has more relevant items per user compared to QK-video and Jester (see App. 5.7.1).

model as per a Fair measure. Existing Fair+Rel measures tend to have the same best model as either Fair or Rel measures (73.3% of the time), instead of having a more balanced evaluation of both aspects. These findings are expected as existing joint evaluation measures use relevance in their formulation differently than the Rel measures. Overall, the best model found with DPFR is less skewed towards relevance or fairness.

5.5.4.2 Correlation of measures

For each dataset, we compute the Kendall's[10] τ [113] correlations between the ranking given by DPFR and by the joint evaluation baselines (see Fig. 5.3). Rankings are considered equivalent if $\tau \geq 0.9$ [143, 223]. We see similar agreement trends in datasets where recommenders have higher Rel scores (Lastfm and Jester) or lower (Amazon-lb and QK-video).

Overall, most times DPFR orders models differently ($\tau < 0.9$) than all Fair+Rel measures except AI-F. We see similar trends between IAA and II-F, and between MME and AI-F. IBO can have similar trends as MME (except for Amazon-lb and QK-video). For all datasets, IAA and II-F have overall either weak or negative τ with DPFR (e.g., $[-0.2, 0.25]$ for ML-10M and $[-0.62, -0.35]$ for QK-video). A notable exception is DPFR with {MAP, R, NDCG} $\times$ Ent for Lastfm and Jester, where we see moderate correlations, $\tau \in [0.42, 0.68]$.

Ent differs from this trend because DPFR with {MAP, R, NDCG} $\times$ {Jain, Gini} has PF gradients of greater magnitude. This only affects Lastfm and Jester (they have higher Rel scores than the other datasets). DPFR with P has different patterns from other Rel measures: the raw DPFR scores of pairs involving P are lower on average, as the scores from Oracle do not start from 1, but much less, and therefore closer to the models' scores (Fig. 5.2). Meanwhile, IBO has varying τ across datasets: a huge range of τ, i.e., $[0.00, 0.9]$ for Lastfm, weak correlations $[0.01, 0.13]$ for Amazon-lb, and moderate to strong correlations $[0.67, 0.98]$ for ML-10M. These variations might be because IBO is based on the number of items satisfying a certain criterion, rather than an average of scores across users and/or items, i.e., how other Fair+Rel measures are defined.

Among the joint measures, AI-F correlates the strongest with DPFR, as both AI-F and DPFR, indirectly or directly, consider the recommendation frequency of each item and compare it with that of other items. However, the rank orderings given by AI-F are not equivalent to DPFR, as $\tau < 0.9$ for 5/6 datasets (excl. Amazon-lb). For the same measure-pair and between datasets, the τ of AI-F and DPFR also varies

[10] Ties are handled, unlike in Spearman's ρ.

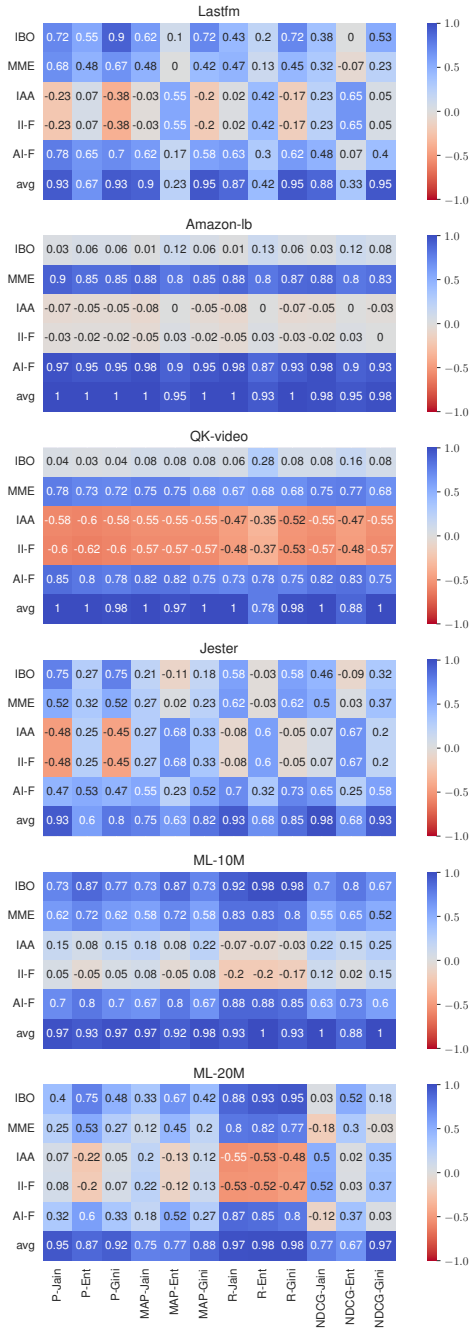


Figure 5.3: Kendall's τ correlation heatmap between the rank ordering of existing joint evaluation measures (including the average of Fair and Rel scores, avg), and DPFR.

a lot. E.g., $\tau = 0.07$ for NDCG-Ent for Lastfm, but $\tau = 0.9$ for Amazon-lb. We thereby do not recommend using any of the Fair+Rel measures (none correlates with Pareto optimality).

Taking the mean of Fair and Rel scores (avg) at a glance seems to correlate highly with DPFR. However, while it gives equivalent rankings ($\tau \geq 0.9$) in some cases (e.g., for Amazon-lb, most of ML-10M and QK-video, and half of ML-20M), it only does so for (1) datasets with lower Rel scores (Amazon-lb, QK-video), i.e., in cases where all models perform poorly, we have low variance in Rel, which leads to fairness dominating both avg and DPFR; (2) datasets with low variance in Fair scores (ML-*). In such cases, quantifying the evaluation jointly is challenging as one aspect dominates over the other. In the other datasets, the rank ordering given by the average is inconsistent: sometimes $\tau \geq 0.9$ for one dataset, but not for the others. This inconsistency between datasets holds for all measure pairs, except for P-Jain and NDCG-Gini. Due to these inconsistencies, we discourage using the arithmetic mean.

Overall, our correlation analysis shows that existing joint Fair+Rel evaluation measures cannot be used as a reliable proxy for DPFR.

5.5.4.3 Best model disagreement

We take a closer look at how DPFR relates to computing averages, as they are similar approaches in terms of combining scores from a measure pair. As comparing the raw scores of DPFR and the average is invalid, we instead count the disagreement between the best model based on DPFR and the mean of Fair and Rel scores (Tab. 5.2). The aim is to study whether one would come to the same conclusion regarding the best model, using the two different joint evaluation approaches.

Among the 12 measure pairs that are fit for DPFR, we find that **the best model according to DPFR is not always the same according to the average of Fair and Rel scores of the same pair**; in one case the disagreement is up to 58.33% (for Lastfm). The disagreement is generally much higher in the more complex rank-based measures (0–83.33%) compared to simpler set-based Rel measures (0–50.00%). Therefore, there are many cases where the mean of Fair and Rel scores is not the best case, especially for Lastfm and Jester where Rel scores are higher and vary more. In these two datasets, more often than not, DPFR leads to different conclusions than a simple average. Yet, sometimes the average agrees with DPFR in the best model: for QK-video, disagreement is low (8.33%), and there is a perfect agreement on the best model for Amazon-lb; we posit that these are due to equally poor and low variance in the Rel scores. This is in line with our correlation anal-

Table 5.2: The percentage of best model disagreement when taking the mean of FAIR and REL scores as opposed to using DPFR, separated by the REL measure type. P@k and R@k are set-based, NDCG and MAP are rank-based. We only consider the 12 measure pairs with a nonzero, defined gradient (§5.5.2).

	Set-based	Rank-based	All
Lastfm	50.00	66.67	58.33
Amazon-lb	0.00	0.00	0.00
QK-video	16.67	0.00	8.33
Jester	16.67	83.33	50.00
ML-10M	0.00	66.67	33.33
ML-20M	0.00	50.00	25.00
All datasets	13.89	44.44	29.17

ysis. As there is a huge range of variability across datasets (0–58.33%), we do not recommend using a simple average to get the same result as DPFR, as it is unreliable and inconsistent. Generally, averaging fails to reach the same conclusion as DPFR almost half the time for the rank-based REL measures,[11] especially when the REL and FAIR scores vary highly.

5.5.5 Efficient DPFR

5.5.5.1 The efficiency of the PF generation

We study the efficiency of DPFR by comparing the PF, an estimated version of PF on a subset of points, and the FAIR+REL measures. The estimated version of PF uses 3–12 points as per §5.3.4. We compute the amount of points in the estimated PF as % of those in the PF, and the resulting computation times. One point in the PF translates to one round of computing all FAIR and REL measures, so fewer points mean faster. For brevity, we report the estimated PF with only 3, 6, and 12 points in Tab. 5.3 (see App. 5.7.4, Tab. 5.11 for extended results).

The PF (Fig. 5.2) has hundreds to tens of thousands of points each, while the estimated PF only contains 0.02–2.40% of the points, which means reduced computational complexity for the PF generation. In terms of actual computation time (Tab. 5.3), computing the PF with ORACLE2FAIR take 0.56–75.77 mins to compute,

[11]Note that the published version of the work does not contain the phrase "for the rank-based REL measures". This phrase has been added for clarity.

Table 5.3: Efficiency and effectiveness of PF, estimated PF (Est. PF), and FAIR+REL measures: % of data points in the Est. PF (% pts), computation time. Dist is the average distance between midpoints in the Est. PF and PF over 12 measure pairs. Minimum agreement (Min τ) is the Kendall's τ correlation between DPFR with PF and Est. PF. Both PFs compute 11 REL and FAIR measures simultaneously. The times for other evaluation measures are averaged (Avg/model) and summed (All models) over 16 model combinations.

			Lastfm	Amazon-lb	QK-video	Jester	ML-10M	ML-20M
#pts		PF	4882	847	499	16202	2781	3783
%pts		Est. PF (12 pts)	0.25	1.42	2.40	0.07	0.43	0.32
		Est. PF (6 pts)	0.12	0.71	1.20	0.04	0.22	0.16
		Est. PF (3 pts)	0.06	0.35	0.60	0.02	0.11	0.08
↓ Computation time (mins.)		PF	19.18	0.56	10.49	847.42	28.99	75.77
		Est. PF (12 pts)	2.02	0.19	4.23	552.16	1.90	2.60
		Est. PF (6 pts)	2.00	0.19	4.12	551.72	1.84	2.54
		Est. PF (3 pts)	2.01	0.19	4.07	552.26	1.82	2.52
	Avg/model	IBO	<0.3s	<0.3s	0.01	0.01	<0.3s	0.01
		MME	2.04	0.03	19.51	0.09	15.25	89.13
		IAA	<0.3s	<0.3s	0.01	0.02	0.01	0.02
		II-F	<0.3s	<0.3s	0.01	0.01	0.01	0.02
		AI-F	<0.3s	<0.3s	0.01	0.01	<0.3s	0.01
	All models	IBO	0.02	<0.3s	0.10	0.12	0.06	0.15
		MME	32.63	0.49	312.14	1.38	244.03	1426.10
		IAA	0.03	<0.3s	0.10	0.36	0.13	0.30
		II-F	0.04	<0.3s	0.16	0.14	0.1	0.25
		AI-F	0.03	<0.3s	0.11	0.13	0.07	0.17
↑ Min τ		Est. PF (12 pts)	0.95	1.00	1.00	0.98	0.98	0.97
		Est. PF (6 pts)	0.90	0.97	1.00	0.98	0.95	0.92
		Est. PF (3 pts)	0.78	0.98	1.00	1.00	0.97	0.75
↓ Dist.		Est. PF (12 pts)	0.01	0.02	0.00	0.00	0.01	0.01
		Est. PF (6 pts)	0.03	0.05	0.00	0.00	0.02	0.02
		Est. PF (3 pts)	0.03	0.05	0.00	0.00	0.03	0.05

but only 0.19–4.07 mins for the PFs estimated with 3 points for all datasets except Jester. For Jester, it takes ∼14 hours and the estimation only takes ∼9 hours. However, this is expected for Jester as it has 62K users in the test split, as opposed to the 3.5K or fewer in the other datasets (i.e., see Tab. 5.4 in App. 5.7.1). While computing the estimated PFs is on average slower than computing the joint measures IBO, IAA, II-F, and AI-F, it is expected as the (estimated) PFs compute 11 measures simultaneously. Yet, in most cases (except Amazon-lb and Jester), the estimated PFs are still faster to compute than the time to compute MME for one model per dataset, let alone to compute MME for all models. For ML-20M, computing the estimated PF is even up to 35 times faster than computing MME of one model.

5.5.5.2 The effectiveness of efficient DPFR

To what extent is the DPFR from the efficiently generated PF (estimated PF) a reasonable proxy for fairness-relevance joint evaluation using the PF, in terms of giving a similar ordering of models? We compare the DPFR from the PF and estimated PF using Kendall's τ correlations. Further, as DPFR is computed based on a α-based reference point lying on the PF, to quantify possible accuracy loss of the estimated PF, in Tab. 5.3 we show the error of the midpoint estimation. This error is the Euclidean distance between the reference point in the PF and estimated PF (i.e., the midpoint in our case), as per [225].

We first analyse the error of the midpoint estimation. For the 12 measure pairs and 6 datasets, the midpoint coordinates on average do not move much: the distance is 0.00–0.05, even when the PF is only estimated with 3 points. Ergo, the correlations between the rank ordering of models given by the DPFR of PF and its estimation, are still equivalent ($\tau \geq 0.9$) when estimated with 6 or 12 points [143, 223]. Even the 3-point estimation maintains high agreement ($\tau \in [0.75, 1]$), with only 5 cases having $\tau < 0.9$ across 6 datasets and 12 measure-pairs. Therefore, it is possible to only compute a small number of points in the PF, e.g., 6 or 12 points, and still make a reliable PF estimation, evidenced by the small shift of the PF midpoint and the rank ordering of the models remaining equivalent ($\tau \geq 0.9$), if not identical ($\tau = 1$) for all measure pairs and datasets.

5.6 Discussion and Conclusions

Recommendation evaluation has long used measures that quantify only relevance, but has recently shifted to include fairness. However, there exists no de-facto way to robustly quantify these two aspects. We propose a novel approach (DPFR) that

uses fairness and relevance measures under a joint evaluation scheme for RSs. DPFR computes the empirical best possible recommendation, jointly accounting for a given pair of relevance and fairness measures, in a principled way according to Pareto-optimality. DPFR is modular, tractable, and intuitively understandable. It can be used with existing measures for relevance and fairness, and allows for different trade-offs of relevance and fairness. We empirically show that existing evaluation measures of fairness w.r.t. relevance [19, 23, 56, 191, 234] behave inconsistently: they disagree with optimal solutions based on DPFR computed on more robust and well-understood measures of relevance, such as NDCG, and fairness, such as Gini. We uncover some weaknesses of these measures, but more research is warranted to study their behaviour. Admittedly, existing joint measures are not originally defined to be aligned with existing relevance and fairness measures [85, 101, 147, 148, 167, 264], so it is not surprising that they have different results from DPFR. However, existing measures show varying performance also from each other and from well-understood relevance and fairness measures. Thus, DPFR can be a viable alternative for robust, interpretable, and provenly optimal evaluation in offline scenarios. We also show that DPFR can be computed fast while reaching equivalent conclusions. Overall, DPFR demonstrates distinct benefits in mitigating false conclusions by up to 58% compared to basic aggregation methods like averaging.[12] Surprisingly, simple averaging aligns more with our Pareto-optimal based DPFR, than existing joint measures. We recommend combining either MAP-Ent or NDCG-Ent: the conclusions are distinguishable from simply averaging, or taking the best model based on fairness or relevance measures.

Acknowledgements

The work is supported by the Algorithms, Data, and Democracy project (ADD-project), funded by Villum Foundation and Velux Foundation. Qiuchi Li contributed to the idea of computing the reference point. We thank the DIKU IR Lab and the anonymous reviewers, who have provided helpful feedback to improve earlier versions of the manuscript.

[12]The published version concluded "by up to 50%", but the performance of DPFR is better than previously reported.

5.7 Appendix

5.7.1 Extended Dataset Statistics

Tab. 5.4 presents the statistics of each dataset split. For several datasets (e.g., Amazon-lb and ML-*), the number of users in the test split is significantly less than the number of users in the train split. Tab. 5.5 presents the statistics of items in the test split, per user.

Table 5.4: Number of [users, items, and interactions] in the train, validation, and test split after preprocessing.

	train	val	test
Lastfm	[1842, 2821, 42758]	[1831, 2448, 14248]	[1836, 2476, 14237]
Amazon-lb	[1054, 552, 8860]	[470, 204, 1811]	[437, 209, 1726]
QK-video	[4656, 6245, 34345]	[3470, 4095, 8726]	[3514, 4101, 8706]
Jester	[63724, 100, 1294511]	[62137, 100, 427623]	[62167, 100, 427926]
ML-10M	[49378, 6838, 4944064]	[2695, 7828, 296914]	[1523, 7880, 121707]
ML-20M	[89917, 8719, 9882504]	[4987, 10742, 472243]	[2178, 13935, 233394]

Table 5.5: Statistics of items in the test split, per user, i.e., the number of relevant items per user.

	mean	min	median	max
Lastfm	7.75	1	8	19
Amazon-lb	3.95	1	3	16
QK-video	2.48	1	2	16
Jester	6.88	1	6	29
ML-10M	79.91	1	46	1632
ML-20M	107.16	1	53	2266

5.7.2 Algorithms for Generating Pareto Frontier

We present the pseudocodes of the algorithms for generating the Pareto Frontier: the Oracle (Algorithm 1) and Oracle2Fair (Algorithm 2). We provide the worst-case time complexity analysis for both algorithms and discuss a possible edge case. We

denote as k the recommendation cut-off, as m the number of users, as n the number of items, as H the maximum number of items in a user's interaction history (H_u) across all users, and as R the maximum number of items in a user's test split (R_u^*) across all users. We use the binary logarithm ($\log_2$). For brevity, we omit lines with $O(1)$, and we state the reasoning for each line (L).

5.7.2.1 Time complexity of the Oracle (Algorithm 1)

The line-by-line time complexity analysis of Oracle (Algorithm 1) is as follows:

- L1: $O(km)$, creating list of size k for m users
- L2–17: $O(R^2m + Rm\log m + Rmk\log k)$, resulting from at most R iterations of:
  - L3: $O(m)$, linear search on list of size m
  - L4–7: $O(mR)$, at most m times looking up at most R values
  - L8: $O(m\log m)$, sorting a list of size m
  - L11–16: $O(mk\log k)$, at most m times sorting list of size k
- L18: $O(m)$, linear search on list of size m
- L19: $O(m)$, m assignment operations
- L21–36: $O(k^2m^2)$, resulting from at most m iterations of:
  - L22–29: $O(kH)$, at most k times of linear search on list of size H
  - L30–L35: $O(k^2m)$, k times of counting in a list of size km.
    In most cases, the number of users $m >> H$, so the time complexity of this block is dominated by $O(k^2m)$.
- L37: $O(km)$ computing relevance/fairness measures for m users based on recommendation list of size k

Overall, the time complexity of Oracle is dominated by that of L2 and L21–36. Hence, the time complexity of Oracle is $O(R^2m + Rm\log m + Rmk\log k + k^2m^2)$.

Algorithm 1: Oracle
Create recommendations with the highest relevance

```
Data:
I: all items in the dataset;
H_u: items in train-val split for each user u ∈ U;
R*_u: items in test split (relevant items) for each user u ∈ U;
k: number of recommended items
Result:
rec: most relevant recommendation
result: a list of relevance and fairness scores
itemNotInRec: items that are not in the recommendation
/* Handle users with exactly |R*_u| = k                                  */
foreach u ∈ U where |R*_u| = k do rec[u] ← R*_u;
/* Handle users with |R*_u| > k                                          */
for K = k + 1 to max(|R*_u|) do
    userWithK ← get users where |R*_u| = K
    foreach u ∈ userWithK do
        takenItem[u] ← R*_u ∩ rec
        weight[u] ← sum(countInRec(takenItem[u]))
    end
    sortUserWithK ← sort userWithK by the least weight
    tempRec[u] ← R*_u \ takenItem[u]
    keep only max k items in tempRec[u]
    foreach u ∈ sortUserWithK do
        rec[u] ← tempRec[u]
        numItemToAdd ← k − |tempRec[u]|
        sort takenItem[u] by the least item count
        rec[u].append(takenItem[u][: numItemToAdd])
    end
end
/* Handle users with |R*_u| < k                                          */
remainUser ← get users where |R*_u| < k
foreach u ∈ remainUser do rec[u] ← |R*_u|;
itemNotInRec ← I \ rec
foreach u ∈ remainUser do
    while |rec[u]| < k and itemNotInRec ≠ ∅ do
        for item ∈ itemNotInRec do
            if item ∉ H_u then
                rec[u].append(item)
                itemNotInRec ← itemNotInRec \ {item}
            end
        end
    end
    if |rec[u]| < k then
        while |rec[u]| < k do
            candItem ← least popular item in rec that is not in H_u ∪ R*_u ∪ rec[u]
            rec[u].append(candItem)
        end
    end
end
result ← calculateScores(rec)
```

Algorithm 2: ORACLE2FAIR
After recommending maximally relevant items, iteratively change the recommendation list to increase fairness until maximum fairness is reached

Data: H_u, R_u^*, I, k
Result:
rec: most fair possible recommendation;
$result$: a list of relevance and fairness scores

1 $rec, result, itemNotInRec \leftarrow Oracle(I, H_u, R_u^*)$
`/* Get the most popular item in the recommendations & its frequency count */`
2 $newMostPop \leftarrow mostPop \leftarrow getMostPopItem(rec)$
3 $newCntPop \leftarrow cntPop \leftarrow cnt(mostPop, rec)$
4 $uWithMostPop \leftarrow$ all users with $mostPop \in rec[u]$
5 sort $uWithMostPop$ by largest index of $mostPop$ in $rec[u]$
6 **for** $i \in itemNotInRec$ **do**
7   **if** $cntPop = 1$ **then** break;
8   **if** $newMostPop \neq mostPop$ **then**
9         $mostPop \leftarrow newMostPop$
10   update $uWithMostPop$ following $mostPop$
11   **end**
12   **if** $newMostPop = mostPop$ **then**
13         $candU \leftarrow$ all u in $uWithMostPop$ where $i \notin H_u$
14   **if** $\exists u \in candU$ *with* $i \in R_u^*$ **then**
15   recommend i to the top u from $candU$ with $i \in R_u^*$
16   **end**
17   **else** recommend i to the top u from $candU$;
18   reorder $rec[u]$ so all relevant items are at the top
19         $result.append(calculateScores(rec))$
20         $itemNotInRec \leftarrow itemNotInRec \setminus \{i\}$
21         $newMostPop \leftarrow getMostPopItem(rec)$
22         $newCntPop \leftarrow cnt(mostPop, rec)$
23   **end**
24 **end**
25 **else**
26   do lines 2–5
27     $i \leftarrow leastPop \leftarrow getLeastPopItem(rec)$
28     $m, n \leftarrow |U|, |I|$
29   **while** $cntPop > \lceil km/n \rceil$ **do**
30   **if** $newMostPop \neq mostPop$ **then** do lines 9–10;
31   **if** $newMostPop = mostPop$ **then**
32             $candU \leftarrow$ all u in $uWithMostPop$ where $i \notin H_u \cup rec[u]$
33   do lines 14–19
34   do lines 26–27
35   **end**
36   **end**
37 **end**

5.7.2.2 Time complexity of Oracle2Fair (Algorithm 2)

The line-by-line time complexity analysis of Oracle2Fair (Algorithm 2) is as follows:

- L1: $O(R^2m + Rm\log m + Rmk\log k + k^2m^2)$ (derived in §5.7.2.1)
- L2–4: $O(km)$, m times of counting and linear search on list of size k
- L5: $O(m\log m)$, sorting a list of size m (e.g., with Tim Sort)
- L6–24: $O(max(Hmn, kmn))$, resulting from at most n iterations of:
  - L10: $O(km)$, m times of linear search on list of size k
  - L13: $O(Hm)$, m times linear search on list of size at most H
  - L14: $O(km)$
  - L18: $O(k\log k)$, sorting a list of size k
  - L21–L22: $O(km)$; the term $O(k\log k)$ is dominated by $O(km)$, as typically the cut-off k is much smaller than the number of users m (and hence $\log k < m$).
- L26: $O(km + m\log m)$, combining L2–4 and L5
- L27: $O(km)$, m times counting on list of size k
- L29–34: $O(max(Hkm^2, k^2m^2) + km^2\log m)$, resulting from at most km iterations (derived below) of:
  - L30: $O(km)$, from L10
  - L32: $O(Hm)$, m times linear search on list of size at most H
  - L33: $O(km)$, from L14 and L18
  - L34: $O(km + m\log m)$, from L26–27

We estimate the worst-case complexity of the number of iterations of the code block in L29–34, by assuming that the initial recommendation is the unfairest. The unfairest recommendation happens when the same k unique items are recommended to all m users. This code block aims to reduce the max frequency count of all items to $\lceil km/n \rceil$. Hence, the number of iterations is $\sum^k(m - \lceil km/n \rceil) \leq k(m - \frac{km}{n} - 1) = km - \frac{k^2m}{n} - k \leq km$.

All in all, the time complexity of Oracle2Fair is dominated by that of L1, L6–24, and L29–34. Hence, the overall time complexity of Oracle2Fair is $O(R^2m + Rm\log m + Rmk\log k + k^2m^2 + km^2\log m + kmn)$ if $k \geq H$, or $O(R^2m + Rm\log m + Rmk\log k + Hkm^2 + km^2\log m + Hmn)$ otherwise. To simplify the time complexity, further assumptions need to be made for one or more variables.

5.7.2.3 Possible edge cases

There might be edge cases, for example, datasets where the train/val set of each user contains almost all items in the dataset (each user has rated/clicked most items in the dataset). In this case, if we do not want to re-recommend items in the train/val set to users, some items may have item counts more than $\lceil km/n \rceil$ at the end of the process, and Algorithm 2 might not halt. However, such datasets are rare in recommender systems.

5.7.3 Modifications to the GS Algorithm

The original GS algorithm [227] increases individual item fairness within clusters of similar items. The item similarity is determined based on the item embedding. As our experiments and the Fair measures do not deal with the additional constraint of item similarity, we consider all items as similar. Therefore, we only have a single cluster of items.

On top of that, we also modify GS to increase computational efficiency. In the original GS algorithm, for each pair of candidate items for replacement i and candidate items to be replaced i', the algorithm finds all users that have i in the original recommendation list. The algorithm then computes the loss in relevance (computed using predicted relevance value) if item i is replaced by i'. Until this point, our modified algorithm does the same. The difference is that we save each i, i', u, and the loss associated, while the original algorithm only saves the information for the one user u^*, whose recommendation list will suffer the least loss when we replace i with i'. The original GS then proceeds to make the replacement, update the pool of candidate items for replacement and to be replaced, and go through the entire process again. Initially, we found that with the GS algorithm, around 20% of the initial recommendations are replaced during the process, meaning that for Amazon-lb, there are at least $437 \times 10 \times 0.2 \geq 800$ iterations of the process (Tab. 5.4). The number of iterations is much bigger for ML-10M, which has more than three times the recommendation slots as Amazon-lb, and hence it is very costly to use the GS algorithm as is.

Our modified GS utilises the saved information earlier. After going through all pairs of (i, i'), we sort the saved list from the smallest to the largest loss, and (attempt to) perform the replacement using the first P pairs, where P is 25% of the number of recommendation slots. During the replacement process, if the item that is supposed to be replaced no longer exists in the user's recommendation list, we simply skip the replacement.

5.7.4 Extended Results

We present the actual scores of the recommender models in Tab. 5.6–5.7. In Tab. 5.8, we present the gradient values of the PF, used in determining which pair of measures are suitable for DPFR. In Fig. 5.4 we present the Pareto Frontier (PF) of fairness and relevance together with recommender model scores in Tab. 5.6–5.7 for Amazon-lb, Jester, and ML-*. In Tabs. 5.9 and 5.10 we present the DPFR scores for all datasets. In Tab. 5.11 we present the Kendall's τ correlation scores of the DPFR from estimated PF and the PF.

5.7.5 Further Discussions

5.7.5.1 The impact of replacing frequently recommended items

In this work, we replace frequently recommended items in two cases: during the Pareto Frontier generation (PF) with the Oracle2Fair algorithm (§5.3.2) and as part of the fair rerankers (§5.4). In both cases, none of the replacements significantly affect the overall recommendation performance:

1. Replacement in the Oracle2Fair algorithm. The Oracle2Fair algorithm is used to generate the recommendation lists whose scores make up the PF (not the scores from the model-generated recommendation lists, that we compare to a point in the PF). The replacements are done on lists separate from the model's recommendation lists. Therefore, the replacements in the Oracle2Fair algorithm do not affect the recommendation performance based on relevance and fairness measures.

2. Replacement in the fair rerankers. We look at recommendation performance based on NDCG (relevance) and Gini (fairness). In all six datasets, when comparing the best non-reranked model to its reranked versions (e.g., for Lastfm, it is NCL vs NCL-BC, NCL-CM, and NCL-GS), the decrease in NDCG is not more than 0.26, and the decrease is even below 0.15 in all datasets excluding Jester, (§5.7.4, Tabs. 5.6 and 5.7). For Jester, the 0.26 decrease in NDCG is exchanged for an improvement in Gini by 0.218. Hence, we believe that the impact of item replacement is reasonable.

5.7.5.2 Using rank-based fairness measures

Suppose we have two FAIR measures, e.g., set-based Gini (Gini) and rank-based Gini (Gini-w). In [177], the absolute scores of Gini and Gini-w do not differ considerably and the two measures correlate strongly with Kendall's $\tau \geq 0.9$. Hence, generating the PF with rank-based fairness measures such as Gini-w is not expected to result in significantly different conclusions from the set-based version.

Table 5.6: Relevance (REL), fairness (FAIR), and joint fairness and relevance (FAIR+REL) scores at $k = 10$ of the recommender models for Lastfm, Amazon-lb, and QK-video, without and with re-ranking the the top $k' = 25$ items using Borda Count (BC), COMBMNZ (CM), and Greedy Substitution (GS). The most relevant or most fair score per measure is in bold. ↑ means the higher the better, ↓ the lower the better.

		model	ItemKNN				BPR				MultiVAE				NCL			
		reranking	-	BC	CM	GS	-	BC	CM	GS	-	BC	CM	GS	-	BC	CM	GS
Lastfm	REL	↑ HR	0.765	0.742	0.581	0.750	0.773	0.729	0.587	0.751	0.778	0.693	0.523	0.734	**0.793**	0.726	0.571	0.765
		↑ MRR	0.484	0.333	0.270	0.481	0.492	0.323	0.280	0.488	0.476	0.285	0.232	0.470	**0.503**	0.311	0.260	0.499
		↑ P	0.172	0.147	0.089	0.167	0.178	0.140	0.092	0.169	0.176	0.129	0.076	0.161	**0.184**	0.141	0.087	0.173
		↑ MAP	0.137	0.085	0.053	0.135	0.141	0.080	0.058	0.138	0.138	0.070	0.045	0.132	**0.148**	0.079	0.050	0.144
		↑ R	0.218	0.186	0.114	0.211	0.224	0.180	0.119	0.211	0.224	0.163	0.098	0.205	**0.234**	0.180	0.110	0.220
		↑ NDCG	0.245	0.181	0.119	0.241	0.252	0.173	0.126	0.244	0.247	0.155	0.102	0.235	**0.261**	0.170	0.115	0.252
	FAIR	↑ Jain	0.042	0.101	0.094	0.046	0.058	0.151	0.140	0.067	0.097	**0.236**	0.222	0.115	0.082	0.216	0.215	0.095
		↑ QF	0.474	0.642	**0.679**	0.533	0.362	0.491	0.528	0.402	0.517	0.658	0.678	0.554	0.453	0.622	0.657	0.502
		↑ Ent	0.589	0.727	0.735	0.622	0.610	0.736	0.740	0.646	0.707	0.820	**0.826**	0.740	0.671	0.801	0.810	0.705
		↑ FSat	0.129	0.197	0.216	0.152	0.147	0.211	0.228	0.177	0.202	0.293	**0.321**	0.249	0.178	0.269	0.286	0.221
		↓ Gini	0.904	0.810	0.790	0.879	0.910	0.827	0.818	0.887	0.839	0.715	**0.696**	0.803	0.872	0.748	0.728	0.840
	FAIR+REL	↑ IBO	0.209	0.270	0.256	0.227	0.208	0.263	0.253	0.228	0.261	**0.314**	0.278	0.281	0.242	0.308	0.292	0.265
		↓ MME	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.001	0.001	0.001	**0.000**	0.001	0.001
		↓ IAA	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	**0.004**	0.004	0.004	0.004
		↓ II-F	0.001	0.001	0.002	0.001	0.001	0.001	0.002	0.001	0.001	0.001	0.002	0.001	**0.001**	0.001	0.002	0.001
		↓ AI-F	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.000**	0.000	0.000
Amazon-lb	REL	↑ HR	**0.046**	0.021	0.016	0.043	0.011	0.014	0.021	0.011	0.039	0.007	0.014	**0.046**	0.034	0.021	0.011	0.034
		↑ MRR	0.020	0.011	0.011	0.020	0.003	0.005	0.007	0.003	0.023	0.003	0.004	**0.024**	0.022	0.006	0.003	0.022
		↑ P	0.005	0.002	0.002	0.005	0.001	0.001	0.002	0.001	0.004	0.001	0.002	**0.005**	0.004	0.002	0.001	0.004
		↑ MAP	0.006	0.004	0.004	0.006	0.002	0.003	0.004	0.002	0.006	0.002	0.003	0.006	**0.006**	0.002	0.001	0.006
		↑ R	**0.013**	0.007	0.005	0.013	0.005	0.008	0.010	0.005	0.010	0.005	0.008	0.012	0.012	0.007	0.003	0.011
		↑ NDCG	0.011	0.006	0.005	0.011	0.003	0.005	0.006	0.003	0.010	0.003	0.004	**0.011**	0.011	0.004	0.002	0.011
	FAIR	↑ Jain	0.271	**0.547**	0.431	0.324	0.223	0.432	0.359	0.259	0.035	0.123	0.097	0.043	0.026	0.098	0.080	0.031
		↑ QF	0.650	**0.679**	0.612	0.663	0.549	0.630	0.594	0.571	0.222	0.294	0.286	0.254	0.229	0.315	0.310	0.265
		↑ Ent	0.802	**0.882**	0.839	0.829	0.747	0.839	0.809	0.776	0.418	0.587	0.558	0.469	0.371	0.560	0.534	0.426
		↑ FSat	0.370	**0.538**	0.438	0.435	0.314	0.410	0.376	0.364	0.114	0.159	0.152	0.138	0.091	0.146	0.138	0.115
		↓ Gini	0.665	**0.492**	0.598	0.613	0.747	0.601	0.660	0.703	0.949	0.882	0.899	0.930	0.959	0.898	0.910	0.943
	FAIR+REL	↑ IBO	0.062	0.038	0.029	**0.067**	0.019	0.029	0.038	0.019	0.029	0.019	0.029	0.033	0.038	0.033	0.024	0.033
		↓ MME	0.001	0.001	0.001	0.001	0.001	**0.001**	0.001	0.001	0.003	0.001	0.001	0.003	0.004	0.001	0.001	0.004
		↓ IAA	0.011	0.011	0.011	**0.011**	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011	0.011
		↓ II-F	0.006	0.006	0.006	**0.006**	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006
		↓ AI-F	0.000	**0.000**	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.002	0.000	0.000	0.002
QK-video	REL	↑ HR	0.040	0.046	0.047	0.038	0.099	0.063	0.045	0.089	0.109	0.089	0.061	0.103	**0.130**	0.102	0.077	0.124
		↑ MRR	0.013	0.014	0.013	0.013	0.039	0.018	0.015	0.038	0.039	0.028	0.021	0.038	**0.048**	0.030	0.024	0.047
		↑ P	0.004	0.005	0.005	0.004	0.011	0.007	0.005	0.010	0.012	0.009	0.006	0.011	**0.014**	0.011	0.008	0.013
		↑ MAP	0.005	0.005	0.005	0.005	0.017	0.008	0.006	0.016	0.018	0.012	0.009	0.017	**0.022**	0.013	0.010	0.021
		↑ R	0.014	0.018	0.019	0.014	0.043	0.028	0.019	0.039	0.051	0.039	0.027	0.047	**0.061**	0.045	0.033	0.058
		↑ NDCG	0.009	0.011	0.010	0.009	0.029	0.015	0.011	0.027	0.031	0.022	0.016	0.030	**0.038**	0.025	0.019	0.037
	FAIR	↑ Jain	0.483	**0.815**	0.589	0.567	0.081	0.333	0.379	0.101	0.012	0.038	0.032	0.014	0.020	0.076	0.071	0.023
		↑ QF	0.901	**0.956**	0.790	0.924	0.625	0.809	0.823	0.678	0.100	0.155	0.163	0.127	0.201	0.331	0.365	0.253
		↑ Ent	0.933	**0.979**	0.937	0.950	0.755	0.888	0.903	0.792	0.420	0.557	0.547	0.458	0.507	0.667	0.674	0.549
		↑ FSat	0.443	**0.659**	0.547	0.522	0.212	0.346	0.382	0.259	0.052	0.089	0.090	0.070	0.077	0.140	0.150	0.104
		↓ Gini	0.472	**0.235**	0.442	0.397	0.807	0.613	0.570	0.761	0.982	0.957	0.959	0.976	0.966	0.909	0.902	0.952
	FAIR+REL	↑ IBO	0.033	0.038	0.038	0.035	0.054	0.050	0.036	0.052	0.031	0.042	0.036	0.033	0.043	**0.060**	0.054	0.047
		↓ MME	0.000	**0.000**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		↓ IAA	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	**0.001**	0.001	0.001	0.001
		↓ II-F	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	**0.001**	0.001	0.001	0.001
		↓ AI-F	0.000	**0.000**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 5.7: Relevance (Rel), fairness (Fair), and joint fairness and relevance (Fair+Rel) scores at $k = 10$ of the recommender models for Jester and ML-*, without and with re-ranking the the top $k' = 25$ items using Borda Count (BC), COMBMNZ (CM), and Greedy Substitution (GS) evaluated at $k = 10$. The most relevant or most fair score per measure is in bold. ↑ means the higher the better, ↓ the lower the better.

		model	ItemKNN				BPR				MultiVAE				NCL			
		reranking	-	BC	CM	GS	-	BC	CM	GS	-	BC	CM	GS	-	BC	CM	GS
Jester	Rel	↑ HR	0.933	0.888	0.652	0.932	0.929	0.876	0.742	0.928	**0.944**	0.899	0.818	0.944	0.939	0.893	0.804	0.939
		↑ MRR	0.632	0.443	0.307	0.632	0.635	0.455	0.322	0.635	**0.661**	0.465	0.370	0.661	0.651	0.479	0.349	0.651
		↑ P	0.334	0.250	0.144	0.333	0.330	0.243	0.163	0.329	**0.351**	0.262	0.194	0.351	0.342	0.257	0.185	0.341
		↑ MAP	0.352	0.198	0.101	0.352	0.348	0.195	0.112	0.348	**0.379**	0.208	0.145	0.379	0.367	0.211	0.133	0.367
		↑ R	0.529	0.393	0.197	0.529	0.524	0.377	0.255	0.523	**0.555**	0.405	0.324	0.555	0.543	0.400	0.305	0.542
		↑ NDCG	0.496	0.336	0.189	0.496	0.493	0.331	0.216	0.492	**0.525**	0.350	0.265	0.524	0.512	0.352	0.249	0.511
	Fair	↑ Jain	0.343	0.556	0.445	0.345	0.377	**0.583**	0.547	0.380	0.295	0.544	0.509	0.297	0.351	0.504	0.534	0.354
		↑ QF*	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.967	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		↑ Ent	0.702	0.854	0.784	0.705	0.754	**0.875**	0.857	0.757	0.648	0.852	0.839	0.651	0.722	0.838	0.855	0.725
		↑ FSat	0.267	**0.378**	0.289	0.267	0.244	0.344	0.333	0.244	0.256	0.344	0.300	0.256	0.222	0.344	0.311	0.222
		↓ Gini	0.687	0.502	0.595	0.685	0.632	**0.467**	0.495	0.629	0.738	0.506	0.520	0.735	0.668	0.528	0.502	0.665
	Fair+Rel	↑ IBO	0.600	**0.930**	0.740	0.600	0.840	0.910	0.780	0.840	0.500	0.870	0.810	0.500	0.740	0.920	0.780	0.740
		↓ MME	0.003	0.003	0.006	0.003	0.004	**0.002**	0.005	0.004	0.008	0.003	0.004	0.008	0.004	0.003	0.006	0.004
		↓ IAA	0.081	0.093	0.104	0.081	0.081	0.094	0.103	0.081	0.078	0.092	0.100	**0.078**	0.079	0.092	0.101	0.079
		↓ II-F	0.028	0.035	0.040	0.028	0.029	0.035	0.040	0.029	0.027	0.035	0.038	**0.027**	0.028	0.034	0.038	0.028
		↓ AI-F	0.002	0.002	0.003	0.002	0.002	**0.001**	0.002	0.002	0.003	0.002	0.002	0.003	0.002	0.002	0.002	0.002
ML-10M	Rel	↑ HR	0.487	0.480	0.443	0.481	0.512	0.462	0.386	0.485	0.417	0.438	0.387	0.410	**0.521**	0.473	0.402	0.513
		↑ MRR	0.282	0.242	0.225	0.279	0.299	0.208	0.185	0.295	0.237	0.231	0.191	0.235	**0.302**	0.216	0.203	0.301
		↑ P	0.137	0.128	0.105	0.133	0.146	0.114	0.088	0.132	0.107	0.111	0.096	0.105	**0.154**	0.123	0.094	0.149
		↑ MAP	0.089	0.074	0.060	0.086	0.095	0.061	0.047	0.088	0.067	0.067	0.054	0.066	**0.101**	0.067	0.052	0.099
		↑ R	0.022	0.022	0.018	0.022	0.025	0.019	0.012	0.023	0.020	0.021	0.016	0.021	**0.026**	0.020	0.013	0.026
		↑ NDCG	0.150	0.133	0.113	0.147	0.160	0.115	0.092	0.150	0.119	0.121	0.100	0.118	**0.167**	0.123	0.100	0.164
	Fair	↑ Jain	0.011	0.026	0.027	0.012	0.037	0.100	**0.115**	0.044	0.003	0.005	0.006	0.004	0.024	0.063	0.069	0.027
		↑ QF	0.044	0.062	0.068	0.047	0.145	0.199	**0.216**	0.160	0.014	0.021	0.025	0.016	0.086	0.123	0.132	0.094
		↑ Ent	0.407	0.503	0.514	0.418	0.596	0.697	**0.716**	0.624	0.238	0.302	0.324	0.258	0.519	0.625	0.638	0.544
		↑ FSat	0.044	0.062	0.068	0.047	0.145	0.199	**0.216**	0.160	0.014	0.021	0.025	0.016	0.086	0.123	0.132	0.094
		↓ Gini	0.987	0.973	0.971	0.985	0.945	0.895	**0.879**	0.932	0.997	0.994	0.993	0.996	0.969	0.936	0.930	0.963
	Fair+Rel	↑ IBO	0.031	0.043	0.046	0.034	0.069	0.089	**0.091**	0.076	0.012	0.016	0.018	0.014	0.054	0.073	0.074	0.058
		↓ MME	0.001	0.001	0.001	0.001	0.001	**0.001**	0.001	0.001	0.003	0.002	0.001	0.002	0.001	0.001	0.001	0.001
		↓ IAA	0.008	0.009	0.009	0.008	0.008	0.009	0.009	0.008	0.009	0.009	0.009	0.009	**0.008**	0.009	0.009	0.008
		↓ II-F	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.000**	0.000	0.000	0.000
		↓ AI-F	0.000	0.000	0.000	0.000	0.000	**0.000**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ML-20M	Rel	↑ HR	0.488	0.473	0.420	0.483	**0.505**	0.444	0.392	0.483	0.489	0.432	0.391	0.465	**0.505**	0.453	0.388	0.493
		↑ MRR	0.280	0.237	0.213	0.278	0.293	0.205	0.190	0.290	0.259	0.193	0.180	0.256	**0.293**	0.206	0.193	0.292
		↑ P	0.142	0.131	0.106	0.139	0.145	0.116	0.094	0.136	0.141	0.112	0.091	0.128	**0.150**	0.121	0.094	0.141
		↑ MAP	0.092	0.077	0.061	0.090	0.096	0.063	0.052	0.092	0.089	0.060	0.049	0.082	**0.100**	0.068	0.053	0.095
		↑ R	0.019	0.017	0.014	0.019	0.019	0.014	0.012	0.018	0.019	0.014	0.011	0.018	**0.020**	0.016	0.011	0.020
		↑ NDCG	0.154	0.135	0.112	0.151	0.158	0.116	0.098	0.152	0.148	0.111	0.093	0.139	**0.163**	0.121	0.099	0.157
	Fair	↑ Jain	0.008	0.017	0.018	0.009	0.028	0.068	**0.081**	0.033	0.029	0.070	0.074	0.034	0.018	0.044	0.049	0.021
		↑ QF	0.035	0.047	0.051	0.037	0.114	0.154	**0.165**	0.125	0.117	0.146	0.154	0.126	0.074	0.103	0.112	0.082
		↑ Ent	0.399	0.483	0.491	0.411	0.581	0.670	**0.690**	0.606	0.591	0.669	0.680	0.615	0.517	0.608	0.624	0.541
		↑ FSat	0.035	0.047	0.051	0.037	0.114	0.154	**0.165**	0.125	0.117	0.146	0.154	0.126	0.074	0.103	0.112	0.082
		↓ Gini	0.991	0.982	0.981	0.990	0.960	0.926	**0.914**	0.951	0.957	0.927	0.920	0.948	0.976	0.953	0.947	0.971
	Fair+Rel	↑ IBO	0.021	0.031	0.033	0.022	0.049	0.064	**0.067**	0.054	0.052	0.064	0.065	0.056	0.039	0.051	0.054	0.042
		↓ MME	0.001	0.001	0.001	0.001	0.001	0.000	**0.000**	0.001	0.001	0.000	0.000	0.001	0.001	0.001	0.001	0.001
		↓ IAA	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	**0.007**	0.007	0.007	0.007
		↓ II-F	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.000**	0.000	0.000	0.000
		↓ AI-F	0.000	0.000	0.000	0.000	0.000	0.000	**0.000**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

*QF = 1 means that all items in the dataset appear in the recommendation across all users.

†The scores of QF are the same as FSat for ML-*, as QF is computed based on the percentage of items in the dataset that are recommended, which in this dataset is equivalent to FSat: the percentage of items in the dataset that are recommended at least $\lfloor \frac{km}{n} \rfloor = 1$ time.

Table 5.8: The gradient values of the PF, based on the extreme points (starting and ending points). We consider a gradient to be 'good' if it is not zero or undefined (-).

	Lastfm	Amazon-lb	QK-video	Jester	ML-10M	ML-20M	# good	conclusion
HR-Ent	-97.57	-1.86	-0.31	-	-14.74	-6.95	5	inconsistent
HR-FSat	-1439.17	-19.92	0.00	-	-30.48	-18.97	4	inconsistent
HR-Gini	561.63	6.23	3.71	-	117.19	43.44	5	inconsistent
HR-Jain	-979.86	-18.77	-5.80	-	-157.73	-78.22	5	inconsistent
HR-QF	0.00	0.00	0.00	-	-30.48	-18.97	2	inconsistent
MAP-Ent	-0.17	-0.17	-0.03	-0.07	-0.14	-0.18	6	always good
MAP-FSat	-2.46	-1.81	0.00	-44.47	-0.29	-0.48	5	inconsistent
MAP-Gini	0.96	0.56	0.34	1.42	1.12	1.10	6	always good
MAP-Jain	-1.68	-1.70	-0.54	-0.37	-1.51	-1.98	6	always good
MAP-QF	0.00	0.00	0.00	0.0	-0.29	-0.48	2	inconsistent
MRR-Ent	-97.57	-1.86	-0.31	-	-14.74	-6.95	5	inconsistent
MRR-FSat	-1439.17	-19.92	0.00	-	-30.48	-18.97	4	inconsistent
MRR-Gini	561.63	6.23	3.71	-	117.19	43.44	5	inconsistent
MRR-Jain	-979.86	-18.77	-5.80	-	-157.73	-78.22	5	inconsistent
MRR-QF	0.00	0.00	0.00	-	-30.48	-18.97	2	inconsistent
NDCG-Ent	-0.24	-0.22	-0.04	-0.1	-0.20	-0.25	6	always good
NDCG-FSat	-3.50	-2.32	0.00	-68.56	-0.42	-0.68	5	inconsistent
NDCG-Gini	1.37	0.73	0.47	2.2	1.62	1.55	6	always good
NDCG-Jain	-2.38	-2.19	-0.73	-0.57	-2.18	-2.79	6	always good
NDCG-QF	0.00	0.00	0.00	0.0	-0.42	-0.68	2	inconsistent
P-Ent	-0.20	-0.33	-0.07	-0.08	-0.16	-0.20	6	always good
P-FSat	-2.95	-3.53	0.00	-51.41	-0.33	-0.55	5	inconsistent
P-Gini	1.15	1.10	0.89	1.65	1.26	1.26	6	always good
P-Jain	-2.01	-3.33	-1.40	-0.43	-1.70	-2.27	6	always good
P-QF	0.00	0.00	0.00	0.0	-0.33	-0.55	2	inconsistent
R-Ent	-0.17	-0.17	-0.03	-0.07	-0.26	-0.30	6	always good
R-FSat	-2.57	-1.83	0.00	-48.04	-0.53	-0.82	5	inconsistent
R-Gini	1.00	0.57	0.34	1.54	2.04	1.88	6	always good
R-Jain	-1.75	-1.73	-0.54	-0.4	-2.75	-3.39	6	always good
R-QF	0.00	0.00	0.00	0.0	-0.53	-0.82	2	inconsistent

Table 5.9: ↓DPFR scores at $k = 10$ of the recommender models for Lastfm, Amazon-lb, and QK-video, without and with re-ranking the the top $k' = 25$ items using Borda Count (BC), COMBMNZ (CM), and Greedy Substitution (GS). The best score per measure pair is in bold.

	model	ItemKNN				BPR				MultiVAE				NCL			
	reranking	-	BC	CM	GS	-	BC	CM	GS	-	BC	CM	GS	-	BC	CM	GS
Lastfm	P-Jain	0.853	0.819	0.861	0.853	0.837	0.784	0.824	0.834	0.805	**0.728**	0.776	0.799	0.813	0.735	0.773	0.809
	P-Ent	0.585	0.524	0.571	0.567	0.567	0.525	0.566	0.551	0.510	0.501	0.550	0.505	0.524	**0.497**	0.544	0.513
	P-Gini	0.869	0.802	0.821	0.850	0.871	0.820	0.841	0.856	0.811	**0.737**	0.759	0.789	0.835	0.756	0.775	0.813
	MAP-Jain	1.044	1.042	1.072	1.042	1.029	1.015	1.040	1.026	1.005	**0.974**	1.004	0.998	1.008	0.979	1.003	1.003
	MAP-Ent	0.811	0.802	0.830	0.797	0.797	0.803	0.824	0.784	0.759	0.791	0.815	**0.753**	0.764	0.787	0.813	0.755
	MAP-Gini	1.038	1.009	1.022	1.021	1.039	1.025	1.036	1.025	0.991	**0.961**	0.972	0.970	1.007	0.975	0.986	0.987
	R-Jain	0.968	0.947	1.004	0.969	0.952	0.917	0.970	0.954	0.923	**0.875**	0.936	0.923	0.927	**0.875**	0.930	0.928
	R-Ent	0.720	0.683	0.747	0.708	0.703	0.686	0.741	0.696	**0.656**	0.675	0.738	0.660	0.664	0.665	0.729	0.661
	R-Gini	0.968	0.918	0.955	0.953	0.969	0.935	0.971	0.959	0.914	**0.868**	0.907	0.900	0.932	0.878	0.918	0.917
	NDCG-Jain	0.989	0.995	1.046	0.989	0.973	0.968	1.013	0.972	0.948	0.933	0.985	0.945	0.949	**0.932**	0.978	0.947
	NDCG-Ent	0.760	0.758	0.813	0.747	0.743	0.762	0.805	0.733	0.703	0.756	0.807	0.702	0.706	0.746	0.797	**0.700**
	NDCG-Gini	1.000	0.976	1.009	0.984	1.000	0.994	1.023	0.987	0.949	0.934	0.966	**0.932**	0.964	0.943	0.974	0.947
Amazon-lb	P-Jain	0.529	**0.352**	0.415	0.489	0.571	0.415	0.464	0.542	0.733	0.656	0.679	0.726	0.742	0.678	0.694	0.736
	P-Ent	0.340	**0.307**	0.324	0.327	0.376	0.325	0.339	0.358	0.636	0.493	0.517	0.590	0.678	0.515	0.537	0.629
	P-Gini	0.633	**0.486**	0.575	0.587	0.708	0.578	0.630	0.669	0.896	0.834	0.850	0.878	0.906	0.849	0.860	0.890
	MAP-Jain	0.998	**0.904**	0.936	0.974	1.025	0.936	0.962	1.007	1.129	1.078	1.093	1.123	1.134	1.093	1.105	1.130
	MAP-Ent	0.830	**0.818**	0.825	0.825	0.848	0.825	0.830	0.840	0.990	0.906	0.919	0.961	1.017	0.919	0.933	0.985
	MAP-Gini	0.993	**0.906**	0.958	0.963	1.045	0.959	0.991	1.018	1.180	1.135	1.147	1.166	1.187	1.146	1.156	1.175
	R-Jain	0.984	**0.893**	0.927	0.960	1.015	0.924	0.948	0.997	1.119	1.069	1.082	1.112	1.123	1.082	1.096	1.119
	R-Ent	0.817	**0.809**	0.818	0.811	0.839	0.815	0.818	0.831	0.981	0.898	0.909	0.950	1.007	0.909	0.925	0.975
	R-Gini	0.981	**0.897**	0.951	0.951	1.037	0.949	0.980	1.010	1.172	1.127	1.138	1.157	1.178	1.137	1.149	1.166
	NDCG-Jain	1.022	**0.937**	0.966	1.000	1.051	0.967	0.990	1.034	1.148	1.102	1.116	1.142	1.153	1.116	1.127	1.149
	NDCG-Ent	0.868	**0.859**	0.866	0.863	0.889	0.867	0.871	0.881	1.023	0.945	0.957	0.994	1.049	0.956	0.970	1.018
	NDCG-Gini	1.028	**0.947**	0.997	1.000	1.083	0.999	1.029	1.056	1.211	1.170	1.181	1.197	1.218	1.180	1.190	1.206
QK-video	P-Jain	0.563	**0.297**	0.468	0.487	0.941	0.700	0.658	0.922	1.007	0.983	0.989	1.005	0.999	0.945	0.950	0.996
	P-Ent	0.245	**0.236**	0.243	0.241	0.335	0.258	0.254	0.310	0.623	0.499	0.510	0.588	0.542	0.403	0.400	0.504
	P-Gini	0.521	**0.327**	0.495	0.456	0.833	0.649	0.611	0.789	1.002	0.978	0.981	0.996	0.985	0.931	0.925	0.972
	MAP-Jain	1.106	**0.996**	1.061	1.070	1.332	1.180	1.157	1.318	1.379	1.365	1.371	1.378	1.371	1.338	1.343	1.369
	MAP-Ent	0.979	**0.977**	0.979	0.978	0.995	0.980	0.980	0.987	1.125	1.066	1.073	1.106	1.079	1.024	1.025	1.061
	MAP-Gini	1.082	**1.003**	1.070	1.052	1.254	1.147	1.127	1.225	1.372	1.358	1.362	1.367	1.357	1.324	1.322	1.347
	R-Jain	1.097	**0.982**	1.047	1.060	1.312	1.162	1.145	1.301	1.355	1.345	1.357	1.356	1.343	1.314	1.326	1.342
	R-Ent	0.969	0.962	0.964	0.968	0.969	**0.959**	0.966	0.964	1.096	1.040	1.056	1.079	1.043	0.993	1.002	1.027
	R-Gini	1.073	**0.989**	1.056	1.043	1.233	1.129	1.114	1.207	1.348	1.338	1.348	1.346	1.329	1.300	1.304	1.321
	NDCG-Jain	1.106	**0.996**	1.060	1.070	1.326	1.178	1.156	1.314	1.373	1.361	1.369	1.372	1.362	1.332	1.340	1.361
	NDCG-Ent	0.980	**0.976**	0.978	0.979	0.988	0.977	0.980	0.982	1.118	1.061	1.071	1.099	1.069	1.017	1.021	1.051
	NDCG-Gini	1.083	**1.002**	1.069	1.053	1.248	1.145	1.127	1.221	1.366	1.354	1.360	1.362	1.349	1.319	1.318	1.340

Table 5.10: ↓DPFR scores at $k = 10$ of the recommender models for Jester and ML-*, without and with re-ranking the the top $k' = 25$ items using Borda Count (BC), COMBMNZ (CM), and Greedy Substitution (GS). The best score per measure pair is in bold.

	model	ItemKNN				BPR				MultiVAE				NCL			
	reranking	-	BC	CM	GS	-	BC	CM	GS	-	BC	CM	GS	-	BC	CM	GS
Jester	P-Jain	0.709	0.566	0.719	0.707	0.679	**0.550**	0.631	0.677	0.747	0.569	0.638	0.745	0.698	0.604	0.625	0.696
	P-Ent	0.401	0.382	0.507	0.400	0.367	0.381	0.462	**0.366**	0.433	0.372	0.440	0.430	0.381	0.381	0.442	0.380
	P-Gini	0.724	0.602	0.739	0.722	0.675	**0.578**	0.651	0.673	0.766	0.598	0.651	0.763	0.704	0.619	0.642	0.701
	MAP-Jain	0.915	0.908	1.048	0.914	0.894	0.897	0.987	**0.892**	0.932	0.905	0.976	0.930	0.898	0.923	0.974	0.897
	MAP-Ent	0.704	0.805	0.914	0.703	0.687	0.804	0.889	0.686	0.704	0.795	0.859	0.703	0.681	0.794	0.868	**0.680**
	MAP-Gini	0.927	0.930	1.062	0.926	0.891	0.915	1.001	**0.889**	0.947	0.924	0.985	0.945	0.903	0.933	0.986	0.901
	R-Jain	0.774	0.704	0.927	0.773	0.748	**0.700**	0.821	0.746	0.802	0.702	0.787	0.800	0.759	0.732	0.787	0.758
	R-Ent	0.507	0.566	0.773	0.506	0.483	0.576	0.699	**0.482**	0.521	0.555	0.636	0.519	0.484	0.563	0.651	0.483
	R-Gini	0.788	0.733	0.943	0.787	0.745	**0.723**	0.837	0.743	0.819	0.727	0.798	0.817	0.765	0.745	0.801	0.763
	NDCG-Jain	0.823	0.793	0.977	0.822	0.798	**0.782**	0.900	0.797	0.845	0.788	0.878	0.844	0.807	0.810	0.878	0.805
	NDCG-Ent	0.579	0.673	0.833	0.578	0.558	0.673	0.791	0.557	0.586	0.660	0.745	0.584	0.556	0.661	0.758	**0.555**
	NDCG-Gini	0.837	0.818	0.992	0.835	0.795	0.802	0.914	**0.793**	0.862	0.810	0.887	0.860	0.812	0.821	0.890	0.810
ML-10M	P-Jain	1.072	1.067	1.082	1.075	1.047	**1.023**	1.031	1.051	1.098	1.094	1.104	1.099	1.052	1.044	1.059	1.053
	P-Ent	0.884	0.831	0.844	0.880	0.765	**0.749**	0.766	0.764	1.023	0.974	0.970	1.010	0.801	0.771	0.791	0.791
	P-Gini	1.080	1.075	1.088	1.081	1.042	**1.025**	1.032	1.041	1.106	1.102	1.111	1.107	1.055	1.050	1.065	1.053
	MAP-Jain	1.172	1.172	1.182	1.173	1.149	**1.134**	1.135	1.150	1.193	1.192	1.201	1.194	1.155	1.153	1.161	1.154
	MAP-Ent	0.995	0.955	0.962	0.991	0.893	**0.886**	0.893	0.888	1.119	1.076	1.073	1.106	0.924	0.906	0.915	0.914
	MAP-Gini	1.173	1.173	1.182	1.174	1.139	**1.130**	**1.130**	1.135	1.195	1.193	1.202	1.195	1.152	1.153	1.160	1.149
	R-Jain	0.860	0.846	0.847	0.859	0.835	0.780	**0.768**	0.829	0.868	0.866	0.867	0.867	0.847	0.813	0.810	0.844
	R-Ent	0.653	0.570	0.563	0.643	0.492	0.423	**0.415**	0.472	0.807	0.748	0.730	0.788	0.554	0.473	0.468	0.534
	R-Gini	0.889	0.877	0.876	0.888	0.850	0.805	**0.793**	0.838	0.900	0.897	0.898	0.899	0.872	0.843	0.840	0.866
	NDCG-Jain	1.140	1.142	1.156	1.142	1.115	**1.107**	1.114	1.118	1.168	1.166	1.180	1.169	1.119	1.125	1.138	1.119
	NDCG-Ent	0.972	0.932	0.944	0.968	0.863	0.866	0.881	**0.860**	1.103	1.060	1.062	1.091	0.894	0.884	0.901	0.885
	NDCG-Gini	1.151	1.153	1.166	1.153	1.114	**1.112**	1.119	**1.112**	1.180	1.177	1.190	1.180	1.127	1.134	1.147	1.124
ML-20M	P-Jain	1.014	1.015	1.033	1.016	0.998	**0.992**	1.000	1.001	1.000	0.993	1.006	1.006	1.002	1.004	1.020	1.006
	P-Ent	0.885	0.842	0.858	0.880	0.776	**0.761**	0.773	0.771	0.775	0.765	0.780	0.774	0.807	0.783	0.800	0.801
	P-Gini	1.056	1.057	1.073	1.057	1.031	**1.026**	1.033	1.030	1.031	1.029	1.039	1.033	1.040	1.042	1.057	1.042
	MAP-Jain	1.118	1.124	1.137	1.119	**1.103**	**1.103**	1.104	**1.103**	1.107	1.105	1.111	1.109	1.106	1.114	1.123	1.108
	MAP-Ent	0.996	0.963	0.973	0.991	0.900	0.894	0.898	**0.892**	0.902	0.897	0.905	0.897	0.926	0.913	0.921	0.919
	MAP-Gini	1.151	1.156	1.167	1.151	1.127	1.129	1.129	**1.124**	1.130	1.131	1.136	1.129	1.135	1.143	1.151	1.135
	R-Jain	0.775	0.768	0.769	0.775	0.757	0.723	**0.712**	0.753	0.756	0.721	0.718	0.752	0.766	0.744	0.741	0.764
	R-Ent	0.645	0.573	0.568	0.634	0.490	0.426	**0.413**	0.471	0.483	0.426	0.421	0.465	0.542	0.471	0.462	0.523
	R-Gini	0.836	0.829	0.829	0.836	0.808	0.778	**0.768**	0.800	0.805	0.779	0.774	0.797	0.822	0.803	0.799	0.818
	NDCG-Jain	1.081	1.089	1.107	1.083	**1.065**	1.072	1.078	1.066	1.072	1.075	1.086	1.075	1.068	1.083	1.097	1.071
	NDCG-Ent	0.971	0.939	0.955	0.966	0.871	0.873	0.883	**0.865**	0.876	0.878	0.891	0.873	0.898	0.892	0.906	0.892
	NDCG-Gini	1.121	1.129	1.144	1.122	1.097	1.105	1.110	**1.095**	1.102	1.109	1.118	1.102	1.105	1.119	1.132	1.105

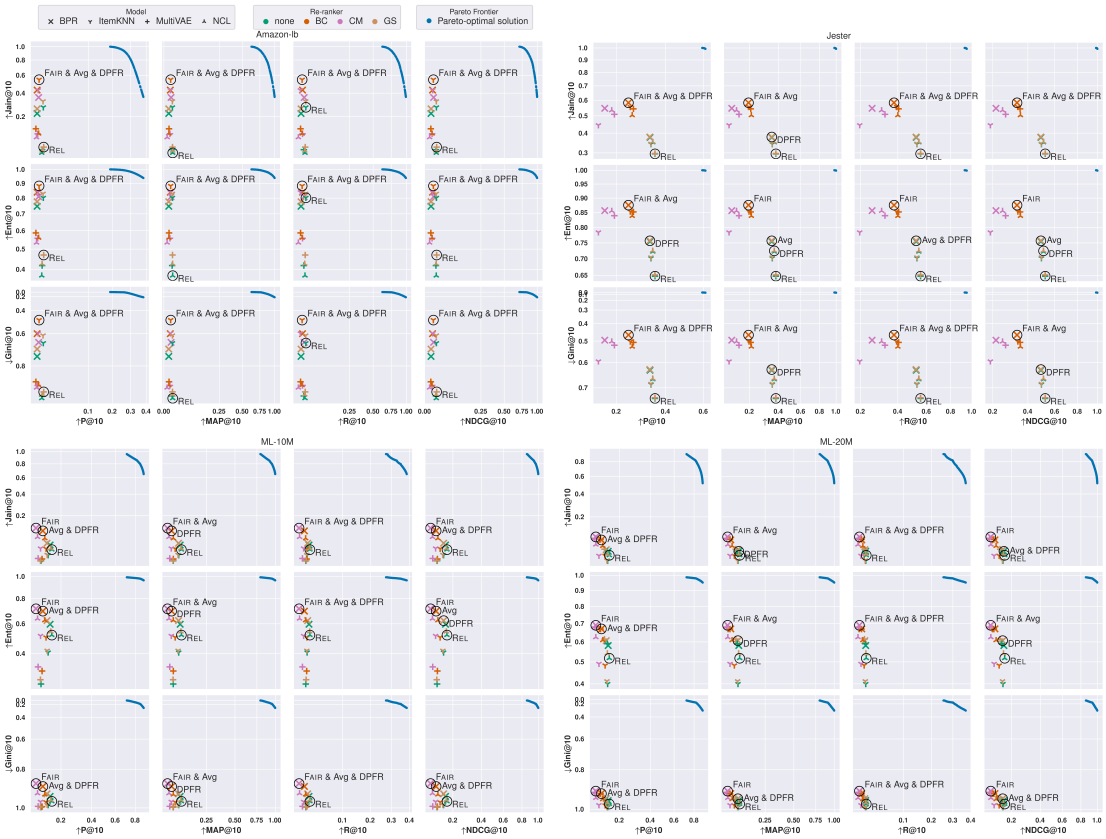


Figure 5.4: Pareto Frontier of fairness and relevance (in blue), together with recommender model scores for Amazon-lb, Jester, and ML-*. FAIR measures are on the y-axis and REL measures are on the x-axis. We implement exponential-like scales to enhance the visibility of the model plots. The REL, FAIR, Avg, and DPFR denote the best model based on each evaluation approach.

Table 5.11: Range of agreement τ between estimated PF and PF across 12 measure pairs, using the estimated PF with 3–12 points.

#pts	Lastfm	Amazon-lb	QK-video	Jester	ML-10M	ML-20M
3	0.78–1.00	0.98–1.00	1.00–1.00	1.00–1.00	0.97–1.00	0.75–1.00
4	0.88–1.00	0.98–1.00	0.98–1.00	0.98–1.00	0.98–1.00	0.93–1.00
5	0.78–1.00	0.98–1.00	1.00–1.00	1.00–1.00	0.97–1.00	0.92–1.00
6	0.90–1.00	0.97–1.00	1.00–1.00	0.98–1.00	0.95–1.00	0.92–1.00
7	0.88–1.00	1.00–1.00	1.00–1.00	1.00–1.00	0.98–1.00	0.93–1.00
8	0.90–1.00	0.98–1.00	1.00–1.00	0.98–1.00	1.00–1.00	0.95–1.00
9	0.98–1.00	1.00–1.00	1.00–1.00	1.00–1.00	0.97–1.00	0.98–1.00
10	0.88–1.00	1.00–1.00	1.00–1.00	0.98–1.00	1.00–1.00	0.95–1.00
11	0.92–1.00	1.00–1.00	1.00–1.00	1.00–1.00	0.98–1.00	0.97–1.00
12	0.95–1.00	1.00–1.00	1.00–1.00	0.98–1.00	0.98–1.00	0.97–1.00

Chapter 6

Measuring Individual User Fairness with User Similarity and Effectiveness Disparity

Abstract

Individual user fairness is commonly understood as treating similar users similarly. In Recommender Systems (RSs), several evaluation measures exist for quantifying individual user fairness. These measures evaluate fairness via either: (i) the disparity in user effectiveness scores regardless of user similarity, or (ii) the disparity in items recommended to similar users regardless of item relevance. Both disparity in recommendation effectiveness and user similarity are very important in fairness, yet no existing individual user fairness measure simultaneously accounts for both of them. In brief, current user fairness evaluation measures implement a largely incomplete definition of fairness. To fill this gap, we present Pairwise User unFairness (PUF), a novel evaluation measure of individual user fairness that considers both effectiveness disparity and user similarity. Our measure works as follows. If two users get recommendations of similar NDCG, for instance, PUF will give a different fairness score depending on how similar the users are to each other, in terms of their past interactions: the case where similar users get recommendations of similar NDCG is fairer than the case where similar users get recommendations of different NDCG. PUF is the only measure that can express this important distinction. We empirically validate that PUF does this consistently across 4 datasets and 7 rankers, and robustly when varying user similarity or effectiveness. In contrast, all other measures are either almost insensitive to effectiveness disparity or completely insensitive

to user similarity. We contribute the first RS evaluation measure to reliably capture both user similarity and effectiveness in individual user fairness. Our code will be made public upon acceptance.

6.1 Introduction

The evaluation of Recommender Systems (RSs) has always relied heavily on effectiveness as it directly affects user utility and satisfaction. Unsurprisingly, existing work on RS fairness evaluation often uses measures that depend on effectiveness scores, especially when it comes to individual user fairness. *Individual fairness* is traditionally defined as treating similar individuals in a similar manner [63]. We focus on attribute-free fairness for individual users, where we assume no user attribute is available other than the user identifiers and interactions [128, 253]. Sensitive attributes (e.g., race, age) are frequently unavailable due to privacy or data incompleteness.

To measure individual user fairness in RSs, the disparity in recommendation effectiveness across users is often used as a proxy of how similar the recommendation algorithm treats the users. Recommendations across users are deemed fair if similar users have similar effectiveness scores. Yet, no existing individual user fairness measure considers both effectiveness and user similarity (Tab. 6.1).

User similarity and recommendation effectiveness are both important, as the former affects our expectation of how close effectiveness scores should be. For example, given two users whose past interactions are highly similar, the recommendations they receive are deemed fair if their effectiveness is similar, and unfair if the effectiveness differs a lot. This is because fairness means similar users should get similar treatment. However, two dissimilar users cannot expect to get similar treatment, as their past interactions may differ, e.g., in terms of amount, frequency, or item type.

Current individual user fairness measures can be grouped into those that consider: (i) only effectiveness disparity; and (ii) only user similarity and recommendation similarity. For (i), different effectiveness for users is not always unfair. Some users may have only a few past interactions, and others may have very specific tastes for whom only a few items are relevant. Fairness should consider that these users are different and cannot be compared to users who, e.g., consume only popular items. For (ii), recommending similar items to similar users may not be fair if one user is satisfied with their recommendation, but another is not. In this case, the recommendation is not really fair as its effectiveness differs.

To counter the above limitations in current measures, we propose a novel evaluation measure for individual user fairness in RS: Pairwise User unFairness (PUF).

Table 6.1: Overview of individual user fairness measures in RS.

Measure	Reference	Effectiveness	User Similarity
Standard deviation	[20, 128, 167, 183, 236]	✓	-
Gini index	[77, 126]	✓	-
Envy-based	[20, 61, 167]	✓	-
UF	[235]	-	✓
PUF	ours	✓	✓

PUF quantifies individual user fairness based on the disparity in recommendation relevance between user pairs, weighted by the similarity of user pairs. As such, PUF is aligned with the definition of individual fairness and also accounts for recommendation effectiveness, and thus, user utility. Our experiments show that PUF is superior to the current measures in terms of sensitivity to changes in effectiveness scores and user similarity distribution. Overall, we contribute a new evaluation measure for individual user fairness, which considers both user similarity and disparity in recommendation effectiveness, and which does not have the same limitations as existing measures.

6.2 Individual User Fairness

We present the definition of individual user fairness (§6.2.1) and existing evaluation measures of individual user fairness (§6.2.2).

6.2.1 Definition of Individual User Fairness

We define individual user fairness in RSs as per [63]: let u and u' be the representations of two users; L_u and $L_{u'}$ be the recommendation lists of these two users; and $M(\cdot)$ be a function that maps a recommendation list to a score, e.g., its effectiveness score. Any two users whose profile distance is $d(u, u')$ should receive recommendations such that recommendation effectiveness satisfies $D(M(L_u), M(L_{u'})) \leq d(u, u')$, where D is a distance measure. In other words, fairness is achieved when the difference in the users' recommendation effectiveness is at most $d(u, u')$. This definition agrees with the definitions of RS individual user fairness in [209, 252].

6.2.2 Current Individual User Fairness Measures

We present existing evaluation measures of attribute-free individual user fairness in RSs (Tab. 6.1). To our knowledge, these are all such measures for RSs published up to January 2025 [1, 9, 133, 173, 209, 228, 237, 252, 261]. All these measures quantify unfairness, so that the lower their scores, the fairer (denoted by ↓). Some measures are based on general-purpose inequality measures, i.e., standard deviation (§6.2.2.1) and Gini Index (§6.2.2.2); others are based on unique fairness concepts, i.e., envy-freeness (§6.2.2.3) and distance-based (§6.2.2.4). All existing measures except the distance-based measure quantify effectiveness disparity but ignore user similarity. While the existing distance-based measure considers user similarity, it is detached from effectiveness. Hence, no existing individual user fairness measure in RS considers both user similarity and effectiveness disparity.

6.2.2.1 Standard deviation (SD)

↓SD measures how dispersed the data is in relation to the mean. In RSs, SD and variance are often used to quantify individual user fairness. RS is fair if it provides equal prediction accuracy to all users, and fairness is evaluated via the variance of the mean squared error (MSE) of user rating prediction [128, 183]. Other works measure fairness from user recommendation lists, e.g., via the variance of NDCG scores across users [236] or the SD of user utility [20, 167]. The utility of user u, when given the top k recommendation list of user u', $L_{u'}$ is originally defined as follows:

$$\phi_u^{ori}(L_{u'}) = \frac{\sum_{i \in L_{u'}} \hat{r}_{u,i}}{\sum_{i \in R_{u'}^k} \hat{r}_{u,i}} \tag{6.1}$$

where $\hat{r}_{u,i}$ is the predicted relevance score of user u for item i, and R_u^k is the top-k most relevant items of u based on the ground truth. Two issues arise from the above formulation: (i) computing user utility based on predicted relevance $\hat{r}_{u,i}$ may not reflect the true recommendation effectiveness; and (ii) relevance ties for items in R_u^k are not handled, leading to multiple possible compositions of R_u^k. Further, in [20, 167], the predicted relevance score is also taken from a latent factorisation model, which means that recommending items based on decreasing predicted relevance (as commonly done), will always result in $\phi_u^{ori}(L_u) = 1$ for all users, rendering the user utility measure pointless. We resolve the above issues in the utility function, by replacing predicted relevance with ground truth relevance $r_{u,i}$, (e.g., ratings in the

test set) as follows:

$$\phi_u(L_{u'}) = \frac{\sum_{i \in L_{u'}} r_{u,i}}{\sum_{i \in R^k_{u'}} r_{u,i}} \tag{6.2}$$

For binary relevance, $r_{u,i} \in \{0, 1\}$, assuming that each user has at least k relevant items, $\phi_u(L_u) = \frac{1}{k} \sum_{i \in L_u} 1_{\{i \in R_u\}}$ or simply Precision@k (P@k). The indicator function $1_{\{\cdot\}}$ returns 1 if the expression $\cdot$ is true, and 0 otherwise. We denote the set of user u's relevant items as R_u. As such, the SD based on the reformulated utility function is SD-P $= SD_{u \in U} \phi_u(L_u)$, where U is the set of users.

6.2.2.2 Gini Index (Gini)

↓Gini is a well-known inequality measure that quantifies the extent to which a distribution deviates from a perfectly equal distribution. It has been used to measure individual user fairness from different distributions, e.g., P@k (Gini-P), NDCG (Gini-NDCG), F1, or the utility (Eq. (6.1)) score per user [77, 126].

6.2.2.3 Envy-based measures

Fairness has been measured as envy-freeness. Envy is defined as the extra utility a user u would have received if they were given the recommendation list of user u', $L_{u'}$ [20, 167]. Formally, the envy of user u towards u' is computed as:[1]

$$envy(u, u') = \max\{\phi_u(L_{u'}) - \phi_u(L_u), 0\} \tag{6.3}$$

Three fairness evaluation measures are based on this concept, where envy is aggregated differently: Mean Envy (↓ME) [20, 167], Mean Max Envy (↓MME) [61], and Proportion of ϵ-Envious User (↓PEU) [61]:

$$ME = \frac{2}{m(m-1)} \sum_{u \in U} \sum_{\substack{u' \in U \\ u' \neq u}} envy(u, u') \tag{6.4}$$

$$MME = \frac{1}{m} \sum_{u \in U} \max_{u' \in U} \{envy(u, u')\} \tag{6.5}$$

$$PEU = \frac{1}{m} \sum_{u \in U} 1_{\{\max_{u' \in U} envy(u,u') > \epsilon\}} \tag{6.6}$$

[1]The original user utility function (Eq. (6.1)) is used to compute envy in [20, 167]. WLOG, user utility is computed slightly differently in [61] as the work is on online learning.

where $m = |U|$ is the number of users, and ϵ is the envy tolerance hyperparameter. A user u is said to be ϵ-envious of user u' if their gain in user utility (envy) exceeds ϵ.

6.2.2.4 Distance-based measures

In [235], fairness for individual users is defined as any two similar users u, u' receiving similar recommendations. Recommendation disparity is then measured with UnFairness score (↓UF). UF uses both the user similarity, $sim_{UF}(u, u')$, and the pairwise distance between the representation (e.g., embeddings) of items in the recommendation list of user u and of user u', $d_L(u, u')$. User similarity is modelled by both Jaccard (sim_{Jacc}) of the user's set of past interactions H_u and JS-div between item feature (e.g., genre) distributions in the interactions of user u and u', $D_{JS}(u_\alpha, u'_\alpha)$, where α is the item feature. UF is computed as:

$$UF = \log \sum_{(u,u') \in U^*} sim_{UF}(u, u') \times d_L(u, u') \tag{6.7}$$

$$sim_{UF}(u, u') = \gamma \; sim_{Jacc}(u, u') + (1 - \gamma) \; sim_{JS}(u, u') \tag{6.8}$$

$$d_L(u, u') = \frac{1}{k^2} \sum_{i \in L_u} \sum_{i' \in L_{u'}} d_{cos}(\vec{i}, \vec{i'}) \tag{6.9}$$

$$sim_{Jacc}(u, u') = \frac{|H_u \cap H_{u'}|}{|H_u \cup H_{u'}|} \tag{6.10}$$

$$sim_{JS}(u, u') = 1 - D_{JS}(u_\alpha, u'_\alpha) \tag{6.11}$$

where U^* is the set of similar user pairs (u, u') with $sim_{UF}(u, u') \geq t$, and t is the similarity threshold. In [235], $t = mean(sim_{UF}) + SD(sim_{UF})$, where the mean and SD are computed for all user pairs. We denote as γ the weight for sim_{Jacc}, as k the recommendation cut-off, and as $d_{cos}(\vec{i}, \vec{i'})$ the cosine distance between the representations of items i and i'. The log base in Eq. (6.7) [235] is unclear. Note that UF does not consider recommendation effectiveness.

6.3 Pairwise User unFairness (PUF)

We present our evaluation measure of individual user fairness, Pairwise User unFairness (↓PUF). PUF consists of two components: the similarity between users and the disparity in recommendation effectiveness. Next, we describe: (1) how user similarity

is computed; and (2) how to quantify the disparity of recommendation effectiveness, while taking into account user similarities.

(1) Measuring similarity. Measuring similarity between users is an inherently hard problem and there is no single ground truth of what makes two users similar [30]. When there is no user attribute, as per our focus on attribute-free fairness (§6.1), user profiles can be modelled based on their historical interactions [97, 128, 235]. Similarity between users is then computed pairwise based on the user representation (e.g., click/rating matrix, user embedding), for example, with cosine similarity [184] or Jaccard (Eq. (6.10)) [235].

(2) Measuring disparity. PUF quantifies individual user fairness based on disparity in recommendation effectiveness, considering user similarity. PUF measures the mean pairwise difference in the effectiveness score per user, weighted by the user pair similarity. Based on (1) and (2), we define PUF as follows:

$$\text{PUF} = \frac{2}{m(m-1)} \sum_{u \in U} \sum_{\substack{u' \in U \\ u \neq u'}} sim(u, u') \times |S(u) - S(u')| \tag{6.12}$$

where U is the set of all users ($|U| = m$, where $m \geq 2$), $sim(u, u') \in [0, 1]$ is the similarity of users u and u', S is an effectiveness measure, e.g., P@k. S must range or scaled to be in $[0, 1]$, so PUF also ranges in $[0, 1]$. PUF is modular; it can be operated with any similarity or effectiveness measure as long as it fulfils the range requirement.

How PUF differs from UF. While both PUF and UF (Eq. 6.7) [235] consider user similarity, PUF differs from UF as PUF considers the difference in the recommendation relevance, rather than only the disparity based on the representation of the recommended items as in UF. Moreover, recommending different sets of items to two similar users may be considered unfair by UF even if both users like their recommendations, but it is fair based on PUF. In theory, users with similar tastes in the past are likely to have a similar preference in the future [186]. Yet, the recommendation problem is challenging as even two highly similar users may not equally like their recommended items, if they are given identical items. This may be due to, for example, incomplete historical data that is unrepresentative of user taste, diverging user preference, or limited ground truth data. These limitations necessitate a look into the disparity of the recommendation relevance, which is what our measure quantifies.

Overall, our measure, PUF, aligns with the definition of individual user fairness and quantifies fairness through the disparity in recommendation effectiveness, which is more meaningful than the disparity in the representation of recommended items, as effectiveness relates more to user utility.

Table 6.2: Statistics of the preprocessed datasets.

dataset	#users (all/test)	#items	#interactions	sparsity (%)
Lastfm [32]	1,859/1,836	2,823	71,355	98.64%
QK-video [247]	4,656/3,514	6,423	51,777	99.83%
ML-10M [94]	49,378/1,523	9,821	5,362,685	98.89%
ML-20M [94]	89,917/2,178	16,404	10,588,141	99.28%

6.4 Empirical Analysis

We compare PUF to existing effectiveness and individual user fairness measures.

6.4.1 Experimental Setup

Datasets. We use four real-world datasets from three domains: music (Lastfm [32]), video (QK-video [247]), and movie (ML-10M and ML-20M [94]). We obtain Lastfm and ML-* from [260], and QK-video from [247]. We use the 'sharing' interactions in QK-video.

Preprocessing: For all datasets, we remove duplicate interactions and keep the most recent. We remove users and items with < 5 interactions. For ML-* we map ratings ≥ 3 to 1, and discard the rest. The threshold 3 is chosen as the ratings range between $[0.5, 5]$. Lastfm and QK-video have unary ratings, so no mapping is required.

Splitting: We split the preprocessed datasets into train/val/test with a ratio of 6:2:2. The ML-* datasets are temporally split, while Lastfm and QK-video are randomly split as they have no timestamps. We split datasets globally (not user-wise) to avoid data leakage [151]. After splitting, users with < 5 interactions in the train set are removed from all splits to ensure that each user has adequate training data. The final preprocessed dataset statistics are in Tab. 6.2.

Recommenders. We use 7 well-known collaborative filtering recommenders that represent different model types: user- and item-based K-nearest neighbour (U-KNN [186] and I-KNN [55]), Bayesian Personalised Ranking (BPR [185]), Variational Autoencoder with multinomial likelihood (MVAE [135]), Neural Graph Collaborative Filtering (NGCF [226]), Neural Matrix Factorisation (NMF [95]), and Neighbourhood-enriched Contrastive Learning (NCL [139]). All models except U- and I-KNN are trained for 300 epochs with early stopping. We tune hyperparameters with grid search. The configuration with the best NDCG@10 during validation

is the final model. Implementation, training, and tuning are done with RecBole [260].

Evaluation measures. We measure recommendation effectiveness (Eff) with Hit Rate (HR), MRR, Precision (P or Prec), Recall (R), MAP, and NDCG [102]. Recommendation fairness for individual users (Fair) is evaluated with our PUF measure (§6.3) and all existing measures (§6.2.2): standard deviation (SD) of the P@k (SD-P) and NDCG@k (SD-NDCG) scores across all users;[2] Gini Index of P@k (Gini-P) and NDCG@k (Gini-NDCG); envy-based measures (ME, MME, and PEU); and distance-based measure, UF.

We evaluate all runs at $k = 10$. For PEU, we set $\epsilon = 0.05$ [61]. For UF, the number of all user pairs is used as the log base, so that UF $\in [0, 1]$. We also set $\gamma = 1$ to remove the term for category-based similarity.[3] As we use several models with diverse item representations, for a fair comparison between all models, we use the one-hot encoding of the historical interactions for the item representation in Eq. (6.9) of UF. The similarity threshold t is fixed at $t = mean(sim_{UF}) + SD(sim_{UF})$ as per [235]. We test four variants of PUF (two user similarity measures $\times$ two effectiveness measures). For similarity, we use cosine similarity (sim_{cos}) and Jaccard (sim_{Jacc}, Eq. (6.10)), as both similarity measures are used in UF. In addition, sim_{cos} is used in our U-KNN model, as well as to analyse the similarity between users in [184]. All user similarities are computed from the observed interactions in the train set (i.e., the binarised interaction matrix). We compute PUF with P and NDCG, to represent set- and rank-based measures, as well as for consistency with other Fair measures which are also computed with P and NDCG. To avoid extremely small values, we min-max normalise the pairwise user similarity for each dataset.

6.4.2 Comparison of All Evaluation Measures

We present the Eff and Fair measures evaluation results (§6.7.1.1, Tab. 6.4), To identify possible empirical limitations of existing Fair measures and PUF, we analyse their score range and computational efficiency.

Score range. A measure that does not exhibit a wide range of scores across various fairness levels is less useful in distinguishing changes in fairness. Such measures may create an illusion of a negligible difference in fairness, due to their compressed empirical range [178]. Across datasets, the observed range (not the theoretical range) of each Fair measure varies, except for MME which is consistently extremely small ($\leq 5 \times 10^{-3}$). Changing a user's recommendation list with another user's as in

[2] We do not compute SD of MSE [128, 183] as the ratings are binarised.

[3] We vary the γ parameter in §6.4.8.

Table 6.3: Mean computation time (s) of measures per model.

		Lastfm	QK-video	ML-10M	ML-20M
FAIR (existing)	SD-P	<0.01	0.01	<0.01	<0.01
	SD-NDCG	<0.01	0.02	0.01	<0.01
	Gini-P	<0.01	0.01	<0.01	<0.01
	Gini-NDCG	<0.01	0.03	0.02	<0.01
	ME	233.88	1867.81	871.17	2020.60
	MME	233.91	1867.88	871.20	2020.70
	PEU	233.91	1867.88	871.20	2020.70
	UF	1297.78	873.52	1020.16	1358.52
FAIR (PUF)	PUF-Prec-Cos	5.30	20.68	4.00	8.25
	PUF-Prec-Jacc	9.79	35.63	7.22	15.65
	PUF-NDCG-Cos	4.32	16.01	3.38	6.72
	PUF-NDCG-Jacc	9.56	35.10	7.40	15.19

Eq. (6.3) does not generally result in a large increase in the user's P@k (envy), which translates to low MME. ME and PEU are unaffected by this issue even if they are based on Eq. (6.3), as ME accounts for envy across all users rather than the maximum envy per user, and PEU employs an envy threshold. In short, MME is the least sensitive as it fails to discriminate between models for the same and across datasets, while other FAIR measures, including PUF, are more sensitive. The most sensitive measures, i.e., the ones with the widest observed range, are Gini and PEU.

Measure computation time. We report the average computation time of the FAIR measures in Tab. 6.3. All runs are done with AMD EPYC 75F3 for a fair comparison. User pairwise similarity is computed once per dataset for each PUF variant. We find that half of the existing measures, i.e., ME, MME, PEU, and UF are computationally expensive (>20 minutes), while PUF is significantly faster (<40 s), despite also being a pairwise measure. ME/MME/PEU/UF need additional extensive computation per pair, which makes them expensive: UF does nested pairwise comparison in Eq. (6.7) and Eq. (6.9), while ME/MME/PEU recompute the effectiveness score for each user pair (Eq. (6.3)). In short, PUF is computationally more efficient and thus more practical to compute than most existing FAIR measures.

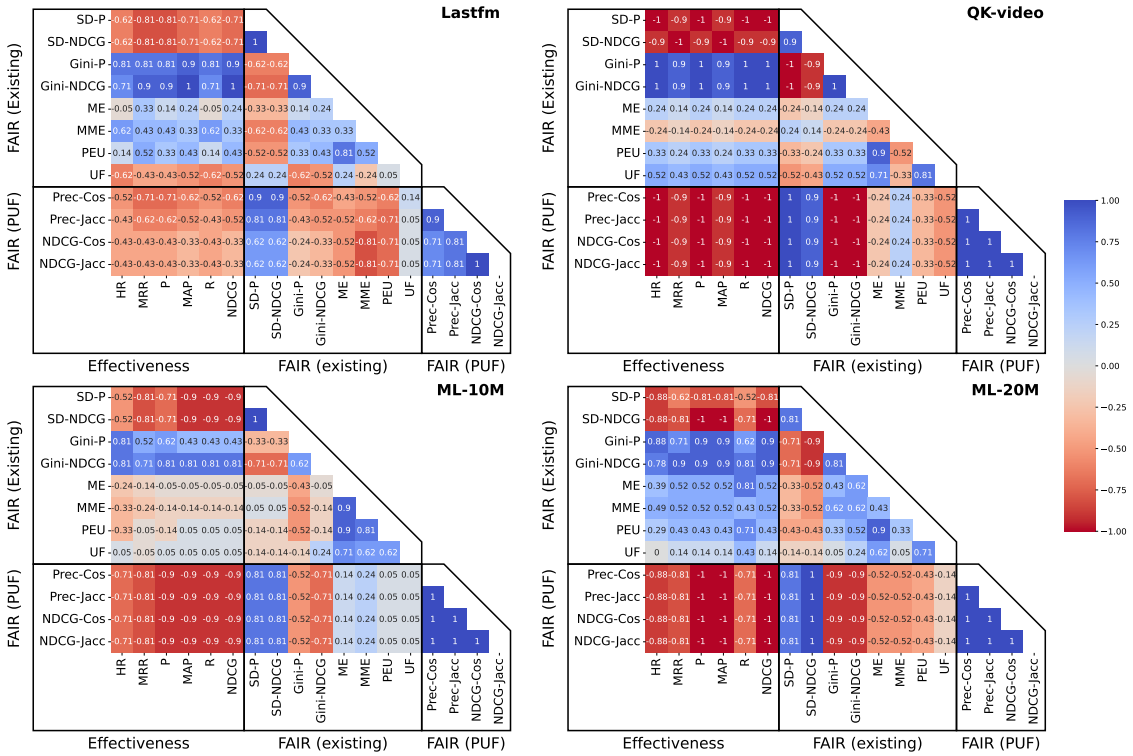


Figure 6.1: Kendall's τ correlation between EFF measures, existing FAIR measures, and PUFs for all recommenders.

6.4.3 Measure Agreement

An important aspect when comparing evaluation measures is how much they agree when their scores are used to rank models from best to worst. If one measure can be used to estimate the rank ordering given by another, there is no point in using both measures if we are only interested in ranking models. To study this, we compute Kendall's τ correlation for all EFF and FAIR measures. Kendall's τ can handle ties (unlike Spearman's ρ) and is more robust to nonlinear relationships (unlike Pearson's coefficient). If two measures have $\tau \geq 0.9$, we consider their rankings equivalent [223]. Fig. 6.1 shows the agreement between (i) EFF and FAIR measures, and (ii) among FAIR measures.

Agreement between Eff and Fair measures. Across datasets, the agreement between FAIR and EFF measures varies from strong disagreement (e.g., SD with $\tau \in [-1, -0.52]$) to moderate-to-strong agreement (e.g., Gini with $\tau \in [0.43, 1]$). Our PUF consistently disagrees with EFF measures, even if the disagreement is weaker

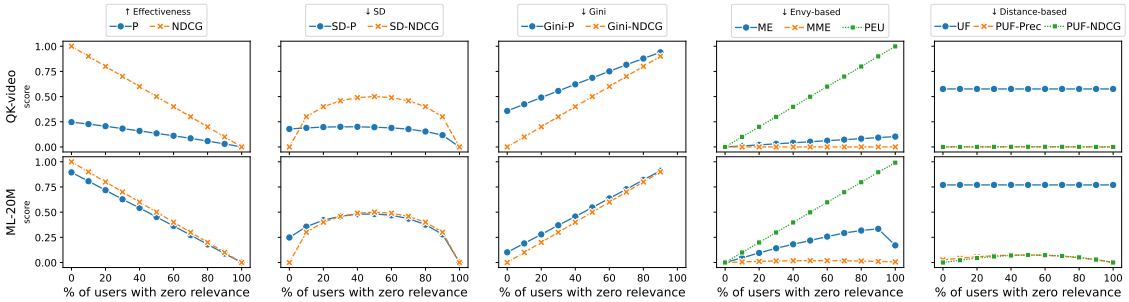


Figure 6.2: Effectiveness (Eff) and fairness (Fair) scores of QK-video and ML-20M, when artificially varying % of users with all irrelevant items (zero relevance), and the rest of the users receiving all relevant items. All PUF variants overlap. Gini is missing points at 100% users with zero relevance as it is undefined when each user has zero Eff scores.

for Lastfm, $\tau \in [-0.71, -0.33]$. As no Fair measure consistently has $|\tau| \geq 0.9$ with any Eff measure for all datasets, their model orderings cannot be precisely inferred from that of Eff measures.

Agreement among Fair measures. Regarding the agreement between PUF and existing Fair measures, we find that only SD aligns with PUF ($\tau \geq 0.62$), while the rest, including UF, tend to disagree or have a weak correlation with PUF, $\tau \in [-1, 0.24]$. Both UF and PUF consider user similarity, yet UF weakly correlates to PUF ($\tau \in [-0.14, 0.14]$) for all datasets except QK-video ($\tau \in [-0.52, -0.43]$). This weak relationship may be due to UF not considering item relevance, while PUF does. In contrast, despite not being similarity-based, SD correlates the strongest to PUF. Yet, most of its rankings are not equivalent to PUF ($\tau < 0.9$), which means that fairness evaluation with SD can lead to a conclusion that misaligns with PUF.

Even if we only compare the best model based on the measures instead of the model rankings, UF and PUF never agree (§6.7.1.1, Tab. 6.4). Thus, using UF and PUF to gauge fairness may differ in the conclusion on which model is the fairest.

Overall, we find that existing Fair measures frequently disagree on model orderings with PUF, regardless of whether the measure accounts for user similarity. While SD has the most similar conclusions to PUF, it still does not give equivalent rankings to PUF.

6.4.4 Varying Relevance Score Distribution

It is important that a fairness measure can capture changes across differences in the recommendation quality across users, as fairness is related to the disparity in recommendation (e.g., its relevance). Thus, any change in Eff scores should also be reflected in the Fair scores. To this end, we study how varying item relevance affects the Eff and Fair measures. All Fair measure equations except for UF explicitly account for effectiveness, but it is unknown how sensitive Fair measures are to changes in Eff scores.

Procedure. The change of relevance scores is artificially done as follows. For all users, we start by recommending k relevant items (based on the test set). For users with more than k relevant items, we select the relevant items randomly. For users with fewer than k relevant items, we fill the remaining slots with random irrelevant items to ensure that each user receives exactly k items. In each iteration, we replace the recommendation of 10% of the users with all irrelevant items and compute the measures. As such, we expect maximum fairness at the start (as all users have the maximum Eff scores[4]) and at the end (as all users have 0 Eff scores). We also expect maximum unfairness when half the users get all relevant items, and the rest get irrelevant items, as this leads to one of the most uneven Eff score distributions. For brevity, we only compute PUF with sim_{Jacc}, as the model orderings given by PUF-*-Cos are equivalent to PUF-*-Jacc (§6.4.3). Here, user similarities are computed based on the observed interactions in the train sets.

Results. Fig. 6.2 shows the results for QK-video and ML-20M, which represent the overall trends in all our datasets; we show the rest in §6.7.1.2. We find that PUF fits the expectation of having decreasing fairness followed by increasing fairness, denoted by the inverted parabolas. This trend is more pronounced for the ML-* datasets, as the mean user pairwise similarity is higher than for the other datasets. Among existing Fair measures, only SD follows this expectation, while the others show undesirable tendencies: as Eff decreases, Gini and PEU significantly become less fair, and ME does the same to a minor extent. The overall decrease in fairness in these measures is undesirable, as the measure scores closely resemble decreasing effectiveness instead of the disparity in the Eff scores distribution. Even worse, MME and UF are almost invariant to the change in recommendation effectiveness:

[4]Measures that are computed with P@k do not start exactly at the optimal value, as not all users have k items in their test set [153].

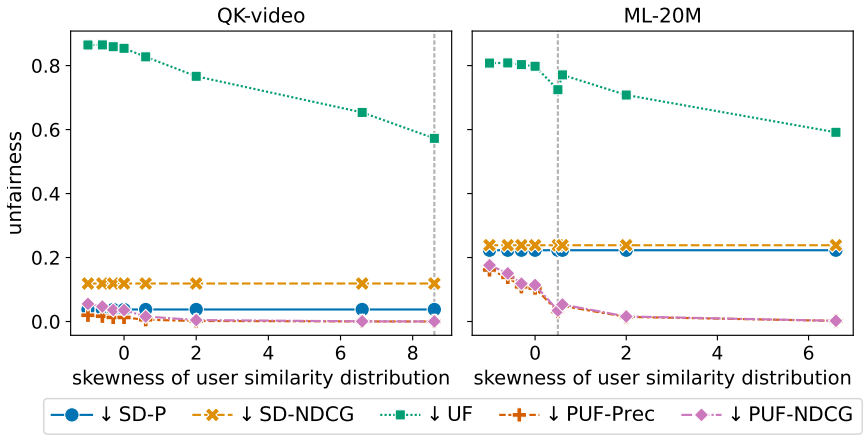


Figure 6.3: Artificially varying the skewness of the user similarity distribution for QK-video and ML-20M. Vertical grey lines denote the skewness corresponding to sim_{Jacc} observed in the dataset. The distribution skewness differs across datasets.

MME tends to have extremely small scores to begin with (§6.4.2), while UF does not depend on item relevance (Eq. (6.7)).

To summarise, we have shown that PUF and SD quantify fairness based on the disparity in recommendation effectiveness. All other existing Fair measures ignore the actual disparity in Eff scores. They merely reflect decreasing effectiveness or are not sensitive to the change in effectiveness. Next, we study whether the change in user similarity distribution is reflected in the Fair measures.

6.4.5 Varying User Similarity Distribution

Individual user fairness is defined based on user similarity. It is important to know if and how user similarity affects fairness. While two or more recommender models should be evaluated under the same similarity distribution, a desirable property of an individual user fairness measure is the ability to distinguish a single model performance across different similarity distributions, that may arise from various ways of quantifying user similarity. To this end, we investigate how the Fair measures respond to varying user similarity distributions, where the distribution is artificially generated.

Procedure. User similarity distributions tend to be right-skewed (many dissimilar users) for random users, and left-skewed (many similar users) for users that are friends [184]. Further, users are often dissimilar, as some users are new to the systems

or do not engage much, leading to discrepancies between the number of interactions among users and potentially affecting user similarity. Considering the above, we create synthetic user similarity scores by sampling from the Weibull distribution [231], which can be used to model skewed distributions. It has been used to model user rating distributions and sampling user neighbour candidates in RSs [2, 114, 115]. Its probability density function is $p(x) = \lambda x^{\lambda-1} \exp(-x^{\lambda})$. To obtain various right- and left-skewed distributions to represent possible user similarity distributions, we set $\lambda \in \{0.5, 1, 2, 5, 10, 50\}$. We also sample from the normal distribution $\mathcal{N}(0, 1)$, which has zero skews (i.e., equal portions of user pairs with similarity below/above the mean). We then min-max normalise the sampled similarity values to rescale them in the $[0, 1]$-range and randomly assign them to user pairs. We analyse non-random assignment in §6.4.6.

While the user similarity is artificial, we use the actual recommendation lists and scores from the NCL model runs as they perform relatively well. To save computation time, we only compute SD to represent all similarity-independent measures; theoretically, these measures will remain constant given no change in their input. We also compute PUF and UF, the two similarity-based measures. We compare the measure scores with the scores corresponding to the user similarity distribution observed in the datasets based on sim_{Jacc} (§6.4.2).

Results. We present the results for QK-video and ML-20M in Fig. 6.3; we show the rest in §6.7.1.3 as their trends are similar. We find that the similarity-based measures PUF and UF become fairer as skewness increases. Increasing skewness means a higher proportion of user pairs with low similarity, hence ↓PUF and ↓UF tend to be lower. Conversely, but as expected, the similarity-independent SD remains constant despite the change in skewness. For highly skewed similarities (skewness > 6), which is a realistic similarity distribution as seen in QK-video, SD-NDCG is somewhat unfair (≈ 0.2) for all datasets. Yet, PUF is almost perfectly fair ($\approx$0). Therefore, simply using SD may lead to the underestimation of fairness. While we only compute SD here, we expect other existing Fair measures to exhibit the same invariance as they are similarity-independent.

Between the two similarity-based measures, PUF is more sensitive towards negatively skewed user similarity distribution than UF. As skewness decreases, the mean user similarity increases at a slower rate. UF only considers user pairs above the mean-based similarity threshold, thus the number of user pairs contributing to the log sum in Eq. (6.7) decreases slower than in the right-skewed distributions. Another concern of UF is its relatively high unfairness compared to PUF and SD, even when most users are dissimilar. This may be because UF computes the pairwise distance of

the representation of the recommended items of a user pair (Eq. (6.9)). Minimising this distance is difficult as each k recommended item of a user must have a similar representation to each of the other user's k items, regardless of item relevance.

In short, we show that PUF can distinguish fairness levels across various similarity distributions, while non-similarity-based measures cannot. This distinction indicates the strength of PUF over SD (and indirectly over other non-similarity-based measures, i.e., all existing Fair measures except UF). We find that disregarding user similarity can also lead to the misinterpretation of fairness level.

In the following part, we further compare the two similarity-based measures and highlight the strengths of PUF over UF.

6.4.6 PUF and UF under Extreme Cases

We compare the similarity-based measures (PUF and UF) under extreme scenarios: can their scores reflect the difference in maximum and minimum fairness, across differences in the recommendation quality? Given an artificial set of pairwise user similarities and an artificial set of pairwise Eff score differences, to simulate the fairest case (MostFair), we sort these values and assign a higher similarity value to user pairs with lower Eff score difference. Separately, we assign a higher similarity score to pairs with higher Eff score difference to mimic the unfairest case (MostUnfair). For MostFair, a desirable measure would score close to 0 (the fairest). For MostUnfair, it should exhibit an inverted U-shape when the Eff score distribution is varied (i.e., similar to Fig. 6.2), as the maximum unfairness happens when the Eff distribution is the most uneven.

Procedure. We use an artificial, right-skewed user similarity distribution sampled from the Weibull distribution with $\lambda = 2$ (§6.4.5). We use the P@k and NDCG@k scores per user from the artificial runs in §6.4.4. To compute PUF- and UF-NDCG, we assign the user similarities to user pairs following the sorted pairwise difference of NDCG, and likewise for PUF- and UF-Prec.

Results. The results for QK-video and ML-20M are shown in Fig. 6.4; the rest are shown in §6.7.1.4 as we find similar trends. We see that the MostFair assignment yields PUF scores that are overall close to the fairest (0) across varying recommendation effectiveness, but UF remains constantly unfair ($\sim$0.8) nonetheless. This emphasises a mismatch between fairness computed based on the disparity of item representation and based on Eff score differences. PUF scores are more unfair for the MostUnfair than the MostFair scenario, while UF is almost invariant to

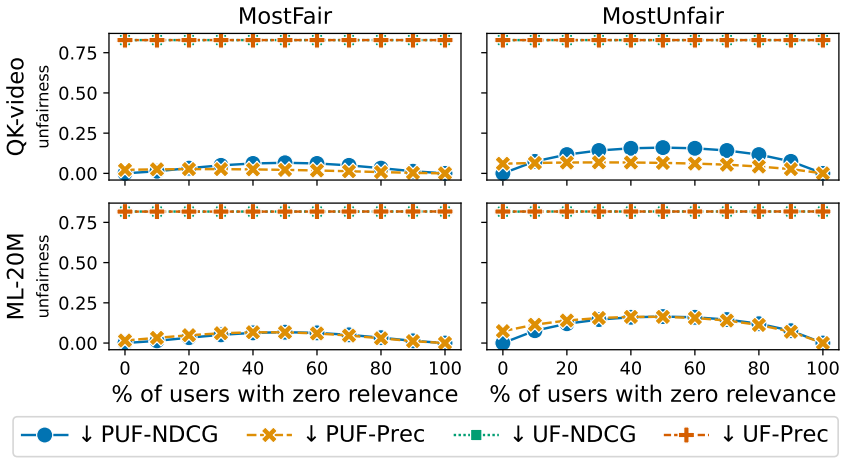


Figure 6.4: Artificially varying the % of users with zero relevance for QK-video and ML-20M. Lower Eff score difference is assigned to user pairs with higher similarity (MostFair), and to lower similarity (MostUnfair). Both UF overlap.

the change in similarity assignment between MostFair and MostUnfair. We observe again the insensitivity of UF towards varying recommendation effectiveness, as seen in Fig. 6.2.

In the MostUnfair case, ↓PUF ranges in $[0, 0.2)$ (which is close to the fairest) for all datasets. This is because the overall user similarity is low, which can happen in real-world scenarios. Hence, the maximum unfairness is expected to be relatively low, as PUF relies on user similarity to quantify fairness.

All in all, PUF correctly scores close to the fairest in the MostFair assignment, while UF fails to do so. PUF is notably more unfair for the MostUnfair rather than the MostFair case, while UF is almost constant for both cases and across varying recommendation effectiveness. In both cases, UF also overestimates effectiveness-based unfairness by constantly scoring much higher than PUF. Overall, PUF correctly quantifies maximal and minimal fairness, while UF does not. Next, we study the effect of using the same similarity measure for generating recommendations and evaluating fairness.

6.4.7 Using the Same Similarity Measure for PUF and Recommendation Model

We ask: if PUF and the recommendation model use the same similarity measure, would the model be the fairest according to PUF? The goal is to validate whether

PUF gives an advantage to models that use the same similarity function as PUF. To answer this, we use cosine similarity (sim_{cos}) for both PUF (PUF-*-Cos) and the U-KNN model (§6.7.1.1, Tab. 6.4). We find that U-KNN is never the fairest according to PUF. Even though U-KNN is the second-best model based on PUF for Lastfm, it is the second worst for ML-10M and it is the unfairest for ML-20M. So, using the same similarity measure for both PUF and the recommender model does not lead to the model being the fairest based on PUF. This may be due to the data splits: let two users have identical items in the train set, but completely different items in the test set. Even if both users get the same recommendation list, their effectiveness scores may differ and PUF cannot be the fairest. Therefore, minimising PUF to optimise for pairwise user fairness while maintaining an overall high mean effectiveness scores for all users, is not a trivial problem.

6.4.8 Using Item Features for PUF and UF

We next study what happens in a scenario where item features are available. PUF and UF are the only measures that can incorporate item features in their similarity functions. We compare PUF to UF when varying the importance of item features to them.

Procedure. We compute PUF and UF with sim_{UF} (Eq. (6.8)) and vary the weighted sum of its two components (Eq. (6.10)–(6.11)) with $\gamma \in \{0, 0.25, 0.5, 0.75, 1\}$. The weight $\gamma = 0$ accounts only for users' item feature distribution, while $\gamma = 1$ accounts only for their past interactions. We evaluate the groundwork runs[5] for the ML-* datasets, with movie genres as features. We cannot use the other datasets as QK-video only has two item categories and Lastfm has none.

Results. We present the results in Fig. 6.5. Overall, we find that PUF scores are fairer with higher weights of the users' past interactions.[6] Users' past interactions tend to be more diverse than their item category distributions, leading to more dissimilar users and less weight of user effectiveness disparity, which means fairer ↓PUF. UF follows the trend of PUF for ML-10M, but differs for ML-20M, where at first it becomes more unfair, then fairer with a higher γ. The non-linear trend of UF for ML-20M is due to its self-imposed thresholding, which results in highly different numbers of similar user pairs across various γ. While the effect of γ on UF can

[5]We do the same analysis for the artificial runs with various relevance score distributions (§6.4.4) and find similar results (§6.7.1.5).

[6]We also compute PUF-Prec; the trends are similar to PUF-NDCG (§6.7.1.5).

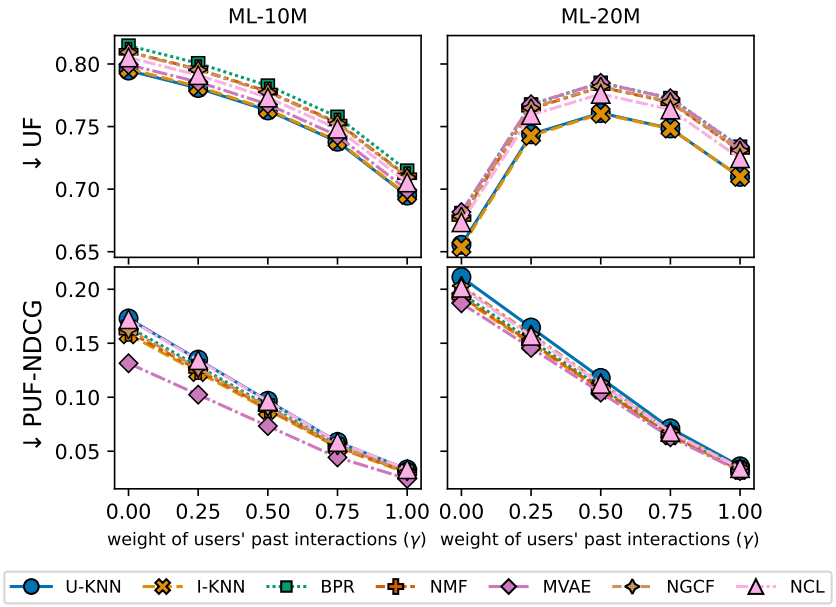


Figure 6.5: UF and PUF computed with user similarity sim_{UF} (Eq. (6.8)), varying the weighted sum of users' past interactions (sim_{Jacc}) and their item feature distribution (sim_{JS}) with γ.

be inconsistent for different datasets, it is more predictable for PUF, which always exhibits near-linear relationship.

For both datasets, UF is always more unfair than PUF, which is consistent with our prior results (§6.4.2, §6.4.4–§6.4.6). For both measures, the ordering of models given for a γ is maintained for another, i.e., the same conclusion of which model is better than another can be reached despite the different scores for different weights.

To sum up, varying the weights of interaction- and feature-based user similarity impacts PUF in a more predictable way than UF, making PUF superior to UF due to its more intuitive behaviour.

6.5 Related Work

In §6.2 we overviewed existing individual user fairness measures. Next, we discuss work that studies RS fairness measures empirically [175, 177, 178] and proposes pairwise individual fairness measures for ranking [71, 227]. Our work is close to [175], which studies item group fairness measures in RS, and [177, 178], which studies

evaluation measures of individual item fairness. Similarly to [175, 178], we find that fairness measures may disagree in their model ordering, and that some measures are more sensitive than others, given decreasing effectiveness and disparity in the recommendations. Out of all individual fairness measures in [177, 178], we often see that the similarity criterion of individual fairness is ignored, with only a few measures incorporating it. Recent work [71, 227] also proposes pairwise individual fairness measures, but these are for individual items, whereas our pairwise measure, PUF, is for individual users. Further, PUF considers user similarity, while the measure in [71] does not. The measure in [227] is similar to UF, as both employ thresholding of user similarities. Instead of applying a threshold, which can be arbitrary, PUF is weighted by user similarities, which introduces degrees of user similarity in the fairness computation. There exists also work on user group fairness (e.g., [64, 265]) or counterfactual fairness (e.g., [38]). Most such work requires sensitive attributes (e.g., gender), but public recommendation datasets with sensitive attributes tend to lack user representations (e.g., only binary genders) [35, 94, 247], and grouping users may require discretising the attribute (e.g., age) [30]. We focus on attribute-free individual fairness, rather than group or counterfactual fairness, to better assess distribution across all individuals [124] (often hidden in group fairness evaluation [71]).

6.6 Discussion and Conclusions

Current evaluation measures of individual user fairness in RSs either exclusively consider disparity in recommendation effectiveness or user similarity, but never both jointly. This means that none of these measures aligns with the definition of individual user fairness and one of the key objectives of RSs: user utility. To address this issue, we introduced PUF, a novel evaluation measure that quantifies user fairness through pairwise difference in effectiveness scores, while considering similarity between users. Note that accounting for both aspects is not trivial. For example, dividing the user similarity by the effectiveness disparity may result in an infinitely large score if the effectiveness disparity is small, and the score is undefined if the recommendation effectiveness is perfectly equal.

We showed that PUF is empirically superior to existing measures, based on the definition of individual fairness that we adopt (§6.2.1). Firstly, PUF is superior to non-similarity-based measures, such as SD. The non-similarity-based measures are insensitive to changes in user similarity distribution, which is a critical flaw as they cannot give a distinct fairness score for different levels of user similarity. Considering

both user similarity and effectiveness disparity is crucial as failing to do so may lead to over/underestimation of fairness. Secondly, PUF is also better than UF, the only existing measure that factors in user similarity. We have shown that PUF scores correspondingly to the disparity in user utility and successfully quantifies contrasting levels of fairness, while UF fails to do both. While the PUF design is simple, it is a novel, intuitive measure that is easy to interpret and robust against various effectiveness disparity or user similarity, which sets it apart from existing metrics, addressing a crucial gap in RS fairness evaluation.

In summary, we recommend using PUF over other existing individual user fairness measures due to its alignment with the definition of fairness, computational efficiency, and sensitivity to varying levels of user similarities and recommendation effectiveness in both typical and extreme cases. PUF paves the way for more elaborate measures in fairness evaluation, such as with graded relevance or other ways of quantifying user similarity (e.g., considering interaction recency).

6.7 Appendix

6.7.1 Extended Empirical Analysis

6.7.1.1 Comparison of all evaluation measures

We present in Tab. 6.4 the effectiveness (Eff) and fairness (Fair) scores of the 7 recommenders for 4 datasets.

6.7.1.2 Varying relevance score distribution

We present the results of artificially changing the relevance score distribution for Lastfm and ML-10M in Fig. 6.6.

6.7.1.3 Varying user similarity distribution

We present the results of artificially changing the user similarity distribution for Lastfm and ML-10M in Fig. 6.7.

6.7.1.4 PUF and UF under extreme cases

We present the results of computing PUF and UF under extreme cases for Lastfm and ML-10M in Fig. 6.8.

Table 6.4: Effectiveness (Eff) and fairness (Fair) scores at $k = 10$ for recommenders. The most effective/fair score per measure is bolded. ↑/↓ means the higher/lower the better.

			U-KNN	I-KNN	BPR	NMF	MVAE	NGCF	NCL
Lastfm	Eff	↑ HR	0.7102	0.7647	0.7729	0.7424	0.7783	0.7707	**0.7930**
		↑ MRR	0.4386	0.4835	0.4916	0.4641	0.4764	0.4954	**0.5026**
		↑ P	0.1532	0.1721	0.1782	0.1649	0.1758	0.1779	**0.1840**
		↑ MAP	0.1190	0.1369	0.1414	0.1291	0.1377	0.1418	**0.1478**
		↑ R	0.1899	0.2180	0.2238	0.2069	0.2241	0.2236	**0.2342**
		↑ NDCG	0.2167	0.2450	0.2520	0.2334	0.2473	0.2530	**0.2611**
	Fair (existing)	↓ SD-P	**0.1485**	0.1529	0.1538	0.1505	0.1518	0.1527	0.1552
		↓ SD-NDCG	**0.2040**	0.2091	0.2092	0.2072	0.2084	0.2085	0.2124
		↓ Gini-P	0.5146	0.4768	0.4668	0.4889	0.4654	0.4642	**0.4554**
		↓ Gini-NDCG	0.5213	0.4789	0.4678	0.4953	0.4735	0.4646	**0.4581**
		↓ ME	0.0897	**0.0827**	0.0904	0.0983	0.0973	0.0897	0.0886
		↓ MME	0.0025	**0.0017**	0.0020	0.0025	0.0020	0.0021	0.0019
		↓ PEU	0.6525	**0.6176**	0.6318	0.6765	0.6531	0.6275	0.6193
		↓ UF	**0.6615**	0.6616	0.6660	0.6667	0.6685	0.6675	0.6681
	Fair (PUF)	↓ PUF-Prec-Cos	0.0097	0.0100	0.0100	**0.0096**	0.0098	0.0099	0.0100
		↓ PUF-Prec-Jacc	0.0073	0.0075	0.0075	**0.0073**	0.0074	0.0075	0.0076
		↓ PUF-NDCG-Cos	0.0134	0.0137	0.0136	**0.0132**	0.0135	0.0135	0.0136
		↓ PUF-NDCG-Jacc	0.0101	0.0103	0.0102	**0.0099**	0.0101	0.0101	0.0102
QK-video	Eff	↑ HR	0.1155	0.0396	0.0993	0.0857	0.1093	0.1238	**0.1298**
		↑ MRR	0.0435	0.0131	0.0390	0.0303	0.0392	**0.0484**	0.0481
		↑ P	0.0122	0.0041	0.0105	0.0088	0.0116	0.0134	**0.0140**
		↑ MAP	0.0188	0.0046	0.0170	0.0128	0.0178	**0.0218**	0.0218
		↑ R	0.0517	0.0141	0.0433	0.0369	0.0506	0.0589	**0.0608**
		↑ NDCG	0.0331	0.0091	0.0290	0.0231	0.0314	0.0375	**0.0381**
	Fair (existing)	↓ SD-P	0.0350	**0.0205**	0.0327	0.0293	0.0343	0.0373	0.0376
		↓ SD-NDCG	0.1089	**0.0539**	0.1058	0.0910	0.1079	0.1191	0.1188
		↓ Gini-P	0.8906	0.9618	0.9060	0.9168	0.8968	0.8850	**0.8790**
		↓ Gini-NDCG	0.9215	0.9733	0.9344	0.9425	0.9265	0.9156	**0.9120**
		↓ ME	0.0874	0.0957	0.0962	0.0757	**0.0412**	0.0899	0.0622
		↓ MME	0.0012	**0.0003**	0.0010	0.0014	0.0011	0.0011	0.0012
		↓ PEU	0.7746	0.9087	0.8426	0.6810	**0.3867**	0.8056	0.5626
		↓ UF	0.5731	0.5751	0.5739	0.5736	**0.5721**	0.5734	0.5726
	Fair (PUF)	↓ PUF-Prec-Cos	0.0001	**0.0000**	0.0001	0.0001	0.0001	0.0001	0.0001
		↓ PUF-Prec-Jacc	0.0001	**0.0000**	0.0001	0.0001	0.0001	0.0001	0.0001
		↓ PUF-NDCG-Cos	0.0003	**0.0001**	0.0003	0.0002	0.0003	0.0003	0.0003
		↓ PUF-NDCG-Jacc	0.0002	**0.0001**	0.0002	0.0002	0.0002	0.0002	0.0002

Table 6.4 (continued): Effectiveness (Eff) and fairness (Fair) scores at $k = 10$ for recommenders. The most effective/fair score per measure is bolded. ↑/↓ means the higher/lower the better.

			U-KNN	I-KNN	BPR	NMF	MVAE	NGCF	NCL
ML-10M	Eff	↑ HR	0.5207	0.4872	0.5121	0.5141	0.4169	0.5128	**0.5213**
		↑ MRR	**0.3105**	0.2818	0.2987	0.2915	0.2372	0.2865	0.3019
		↑ P	0.1531	0.1369	0.1458	0.1433	0.1066	0.1460	**0.1536**
		↑ MAP	**0.1027**	0.0887	0.0953	0.0911	0.0666	0.0930	0.1009
		↑ R	**0.0276**	0.0222	0.0249	0.0246	0.0203	0.0246	0.0263
		↑ NDCG	**0.1691**	0.1501	0.1600	0.1560	0.1191	0.1579	0.1673
	Fair (existing)	↓ SD-P	0.2206	0.2080	0.2125	0.2076	**0.1781**	0.2109	0.2195
		↓ SD-NDCG	0.2376	0.2244	0.2297	0.2241	**0.1960**	0.2262	0.2367
		↓ Gini-P	0.6884	0.7076	0.6906	0.6874	0.7415	0.6880	**0.6853**
		↓ Gini-NDCG	**0.6889**	0.7123	0.6954	0.6948	0.7460	0.6943	0.6911
		↓ ME	0.1017	0.1131	0.1290	0.1307	**0.0726**	0.1284	0.1192
		↓ MME	0.0039	0.0040	0.0046	0.0046	**0.0025**	0.0043	0.0044
		↓ PEU	0.5936	0.6356	0.6927	0.7072	**0.4668**	0.7026	0.6619
		↓ UF	**0.6948**	0.6954	0.7148	0.7103	0.6994	0.7095	0.7053
	Fair (PUF)	↓ PUF-Prec-Cos	0.0493	0.0445	0.0472	0.0458	**0.0363**	0.0469	0.0497
		↓ PUF-Prec-Jacc	0.0298	0.0268	0.0285	0.0276	**0.0217**	0.0284	0.0302
		↓ PUF-NDCG-Cos	0.0547	0.0495	0.0523	0.0507	**0.0409**	0.0515	0.0549
		↓ PUF-NDCG-Jacc	0.0331	0.0298	0.0316	0.0306	**0.0245**	0.0311	0.0333
ML-20M	Eff	↑ HR	**0.5092**	0.4881	0.5051	0.4927	0.4890	0.5083	0.5051
		↑ MRR	0.2970	0.2799	0.2928	0.2781	0.2593	**0.3007**	0.2934
		↑ P	**0.1582**	0.1423	0.1449	0.1433	0.1409	0.1517	0.1500
		↑ MAP	**0.1079**	0.0923	0.0957	0.0934	0.0889	0.1010	0.0997
		↑ R	**0.0213**	0.0193	0.0193	0.0186	0.0185	0.0199	0.0200
		↑ NDCG	**0.1711**	0.1538	0.1584	0.1544	0.1481	0.1649	0.1626
	Fair (existing)	↓ SD-P	0.2327	**0.2145**	0.2157	0.2160	0.2152	0.2239	0.2228
		↓ SD-NDCG	0.2484	0.2290	0.2333	0.2324	**0.2278**	0.2399	0.2385
		↓ Gini-P	**0.6981**	0.7068	0.6991	0.7054	0.7100	0.6983	0.7001
		↓ Gini-NDCG	**0.7021**	0.7110	0.7053	0.7138	0.7203	0.7022	0.7046
		↓ ME	**0.1034**	0.1199	0.1361	0.1412	0.1421	0.1348	0.1337
		↓ MME	**0.0038**	0.0049	0.0046	0.0052	0.0049	0.0046	0.0050
		↓ PEU	**0.5877**	0.6327	0.7029	0.6956	0.7071	0.6910	0.6869
		↓ UF	0.7098	**0.7098**	0.7339	0.7305	0.7336	0.7309	0.7246
	Fair (PUF)	↓ PUF-Prec-Cos	0.0553	0.0502	0.0504	0.0503	**0.0497**	0.0529	0.0525
		↓ PUF-Prec-Jacc	0.0329	0.0298	0.0299	0.0299	**0.0296**	0.0314	0.0312
		↓ PUF-NDCG-Cos	0.0605	0.0548	0.0559	0.0552	**0.0532**	0.0581	0.0576
		↓ PUF-NDCG-Jacc	0.0361	0.0326	0.0333	0.0328	**0.0317**	0.0346	0.0343

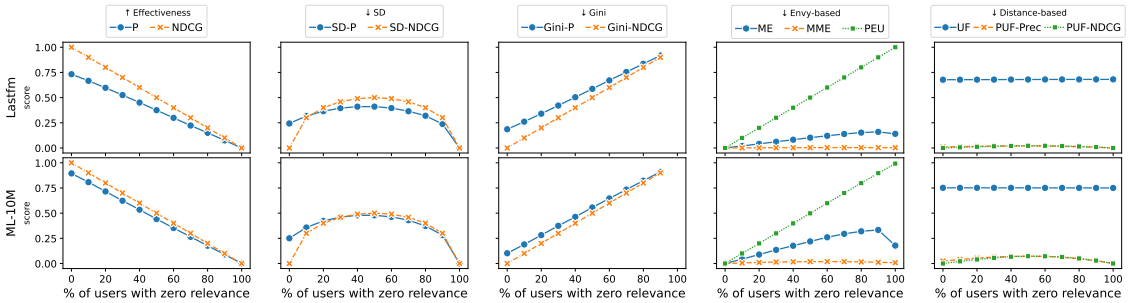


Figure 6.6: Effectiveness (Eff) and fairness (Fair) scores of Lastfm and ML-10M, when artificially varying % of users with all irrelevant items (zero relevance), and the rest of the users receiving all relevant items. All PUF variants overlap. Gini is missing points at 100% users with zero relevance as it is undefined when each user has zero Eff scores.

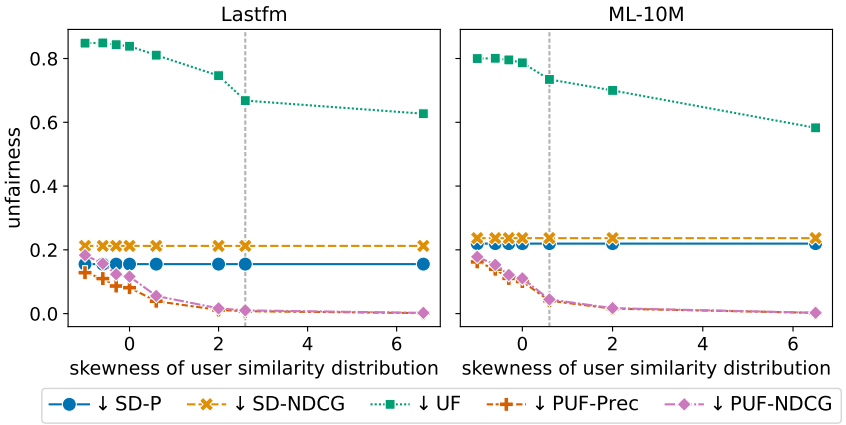


Figure 6.7: Artificially varying the skewness of the user similarity distribution for Lastfm and ML-10M. Vertical grey lines denote the skewness corresponding to sim_{Jacc} observed in the dataset. The distribution skewness differs across datasets.

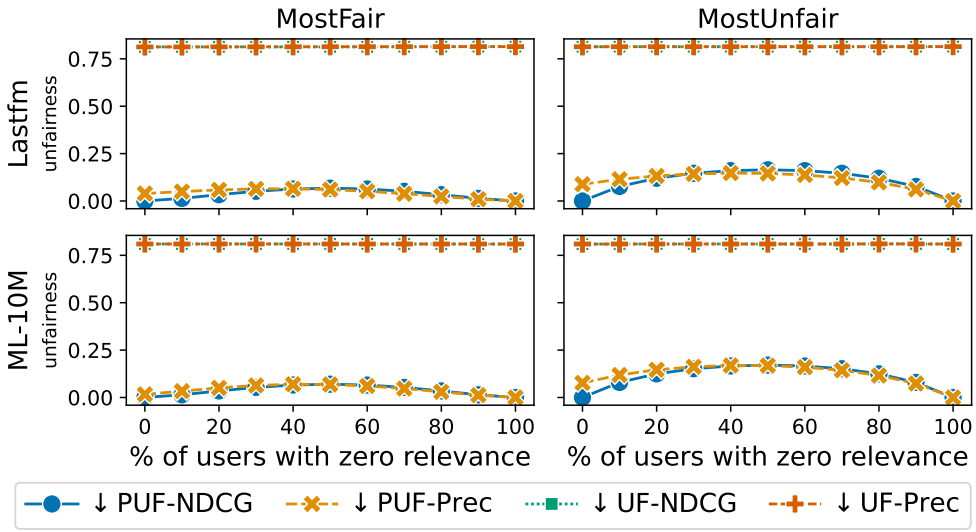


Figure 6.8: Artificially varying the % of users with zero relevance for Lastfm and ML-10M. Lower Eff score difference is assigned to user pairs with higher similarity (MostFair), and to lower similarity (MostUnfair). Both UF overlap.

6.7.1.5 Using item features for PUF and UF

Fig. 6.9 shows the results of computing PUF-Prec, where user similarity sim_{UF} (Eq. (6.8)) is computed with varying weighted sums of user interactions and their item feature distribution. Fig. 6.10 shows the results of computing UF and PUF for varying weights of user interactions in sim_{UF} and artificially varying relevance distribution (i.e., the runs from §6.4.4).

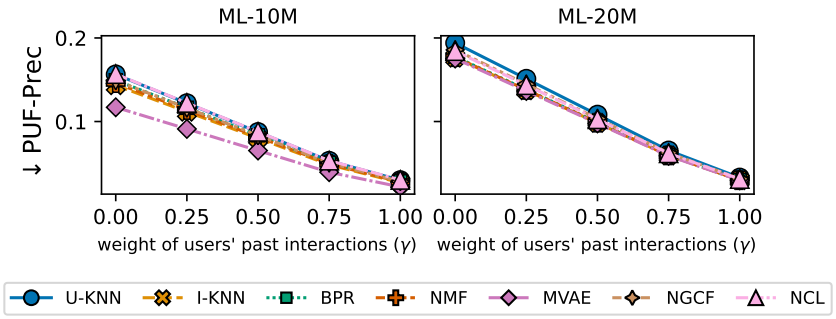


Figure 6.9: PUF-Prec computed with user similarity sim_{UF} (Eq. (6.8)), varying the weighted sum of the users' past interactions (sim_{Jacc}) and their item feature distribution (sim_{JS}) with γ.

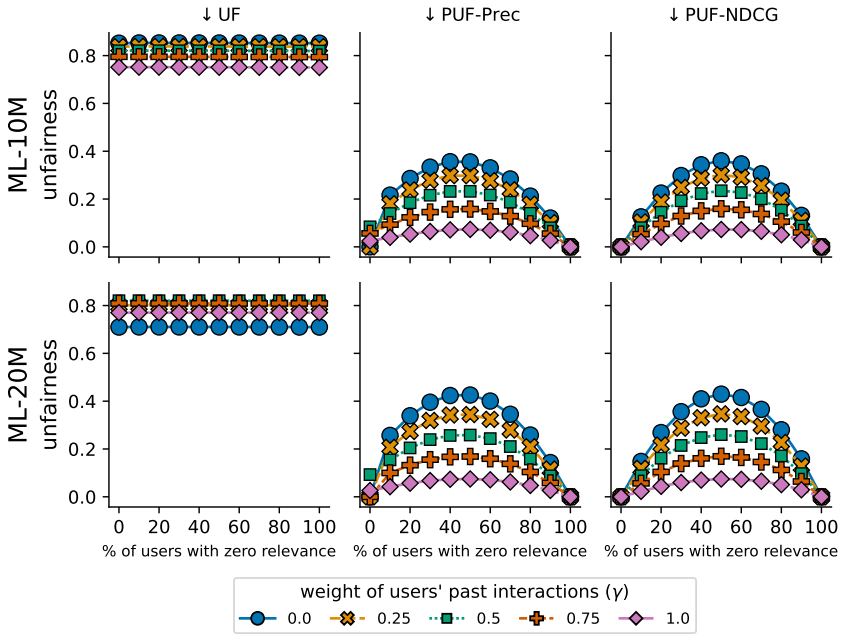


Figure 6.10: Varying the weighted sum of users' past interactions (sim_{Jacc}) and their item feature distribution (sim_{JS}) with γ and artificially varying the % of users with zero relevance.

Chapter 7

Stairway to Fairness: Connecting Group and Individual Fairness

Abstract

Fairness in recommender systems (RSs) is commonly categorised into group fairness and individual fairness. However, there is no established scientific understanding of the relationship between the two fairness types, as prior work on both types has used different evaluation measures or evaluation objectives for each fairness type, thereby not allowing for a proper comparison of the two. As a result, it is currently not known how increasing one type of fairness may affect the other. To fill this gap, we study the relationship of group and individual fairness through a comprehensive comparison of evaluation measures that can be used for both fairness types. Our experiments with 8 RSs across 3 datasets show that recommendations that are highly fair for groups can be very unfair for individuals. Our finding is novel and useful for RS practitioners aiming to improve the fairness of their systems.

7.1 Introduction

Fairness in Recommender Systems (RSs) can be evaluated for groups and for individuals. *Group fairness* typically refers to having equitable outcome across groups (e.g., similar effectiveness between groups of users [64]), while *individual fairness* is commonly defined as treating similar users/items equally [228] (e.g., similar effectiveness for all users [236]). Conceptually, prior work [58, 133, 204] discusses how RSs can be

fair to groups and at the same time unfair to individuals, or vice versa, but no work has empirically studied how this practically occurs in RS fairness evaluation.

Prior work either: (i) evaluates fairness exclusively for groups or individuals [54]; or (ii) evaluates both, but with two different families of measures [75, 166, 171] or for two fairness subjects/objectives [183, 236]. Evaluating group and individual fairness with different families of measures makes comparison difficult, as the measure scores may differ in sensitivity, or in theoretical and empirical ranges [177, 178, 198]. Likewise, it is not possible to properly compare group and individual fairness when each is evaluated for a distinct fairness objective, e.g., recommendation effectiveness disparity across individual users vs. exposure disparity between item groups [236]. To address this gap, we evaluate user-side group and individual fairness with the same families of measures that can quantify both, to study the relationship between evaluation measures of user-side group and individual fairness. This work is the first empirical study that compares the 9 existing user-side fairness evaluation measures for groups with those for individuals. We ask:

1. To what extent do group and individual fairness evaluation measures differ in their conclusions?
2. For the same family of measures, how different are the group and individual fairness scores?
3. How do different ways of grouping users affect between- and within-group fairness?
4. How do between- and within-group fairness relate to individual fairness?

Our results show that group fairness measures often hide unfairness within groups and between individuals, highlighting the importance of evaluating fairness beyond the between-group level.

7.2 Methodology

We compare evaluation measures of user-side individual and group fairness in RSs, considering multiple ways of grouping users.[1]

[1]Our code is freely available at: BLINDED FOR REVIEW.

Datasets

To enable group fairness analysis, three datasets with ≥ 3 user profile features are selected (see Tab. 7.1 for statistics).

ML-1M [94] has 1,000,029 movie ratings (1–5) from 6K users. Users with no/unspecified self-reported gender, age, or occupation are removed, and we exclude users under 18 years to avoid processing the data of minors. We focus on recommending preferred movies, so ratings <3 are discarded, and the levels 4 and 5 are mapped to 1.

JobRec [93] has 1.6M job applications from 321K users. Given a user's application history, we focus on recommending job titles that may suit them, keeping only users with information for degree, major, and years of experience. Users with more than 60 years of experience are removed (as this likely indicates erroneous entries).

LFM-1B [196] has 1,088,161,692 music playcounts, from ∼120K users. We focus on recommending new track artists for a user to listen to, other than the ones they have listened to in the past, using the dataset after deduplication based on the artist, with 65M interactions (provided by `RecBole` [260]). The deduplication summarises the total playcount per artist and keeps the last event timestamp. Users without countries, age, or gender information are removed, as are minors (as in ML-1M), and users with age >100 years (as this likely indicates erroneous entries).

Items without name/title are removed from all datasets. To reduce data sparsity, which may affect LLMRec performance [105], we keep users and items with ≥ 5 interactions (5-core filtering) for ML-1M and JobRec. We apply 50-core filtering [145, 232, 259] to LFM-1B, as it is highly sparse with 5-core filtering. The data is temporally split for train/val/test with a ratio of 3:1:1 using a global timeline [151]. From all splits, users and items with $\leq t$ interactions in the train set are removed. A high t can result in very few unique users in the test set, so we choose t such that at least 500 test users remain. For ML-1M and LFM-1B, we set $t = 5$ [240]. For JobRec, we use $t = 2$. We remove users/items in the val and test sets that are not in the train set.

User Grouping

To study group fairness, we cluster users based on their sensitive attributes (see Tab. 7.1 and §7.5.1.1). Users cannot belong to two groups at the same time, e.g., age<50 and age≥50.

For ML-1M, we use gender, age, and occupation as sensitive attributes. Gender is used as is. Age is grouped into: 18–24 years, 25–49 years, and ≥50 years [160]. User

Table 7.1: Preprocessed dataset statistics. n_{G_a} is the number of groups for sensitive attribute a. We exclude empty groups.

	ML-1M [94]	JobRec [93]	LFM-1B [196]
#interaction (all splits)	467,218	210,921	15,024,267
#item (all splits)	3,030	19,912	51,204
#user (test set)	620	523	16,611
sensitive attr. #1 (n_{G_1})	gender (2)	degree (3)	gender (2)
sensitive attr. #2 (n_{G_2})	age (3)	years of experience (3)	age (3)
sensitive attr. #3 (n_{G_3})	occupation (2)	major (6)	country (5)
#intersectional groups	12	36	29
min–max subgroup size	2–279	1–70	1–3,919
median subgroup size	31	7	59

occupation is grouped into: non-working (student [70, 219], homemaker, retired, and unemployed) and working (14 occupations ranging from farmer to executive [218]).

For JobRec, we consider the user's academic degree, years of working experience, and study major as the sensitive attributes. Degree is grouped into: high school, college (associate or vocational degree), and university (bachelor's, master's, and PhD). Years of experience are grouped into: $\leq$5 years, >5–10 years, and >10 years. We group study majors into six fields of study, as per Xu et al. [240], using manual annotation and fuzzy string matching.[2]

For LFM-1B, the sensitive attributes are gender, age, and country. Gender and age are processed as for ML-1M, and the user's country is mapped to the continent.[3] Users from the North/South Americas are grouped together with Antarctica [88, 240].

LLM-based Recommenders

Recent work has utilised Large Language Models (LLMs) as recommenders (LLMRecs), with promising results [100]. Unlike collaborative filtering models, LLMRecs can easily handle fine-grained user attributes, although including sensitive attributes in the prompt can impact effectiveness and fairness [50, 240, 255]. We therefore study the effectiveness and fairness of LLMRecs under few-shot learning. To ensure comparable performance, we use four open-source, similar-sized LLMs released

[2]Details are provided in §7.5.1.1.

[3]We use the country-continent mapping from `https://gist.github.com/achuhunkin/6cb1cbceb23395300aa209aad09e6e5d`, and manually group transcontinental countries.

in July–Nov'24: Llama-3.1-8B-Instruct [90], Qwen2.5-7B-Instruct [214], GLM-4-9B-chat [86], and Ministral-8B-Instruct-2410 [213]. The temperature is fixed at 0 for each LLM to obtain deterministic output.

The LLMs are prompted using in-context learning strategy [100], as this has been shown to outperform sequential and recency-focused prompting.[4] We only prompt for users that exist in the test set, as otherwise it is not possible to evaluate the recommendation effectiveness. In the prompt, we provide the users' train items as the interaction history and the val items as the few-shot samples. A maximum of 10 most recent train and val items each are provided, as having too many items in the prompt may reduce effectiveness [100, 105]. To avoid inflated performance, we do not prompt with a sampled candidate item list [120]. Instead, we add restriction in the prompt to narrow down the item search space, e.g., for ML-1M, the movies should be between certain years, based on the movie release year in the metadata file. For LFM-1B, the prompt also includes the playcount, which is important in music recommendation [88].

Based on the inclusion/exclusion of user sensitive attributes, we create two prompt types [50, 51, 215, 255]: Sensitive (S), which has both interaction history and all three sensitive attributes, and Non-Sensitive (NS), which has only the interaction history.

To evaluate LLMRecs, we perform string matching between the list of recommended items and items in the test set, by using the TF-IDF [108] of the items' character-based n-gram [134]. The recommended item is deemed relevant if similarity ≥ 0.75. Our LLMRecs experiments are carried out with `vllm` [122] and `RecLM-eval` [134].

Evaluation

Recommendation effectiveness (Eff) and fairness (Fair) are measured at $k = 10$ for all LLMRecs. For Eff, the mean Hit Rate (HR), MRR, P@k (P), and NDCG@k [102] are computed over all users. Group and individual Fair measures are computed in two steps: first, computing an Eff score per user/group as a 'base score'; and second, aggregating the 'base score' between users/groups with a Fair measure. P and NDCG are used as base scores to represent set- and rank-based measures.

Group fairness. We compute all existing fairness measures for two or more user groups in RSs,[5] that are published up to March 2025: Average scores of the worst 25% groups (Min [229]), Range [142], SD [142, 255], MAD [77], Gini [75, 84, 166], CV [266], FStat [224], KL [9], and GCE [52, 53]. We also compute the Atkinson Index

[4]The full prompt templates and examples are provided in §7.5.1.2.

[5]Measures that can only be used for exactly two groups are excluded.

(Atk [10]), an income inequality measure that considers within-group variations.[6] The between- and within-group fairness version of the measures are denoted as $\cdot_{\text{b-group}}$ and $\cdot_{\text{w-group}}$ respectively.[7] We provide the measure formulations and technical details in §7.5.1.3.

Individual fairness. Fairness for individual users is quantified with SD [167], Gini [126], and Atk. The subscript $\cdot_{\text{ind}}$ indicates the individual fairness version of the measure. SD and Gini are the only Fair measures that have been used for both individual and group user fairness, while Atk_{ind} can be decomposed into between- and within-group fairness with no residuals [21, 24, 47]. While other group Fair measures can also be used to measure individual fairness, their scores may not be informative, e.g., Min may be zero for most models, as having most users scoring P=0 or NDCG=0 is common.

7.3 Empirical Analysis

Evaluation of all LLMRecs

Eff scores and NDCG-based Fair scores for group fairness and individual fairness of LLMRecs are shown in Tab. 7.2; the P-based Fair measures are presented in §7.5.2.1 as they show similar results. We evaluate group fairness considering the intersectionality of all 3 sensitive user attributes.

Eff and Fair scores. GLM-4-9B has the best Eff and Fair scores for all datasets, except for JobRec, where all LLMRecs have equally low Eff. This could be because the job titles are very specific and contain more noise than movie/artist names (e.g., 'Chicago Concierge Manager', 'Account Executive - 40k - 70K per year'). Generally, as group/individual Fair measures, Gini, Atk, and CV are fairer when NDCG is better; this is because their equations include division by the mean NDCG. As such, these measures can distinguish group fairness between two models that have similar $\text{SD}_{\text{b-group}}/\text{SD}_{\text{ind}}$ but differ in NDCG (e.g., the sensitive and non-sensitive prompts of Ministral-8B for LFM-1B), or vice versa.

Comparison of sensitive and non-sensitive prompts. The Eff and Fair scores between sensitive (S) and non-sensitive (NS) prompts are generally comparable, implying that including sensitive attributes in the prompt has little effect on effectiveness/fairness. An exception to this is Ministral-8B for LFM-1B, which may be due to the LLM's relatively high gender bias [42].

[6] This measure can be transformed into Generalised Entropy [202, 212].

[7] The term *group fairness* refers to between-group fairness; the latter is used when we compare fairness between and within groups.

Table 7.2: Effectiveness (Eff) and fairness (Fair) scores at cut-off $k = 10$ for intersectional groups (Grp.) and for individuals (Ind.) of LLMRecs with non-sensitive (NS) and sensitive (S) prompts. All Fair scores are computed with NDCG. All measures range in [0,1], except for the Grp. measures below them. The best Eff/Fair scores are bolded. Darker green marks scores closer to the best Eff/Fair per measure. ↑ / ↓ means the higher/lower the better.

	LLMRec	GLM-4-9B		Llama-3.1-8B		Ministral-8B		Qwen2.5-7B	
	prompt type	NS	S	NS	S	NS	S	NS	S
					ML-1M				
Eff	↑ HR	**0.377**	0.358	0.260	0.269	0.342	0.340	0.363	0.371
	↑ MRR	**0.189**	0.174	0.101	0.113	0.159	0.173	0.165	0.180
	↑ NDCG	**0.231**	0.215	0.140	0.148	0.200	0.208	0.207	0.221
Fair (Grp.)	↑ Min	**0.166**	0.137	0.086	0.108	0.059	0.107	0.068	0.091
	↓ Range	**0.188**	0.208	0.356	0.376	0.290	0.500	0.328	0.274
	↓ SD	**0.055**	0.061	0.088	0.088	0.085	0.125	0.093	0.077
	↓ MAD	**0.067**	0.076	0.091	0.087	0.099	0.142	0.112	0.090
	↓ Gini	**0.130**	0.161	0.255	0.226	0.248	0.275	0.255	0.195
	↓ Atk	**0.015**	0.020	0.036	0.025	0.033	0.049	0.053	0.039
	↓ CV	**0.233**	0.285	0.540	0.502	0.465	0.525	0.460	0.365
	↓ FStat	**0.468**	0.714	0.841	0.654	0.767	1.278	1.645	1.220
	↓ KL	1.121	1.138	1.674	1.640	**0.866**	1.671	1.198	1.063
	↓ GCE	**0.028**	0.050	0.112	0.104	659.844	659.741	0.239	0.198
Fair (Ind.)	↓ SD	0.330	0.320	**0.262**	0.272	0.309	0.322	0.308	0.324
	↓ Gini	**0.705**	0.721	0.799	0.793	0.736	0.736	0.721	0.715
	↓ Atk	**0.636**	0.655	0.751	0.742	0.672	0.673	0.652	0.644
					JobRec				
Eff	↑ HR	0.054	0.033	0.044	0.021	0.057	0.044	**0.065**	0.057
	↑ MRR	0.037	0.023	0.017	0.009	**0.048**	0.033	0.046	0.045
	↑ NDCG	0.041	0.025	0.024	0.012	**0.050**	0.036	**0.050**	0.048
Fair (Grp.)	↑ Min	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
	↓ Range	0.500	**0.085**	0.500	0.500	0.333	0.170	0.333	0.333
	↓ SD	0.093	**0.024**	0.086	0.082	0.074	0.042	0.072	0.071
	↓ MAD	0.066	**0.019**	0.053	0.035	0.065	0.035	0.065	0.056
	↓ Gini	0.828	0.836	0.857	0.939	0.757	0.789	**0.740**	0.808
	↓ Atk	0.545	0.525	0.676	0.833	0.520	0.498	**0.472**	0.499
	↓ CV	2.392	2.114	2.883	4.549	1.758	1.923	**1.706**	2.102
	↓ FStat	1.034	**0.562**	1.612	2.136	1.003	0.611	0.909	0.928
	↓ KL	3.208	**1.445**	3.969	5.855	1.873	1.747	2.060	2.279
	↓ GCE	1685.925	1979.103	1685.992	2052.498	1612.573	1759.176	**1539.277**	1759.185
Fair (Ind.)	↓ SD	0.183	0.147	0.121	**0.090**	0.211	0.174	0.204	0.203
	↓ Gini	0.956	0.974	0.966	0.985	0.949	0.963	**0.947**	0.951
	↓ Atk	0.948	0.969	0.958	0.980	0.944	0.957	**0.937**	0.944
					LFM-1B				
Eff	↑ HR	0.658	**0.661**	0.609	0.618	0.451	0.317	0.375	0.317
	↑ MRR	**0.409**	0.408	0.347	0.357	0.266	0.174	0.199	0.160
	↑ NDCG	0.462	**0.463**	0.406	0.415	0.310	0.208	0.241	0.198
Fair (Grp.)	↑ Min	0.232	**0.262**	0.254	0.252	0.109	0.060	0.090	0.043
	↓ Range	0.604	0.604	0.884	0.931	0.993	1.000	0.525	**0.378**
	↓ SD	0.149	0.131	0.161	0.144	0.178	0.177	0.119	**0.095**
	↓ MAD	0.141	0.126	0.163	0.137	0.168	0.156	0.128	**0.107**
	↓ Gini	0.167	**0.146**	0.191	0.171	0.284	0.357	0.268	0.306
	↓ Atk	0.003	**0.002**	0.004	0.004	0.006	0.014	0.008	0.019
	↓ CV	0.366	**0.315**	0.393	0.372	0.624	0.841	0.516	0.561
	↓ FStat	2.348	**2.019**	3.224	4.067	2.874	3.638	2.601	4.788
	↓ KL	2.608	2.740	3.447	3.139	3.001	3.489	2.885	**2.354**
	↓ GCE	338.843	225.901	112.993	**112.988**	451.823	451.872	451.824	564.768
Fair (Ind.)	↓ SD	0.380	0.378	0.370	0.371	0.377	0.334	0.342	**0.320**
	↓ Gini	0.462	**0.461**	0.510	0.502	0.638	0.750	0.703	0.753
	↓ Atk	0.361	**0.358**	0.409	0.400	0.563	0.694	0.638	0.695

Group vs individual fairness. For Atk/Gini, the fairest model for intersectional groups and for individual users are always the same, but this is not the case for SD. Thus, if sensitive attribute information is missing, $\text{Atk}_{\text{ind}}/\text{Gini}_{\text{ind}}$ can estimate which model is the fairest for intersectional groups.

Agreement of group and individual fairness measures

Do individual and group FAIR measures reach the same conclusion? If an individual FAIR measure can rank models from the most to the least fair equivalently to a group FAIR measure, one can be a proxy for the other. To assess measure agreement in ranking models, we compute Kendall's τ; we consider two rankings to be "equivalent" if $\tau \geq .9$ [143]. Fig. 7.1 shows group and individual FAIR measure agreement.

Our results show that **no individual Fair measure consistently has equivalent rankings to any group Fair measures**.[8] For all datasets, group FAIR measures tend to have weak-to-strong agreement with $\text{Gini}_{\text{ind}}/\text{Atk}_{\text{ind}}$ and have weak-to-strong disagreement with SD_{ind}. The only group FAIR measure that always correlates strongly to an individual FAIR measure is CV ($\tau \in [.71, .79]$). This may be due to the measures' similar formulation: $\text{Gini}_{\text{ind}}/\text{Atk}_{\text{ind}}$ have division by the mean EFF scores across all users, while CV has division by the mean of mean EFF score per group. **In summary, no existing individual Fair measures make a reliable proxy for group Fair measures, hence the need to evaluate for both.**

Intersectional fairness

As we consider more intersectional attributes to form user groups, the number of groups grows. We posit that achieving intersectional group fairness is harder than for non-intersectional groups as it requires having similar EFF scores for a larger number of groups. Likewise, individual fairness (i.e., having a group for each user) would be worse than group fairness. We study: (i) if/how the number of attributes used for grouping affect fairness; and (ii) the difference between individual and (intersectional) group FAIR scores. To this end, we compute NDCG-based individual FAIR scores across all users, and average NDCG-based group FAIR scores across all ways of grouping users when considering only 1, 2, or 3 attribute(s) at once. SD, Gini, and Atk are computed, as they have been used as both group and individual FAIR

[8] We find similar results for the group-individual fairness agreement between the same family of measures (e.g., $\text{SD}_{\text{b-group}}$ vs SD_{ind}) for all possible groupings (§7.5.2.2).

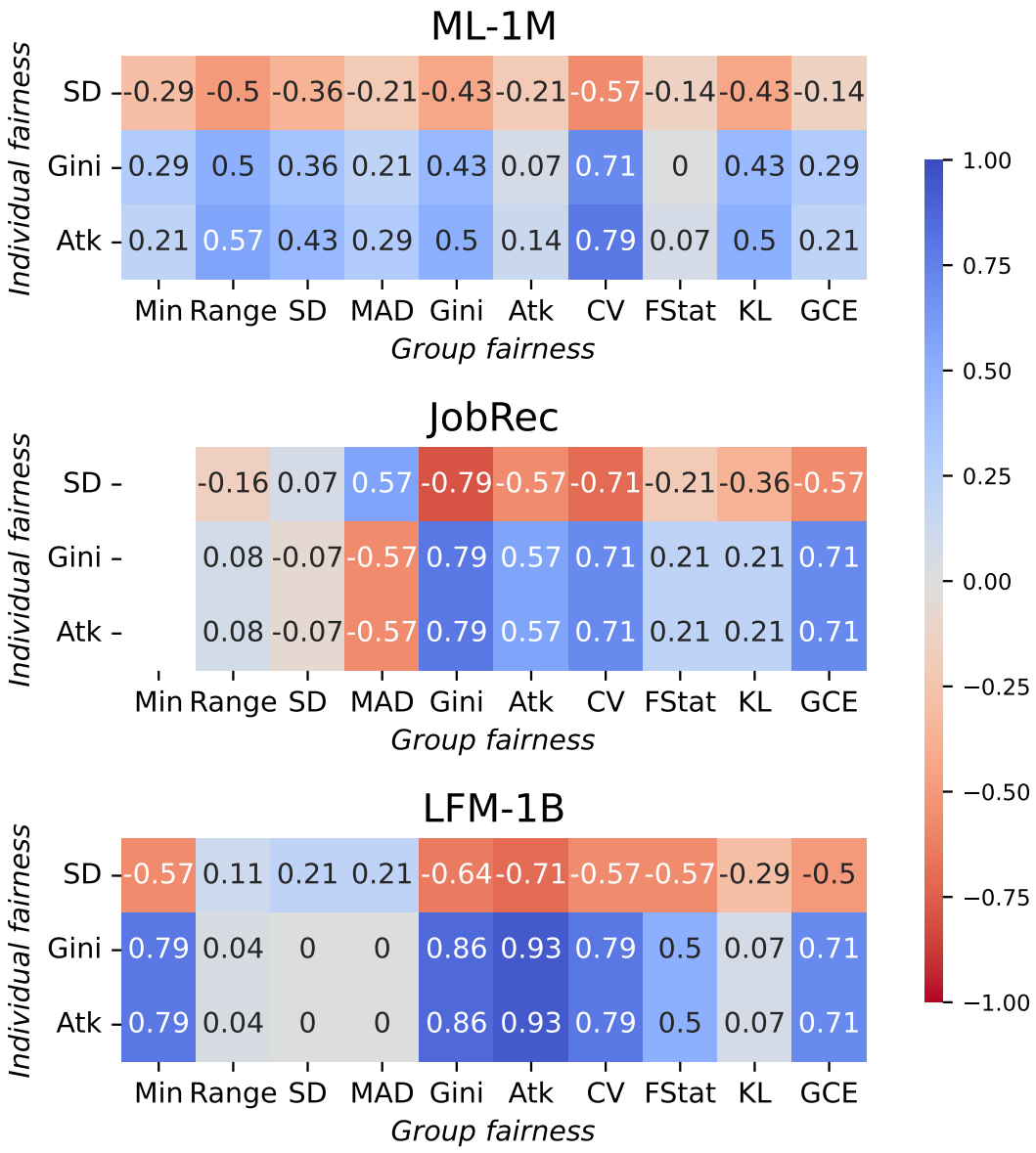


Figure 7.1: Agreement (Kendall's τ) of NDCG-based measures for individual Fair (y-axis) and group Fair (x-axis) in ranking LLMRecs. User groups are based on 3 sensitive attributes. Due to 8-way ties, τ cannot be computed for Min (JobRec).

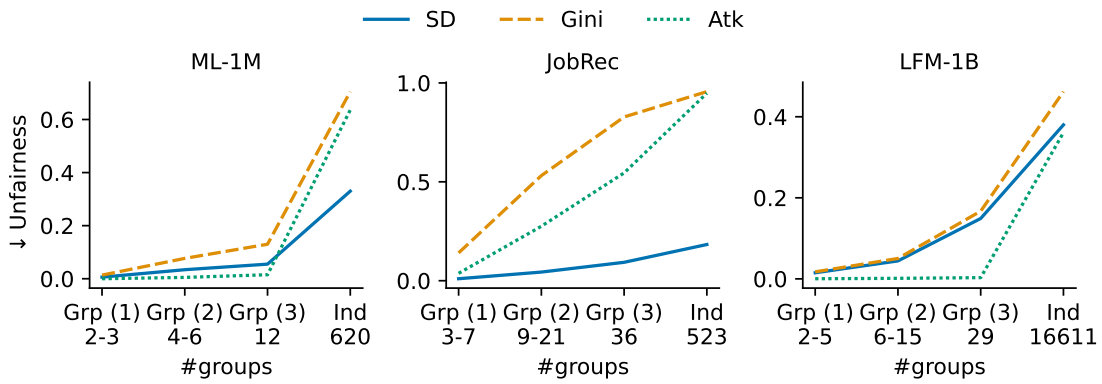


Figure 7.2: NDCG-based Group (Grp) and individual (Ind) Fair scores of GLM-4-9B (NS-prompt). Grp (a) is the mean Fair scores across all ways of grouping users when considering a attribute(s). For $a \in \{1, 2\}$, the # of ways to group users is three each. For $a = 3$, there is only one way to group users.

measures. We evaluate the runs from GLM-4-9B (NS) due to their relatively good NDCG and Fair scores across all datasets. The results are shown in Fig. 7.2.[9]

Regarding (i), our results verify that **fairness worsens as more attributes are used to form groups, highlighting the importance of considering intersectionality in fairness evaluation.** Note that JobRec has higher ↓Gini/↓Atk than ML-1M/LFM-1B as its NDCG is much lower; Gini/Atk accounts for this, but SD does not.

Regarding (ii), we find that **even when the recommendations are relatively fair for (intersectional) groups, they can be much less fair for individuals** (i.e., for ML-1M and LFM-1B). For all measures and datasets, the individual Fair scores are always worse than group Fair scores. These findings imply that group Fair scores may mask unfairness within groups, even if the within-group variation is considered (e.g., in $\text{Atk}_{\text{b-group}}$).

Fairness decomposability

When fairness is evaluated only between groups, within-group unfairness may occur undetected. As the number of groups increases, we analyse: (i) how between- and within- group fairness change; and (ii) how they relate to individual fairness. To this

[9]While we present the measures in the same plot, their distribution differs, so the figures should not be used to quantify the gap between two families of measures.

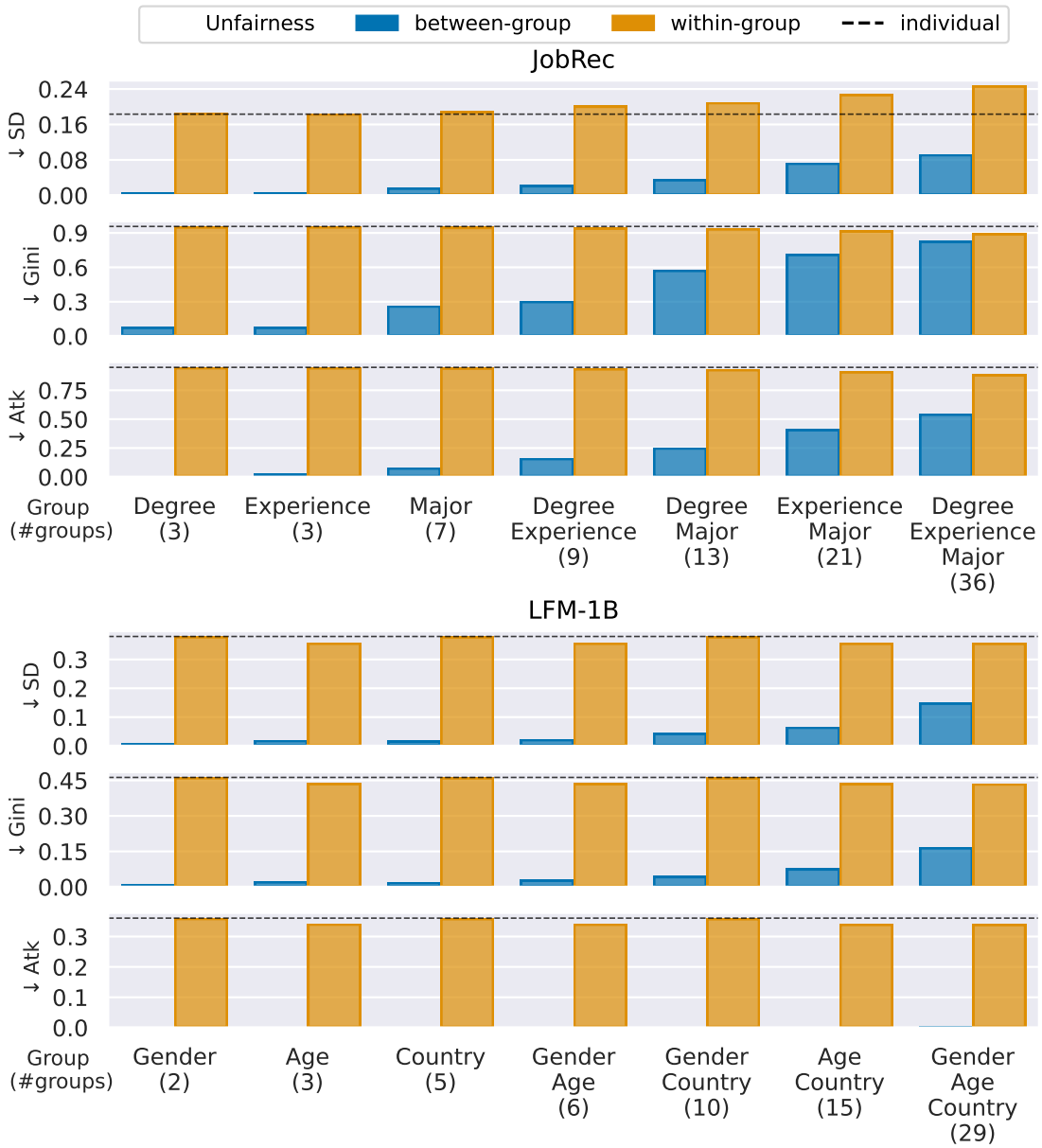


Figure 7.3: NDCG-based individual, between- and within-group unfairness of GLM-4-9B (NS-prompt) for all ways of grouping users in JobRec and LFM-1B.

end, we compute between- and within-group fairness, as well as individual fairness with SD, Gini, and Atk on the NDCG scores from GLM-4-9B (NS). Results for JobRec and LFM-1B are shown in Fig. 7.3; we find similar trends for ML-1M (see §7.5.2.3).

Regarding (i), we find that while between-group fairness generally worsens as the number of groups grows,[10] within-group fairness tends to remain stable, except for SD of JobRec, where it worsens. Regarding (ii), we see that within-group unfairness is almost as high as individual unfairness, and always higher than between-group unfairness. Sometimes, within-group unfairness even exceeds individual unfairness (e.g., SD for JobRec). In summary, **a single RS can vary in group fairness for different ways of grouping users, while being highly unfair for users within groups (regardless of the grouping) and for individual users at the same time.**

7.4 Conclusion

We empirically studied the relationship between group and individual fairness in RSs. We found that RSs which are fair for groups can still be very unfair for individual users, providing the first empirical evidence on the disjointness of these two RS fairness concepts. Overall, we encourage evaluating individual and within-group fairness alongside group fairness, as between-group Fair scores may mask huge disparity in RS effectiveness across users, even for a large number of groups (i.e., 5% of the number of users) and even when the between-group Fair score accounts for within-group variations. Future work will study how strategies that mitigate either group or individual fairness affect each fairness type.

7.5 Appendix

7.5.1 Details on Experimental Setup

7.5.1.1 User grouping

For JobRec, we explain in more detail how we group users based on the study major. The study major is in free text format, so we first performed exploratory data analysis to get a better overview of the its distribution. More than 80% test users with a high school (HS) degree have no specific study major, so we do not use major to divide

[10]Resembling stairs; hence, the title of the paper.

users with a HS degree into smaller groups. Further, an author manually grouped the 235 unique majors of the non-HS test users into six fields of study, taking inspiration from the grouping in Xu et al. [240]. During this process, we also remove users with generic or erroneous majors. We group the train/val users' majors as for the test users when possible, otherwise we perform fuzzy string matching between the annotated (manually grouped) and unannotated majors with `rapidfuzz` [12]. We map the unannotated major to the same group as the most similar annotated major, if similarity ≥ 0.75, otherwise we map it to an 'Others' category.

We summarise the statistics of user sensitive attributes (after preprocessing) in Tab. 7.3. We present the size of intersectional groups for each dataset in Tabs. 7.4–7.6.

7.5.1.2 Prompt templates and examples

We provide the templates and examples for the non-sensitive (NS) prompts in Tab. 7.7 and for the sensitive (S) prompts in Tab. 7.8. In the S-prompt, we replace "a user" in the NS-prompt with a description containing their sensitive attribute, e.g., "a user with a bachelor's degree, majoring in biology, has a total of 10 years of experience'. We use the original user attribute prior to re-grouping as they are more informative and precise (e.g., the user's major is included as 'biology', not 'STEM'). While this may not be the most natural way a user may interact with LLMs (e.g., explicitly disclosing their gender), it is a way to ensure that the LLMs have information on the users' sensitive attribute.

7.5.1.3 Measure formulations and technical details

To enhance reproducibility, we provide the formulations of all Fair measures that we compute. A Fair measure can be used to compute individual fairness, (between-)group fairness, and/or within-group fairness. We denote the individual, between-group, and within-group fairness versions of a measure as $\cdot_{\text{ind}}$, $\cdot_{\text{b-group}}$, and $\cdot_{\text{w-group}}$, respectively. When applicable, we provide the technical details (e.g., parameters) that we used. $\uparrow/\downarrow$ means the higher/lower the better

Average scores of the worst 25% groups (Min). $\uparrow$Min [229] measures fairness between groups. To compute $\text{Min}_{\text{b-group}}$, the Eff score is first calculated per user, and then averaged per user group. The group mean Eff score is then sorted, and the first quartile value is set as a threshold. $\text{Min}_{\text{b-group}}$ averages the mean group Eff score of groups with an Eff score below/equal to the threshold. The range of Min follows the Eff score range.

Range. ↓Range can be used to quantify fairness between groups. It is computed as the range of mean group Eff scores [142].

Standard Deviation (SD). ↓SD[11] can be used to measure between-group fairness [142, 255], within-group fairness, and individual fairness [183]. $\text{SD}_{\text{b-group}}$ is the (population) standard deviation of the group mean Eff scores and SD_{ind} is the (population) standard deviation of users' Eff scores. We compute $\text{SD}_{\text{w-group}}$ as follows:

$$SD_{\text{w-group}} = \sum_{j=1}^{N'} s_j \cdot SD(g_j) \tag{7.1}$$

where g_j is the list of group j users' effectiveness scores and N' is the number of user groups. The range of SD depends on the Eff score range. The weight of group j is s_j, the ratio of the group's total Eff scores to the total across all groups.

Mean Absolute Difference (MAD). ↓MAD [77] measures between-group fairness as the mean absolute pairwise difference of the groups' mean Eff scores. It is computed follows:

$$MAD = \frac{1}{N'(N'-1)} \sum_{j=1}^{N'-1} \sum_{j'=j+1}^{N'} |\bar{g_j} - \bar{g_{j'}}| \tag{7.2}$$

where $\bar{g_j}$ is the group j's mean Eff scores. The range of MAD follows the Eff score range.

Gini Index (Gini). ↓Gini [85] is a general-purpose inequality measure that has been used to measure between-group fairness [84], within-group fairness [75, 166], and individual fairness [126]. Given N non-negative values, $x_1, x_2, \ldots, x_N$, the general-purpose Gini is computed as follows:

$$Gini = \frac{\sum_{j=1}^{N} (2j - N - 1)x_j}{N \sum_{j=1}^{N} x_j} = \frac{\sum_{j=1}^{N} (2j - N - 1)x_j}{N^2 \bar{x}_j} \tag{7.3}$$

[11] We also include prior work that computes Variance under the same family of measure, as Variance is the square of SD.

$\text{Gini}_{\text{b-group}}$ is the Gini of the group mean EFF scores and Gini_{ind} is the Gini of users' EFF scores. We compute $\text{Gini}_{\text{w-group}}$ as follows:

$$Gini_{\text{w-group}} = \sum_{j=1}^{N'} s_j \cdot Gini(g_j) \tag{7.4}$$

The range of Gini is [0,1].

Coefficient of Variation (CV). ↓CV is a general-purpose statistical measure of data dispersion, which has been used to measure between-group fairness [229, 266]. The general-purpose CV is computed as the ratio of the (population) standard deviation to the mean. We compute $\text{CV}_{\text{b-group}}$ as follows:

$$CV_{\text{b-group}} = \frac{SD(\bar{g_1}, \bar{g_2}, \ldots, \bar{g_{N'}})}{mean(\bar{g_1}, \bar{g_2}, \ldots, \bar{g_{N'}})} \tag{7.5}$$

The range of CV is $[0, \infty)$.

F-statistic (FStat). ↓FStat is originally a statistical test that measures the ratio of between-group variance to within-group variance. It has been used to quantify between-group fairness for N' groups and N' users [224], as follows:

$$V = \frac{1}{N} \sum_{j=1}^{N'} |g_j| \cdot (\bar{g_j} - \bar{x}_{\text{eff}})^2 \tag{7.6}$$

$$U = \frac{1}{N} \sum_{j=1}^{N'} \sum_{\ell=1}^{|g_j|} (x_{j,\ell} - \bar{g_j})^2 \tag{7.7}$$

$$FStat_{\text{b-group}} = \frac{V/(N'-1)}{U/(N-N')} \tag{7.8}$$

where $|g_j|$ is the number of users in group j, $\bar{x}_{\text{eff}}$ is the mean EFF scores of all users, and $x_{j,\ell}$ is the EFF score of the ℓ-th user in the j-th group. The range of FStat is $[0, \infty)$.

Kullback-Leibler Divergence (KL). ↓KL can be customised to measure various fairness criteria; we select the version where between-group fairness means that effectiveness should be proportional to group size, as it is a popular choice [9]. KL is computed as follows:

$$KL_{\text{b-group}}(p||q) = \sum_{j=1}^{N'} p_j \log_2 \frac{p_j}{q_j} \tag{7.9}$$

$$p_j = \frac{\bar{g}_j}{\sum_{j=1}^{N'} \bar{g}_j} \tag{7.10}$$

$$q_j = \frac{|g_j|}{\sum_{j=1}^{N'} |g_j|} \tag{7.11}$$

The range of KL is $[0, \infty)$.

Generalised Cross Entropy (GCE). ↓ GCE can be used to quantify various fairness criteria [52, 53]; we choose the version that requires equal utility (effectiveness) across groups. We compute GCE as follows:[12]

$$GCE_{\text{b-group}} = -\frac{1}{B(1-B)} \sum_{j=1}^{N'} p_{ref}^{B} \cdot \left(\frac{\hat{p_j}}{\sum_{j=1}^{N'} \hat{p_j}} \right)^{(1-B)} - 1 \tag{7.12}$$

$$\hat{p_j} = \lambda p_j + (1-\lambda)c \tag{7.13}$$

where B is a parameter, $p_{ref} = 1/N'$, p_j is as per Eq. (7.10) and $\hat{p_j}$ is the smoothened value of p_j. We set $\beta = 2$, and perform smoothing with $\lambda = 0.95$ and $c = 10^{-4}$. The range of GCE is $[0, \infty)$.

Atkinson Index. We adopt Atkinson Index (Atk), a measure of income inequality from the economics domain [10], to quantify group/individual fairness for RS users based on recommendation effectiveness disparity.

General formulation. Given N non-negative values, $x_1, x_2, \ldots, x_N$, the general-purpose Atk is defined as follows:[13]

$$Atk(x_1, x_2, \ldots, x_N) = 1 - \frac{1}{\bar{x}} f(x_1, x_2, \ldots, x_N) \tag{7.14}$$

[12]Note that this is the nonnegative version of the measure in [53].

[13]Atk with $\varepsilon = 1$ is defined differently from Eq. (7.14).

$$f(x_1, x_2, \ldots, x_N) = \left(\sum_{j=1}^{N} \frac{w_j}{\sum_{j=1}^{N} w_j} x_j^{1-\varepsilon} \right)^{\frac{1}{1-\varepsilon}} \tag{7.15}$$

where $\bar{x}$ is the mean of the values, w_j is the weight of the j-th score, and $\varepsilon \geq 0$ is the inequality aversion parameter.[14] Atk with a low ε is more sensitive to disparity among higher scores than among lower scores; we compute all Atk with $\varepsilon = 0.5$ [48]. Atk has a [0,1]-range.

Measuring user fairness with Atk. To measure individual fairness for N users with Atk (referred to as Atk_{ind}), we input the effectiveness score (e.g., NDCG) per user to Eq. (7.14) and set $w_j = 1, \forall j \in \{1, \ldots, N\}$ to weigh each user equally. As a user's effectiveness score could be 0, we choose $\varepsilon < 1$ to avoid division by zero in $x_j^{1-\varepsilon}$ (Eq. (7.15)). $\text{Atk}_{\text{b-group}}$ is computed as follows:

$$Atk_{\text{b-group}} = Atk\left(f(g_1), f(g_2), \ldots, f(g_{N'})\right) \tag{7.16}$$

Essentially, $\text{Atk}_{\text{b-group}}$ measures between-group fairness while accounting for the disparity within each group via f. As f may be unstable for small groups, we reduce the score contributions from small groups by using the group size as w_j to weigh $f(g_j)$. To compute $f(g_j)$ itself, we weigh each user in the group equally.

Decomposition of Atk. Atk is subgroup-decomposable [203], i.e., it can be broken down into between- and within-group components without residual terms [24], establishing a coherent relationship between individual fairness and group fairness [68]. Specifically, it detects if individual unfairness occurs mainly due to between- or within-group effectiveness disparity. Atk_{ind} can be decomposed as follows [21, 24, 47]:

$$\underbrace{1 - Atk_{\text{ind}}}_{\text{individual fairness}} = \underbrace{(1 - Atk_{\text{b-group}})}_{\text{between-group fairness}} \cdot \underbrace{(1 - Atk_{\text{w-group}})}_{\text{within-group fairness}} \tag{7.17}$$

$$Atk_{\text{w-group}} = \sum_{j=1}^{N'} s_j \cdot Atk(g_j) \tag{7.18}$$

7.5.2 Extended Results

7.5.2.1 Evaluation of all LLMRecs

We present Eff and P-based Fair scores for LLMRecs in Tab. 7.9.

[14] If $x_j = 0, \forall j \in \{1, 2, \ldots, N\}$, we define $\downarrow$Atk $= 0$.

7.5.2.2 Agreement of group and individual fairness measures

We show in Fig. 7.4 the agreement of group-individual fairness between the same family of measures (e.g., $SD_{\text{b-group}}$ vs SD_{ind}) for all possible groupings.

7.5.2.3 Fairness decomposability

We show in Fig. 7.5 the individual, between-group, and within-group unfairness of GLM-4-9B (NS-prompt) for all ways of grouping users in ML-1M.

Table 7.3: Statistics of user sensitive attributes after preprocessing.

	ML-1M [94]	JobRec [93]	LFM-1B [196]
sensitive attr. #1	gender (2): M (457); F (163)	degree (3): University (254); High School (157); College (112)	gender (2): M (12,826); F (3,785)
sensitive attr. #2	age (3): 25–49 years (437); 18–24 years (110); ≥50 years (73)	years of experience (3): >10 (219); >5–10 (200); ≤5 (102)	age (3): 18–24 years (8,945); 25–49 years (6,590); ≥50 years (200)
sensitive attr. #3	occupation (2): working (478); non-working (142)	major (6): Business, Management, Finance (154); Social Science, Humanities, Education (103); STEM (41); Health & Medical (29); Arts, Creative, Entertainment (22); Others (17);	country/continent (5): Europe (9,976); America & Antarctica (5,337); Asia (863); Oceania (376); Africa (59)

Table 7.4: Size of intersectional groups (ML-1M).

gender	age	occupation	#user
M	18–24 years	non-working	44
		working	37
	25–49 years	non-working	41
		working	279
	≥50 years	non-working	8
		working	48
F	18–24 years	non-working	21
		working	8
	25–49 years	non-working	26
		working	91
	≥50 years	non-working	2
		working	15

Table 7.5: Size of intersectional groups (JobRec). Combinations resulting in group size of 0 are excluded.

degree	experience (years)	major	#user
High School	≤5	-	31
	>5–10	-	54
	>10	-	70
College	≤5	Business, Management, Finance	7
		Social Science, Humanities, Education	3
		STEM	4
		Health & Medical	8
		Arts, Creative, Entertainment	2
		Others	2
	>5–10	Business, Management, Finance	13
		Social Science, Humanities, Education	6
		STEM	8
		Health & Medical	10
		Arts, Creative, Entertainment	3
		Others	4
	>10	Business, Management, Finance	14
		Social Science, Humanities, Education	12
		STEM	3
		Health & Medical	3
		Arts, Creative, Entertainment	4
		Others	6
University	≤5	Business, Management, Finance	18
		Social Science, Humanities, Education	16
		STEM	7
		Health & Medical	3
		Arts, Creative, Entertainment	1
	>5–10	Business, Management, Finance	47
		Social Science, Humanities, Education	38
		STEM	9
		Arts, Creative, Entertainment	8
	>10	Business, Management, Finance	55
		Social Science, Humanities, Education	28
		STEM	10
		Health & Medical	5
		Arts, Creative, Entertainment	4
		Others	5

Table 7.6: Size of intersectional groups (LFM-1B). Combinations resulting in group size of 0 are excluded.

gender	age	continent	#user
F	18–24 years	Europe	1439
		America & Antarctica	809
		Asia	127
		Oceania	44
		Africa	7
	25–49 years	Europe	551
		America & Antarctica	392
		Asia	59
		Oceania	19
		Africa	1
	≥50 years	Europe	15
		America & Antarctica	11
		Asia	2
		Africa	1
M	18–24 years	Europe	3919
		America & Antarctica	2068
		Asia	339
		Oceania	168
		Africa	25
	25–49 years	Europe	3418
		America & Antarctica	1723
		Asia	276
		Oceania	128
		Africa	23
	≥50 years	Europe	97
		America & Antarctica	59
		Asia	11
		Oceania	3
		Africa	1

Table 7.7: Non-sensitive (NS) prompt templates and examples.

Dataset: ML-1M
Prompt template: "You are a movie recommender. If a user has watched the following movies, listed chronologically from earliest to latest, and rated them positively: *<items in train>*, then you should recommend the following movies: *<items in val>*. Now that the user has watched the recommended movies, you should recommend 10 other movies from the year between 1919 and 2000 (inclusive) that the user is most likely to watch next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not re-recommend movies that have been watched by the user. Do not output other words." **Prompt example:** "You are a movie recommender. If a user has watched the following movies, listed chronologically from earliest to latest, and rated them positively: American Beauty, Magnolia, Boogie Nights, The Tao of Steve, Dark City, Fight Club, then you should recommend the following movies: The Evening Star, The MatchMaker, West Side Story, Dirty Dancing, The Prince of Egypt, Grease, Fargo, Good Will Hunting, Playing by Heart, The Man Without a Face. Now that the user has watched the recommended movies, you should recommend 10 other movies from the year between 1919 and 2000 (inclusive) that the user is most likely to watch next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not re-recommend movies that have been watched by the user. Do not output other words."
Dataset: JobRec
Prompt template: "You are a job title recommender. If a user has applied to job positions with the following job titles, listed chronologically from earliest to latest: *<items in train>*, you should recommend the following job titles: *<items in val>*. Now that the user has applied to positions with those job titles, you should recommend 10 other job titles that the user is most likely to apply for next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not output other words." **Prompt example:** "You are a job title recommender. If a user has applied to job positions with the following job titles, listed chronologically from earliest to latest: Billing/Collections Process Coordinator, Administrative Assistant, Office Manager/Bookkeeper, Bookkeeper, you should recommend the following job titles: Accounts Payable Clerk, Assistant Director of Human Resources. Now that the user has applied to positions with those job titles, you should recommend 10 other job titles that the user is most likely to apply for next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not output other words."
Dataset: LFM-1B
Prompt template: "You are a music artist recommender. If a user has listened to the following artists, listed chronologically from earliest to latest: *<items in train>* and considering that the user listened to the artist *<playcount of items in train>* times respectively, then you should recommend the following artists: *<items in val>*. Now that the user has listened to the recommended artists, you should recommend 10 other artists, who has released songs prior to 2017, that the user is most likely to listen to next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not re-recommend artists that the user has listened to. Do not output other words." **Prompt example:** "You are a music artist recommender. If a user has listened to the following artists, listed chronologically from earliest to latest: Lee Van Dowski, The Oohlas, XTC, Brenmar, Mountain, Bladerunner, Efdemin, Tom Encore, Droid Sector, Xhin, and considering that the user listened to the artist 2, 1, 2, 4, 12, 1, 12, 1, 2, 3 times respectively, then you should recommend the following artists: VNV Nation, Space, Ja Rule, Nneka, Parks, Enzyme X, David Bisbal, Primer 55, A-Sides, Cydonia. Now that the user has listened to the recommended artists, you should recommend 10 other artists, who has released songs prior to 2017, that the user is most likely to listen to next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not re-recommend artists that the user has listened to. Do not output other words."

Table 7.8: Sensitive (S) prompt templates and examples.

Dataset: ML-1M
Prompt template: "You are a movie recommender. If a *<gender>* user whose age is *<age range>* years old and whose occupation is *<occupation>* has watched the following movies, listed chronologically from earliest to latest, and rated them positively: *<items in train>*, then you should recommend the following movies: *<items in val>*. Now that the user has watched the recommended movies, you should recommend 10 other movies from the year between 1919 and 2000 (inclusive) that the user is most likely to watch next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not re-recommend movies that have been watched by the user. Do not output other words." **Prompt example:** "You are a movie recommender. If a female user whose age is 18-24 years old and whose occupation is writer has watched the following movies, listed chronologically from earliest to latest, and rated them positively: American Beauty, Magnolia, Boogie Nights, The Tao of Steve, Dark City, Fight Club, then you should recommend the following movies: The Evening Star, The MatchMaker, West Side Story, Dirty Dancing, The Prince of Egypt, Grease, Fargo, Good Will Hunting, Playing by Heart, The Man Without a Face. Now that the user has watched the recommended movies, you should recommend 10 other movies from the year between 1919 and 2000 (inclusive) that the user is most likely to watch next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not re-recommend movies that have been watched by the user. Do not output other words."
Dataset: JobRec
Prompt template: "You are a job title recommender. If a user with a *<degree type>* degree, majoring in *<major>*, has a total of *<years of experience>* years of experience has applied to job positions with the following job titles, listed chronologically from earliest to latest: *<items in train>*, then you should recommend the following job titles: *<items in val>*. Now that the user has applied to positions with those job titles, you should recommend 10 other job titles that the user is most likely to apply for next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not output other words." **Prompt example:** "You are a job title recommender. If a user with a Bachelor's degree, majoring in Accounting and Information Systems, has a total of 24 years of experience has applied to job positions with the following job titles, listed chronologically from earliest to latest: Billing/Collections Process Coordinator, Administrative Assistant, Office Manager/Bookkeeper, Bookkeeper, then you should recommend the following job titles: Accounts Payable Clerk, Assistant Director of Human Resources. Now that the user has applied to positions with those job titles, you should recommend 10 other job titles that the user is most likely to apply for next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not output other words."
Dataset: LFM-1B
Prompt template: 'You are a music artist recommender. If a *<gender>* user whose age is *<age>* years old and lives in *<country>* has listened to the following artists, listed chronologically from earliest to latest: *<items in train>*, and considering that the user listened to the artist *<playcount of items in train>* times respectively, then you should recommend the following artists: *<items in val>*. Now that the user has listened to the recommended artists, you should recommend 10 other artists, who has released songs prior to 2017, that the user is most likely to listen to next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not re-recommend artists that the user has listened to. Do not output other words." **Prompt example:** "You are a music artist recommender. If a male user whose age is 30 years old and lives in Russia has listened to the following artists, listed chronologically from earliest to latest: Lee Van Dowski, The Oohlas, XTC, Brenmar, Mountain, Bladerunner, Efdemin, Tom Encore, Droid Sector, Xhin, and considering that the user listened to the artist 2, 1, 2, 4, 12, 1, 12, 1, 2, 3 times respectively, then you should recommend the following artists: VNV Nation, Space, Ja Rule, Nneka, Parks, Enzyme X, David Bisbal, Primer 55, A-Sides, Cydonia. Now that the user has listened to the recommended artists, you should recommend 10 other artists, who has released songs prior to 2017, that the user is most likely to listen to next. You should order them by probability and compact them in one line split by commas. Do not output the probability. Do not re-recommend artists that the user has listened to. Do not output other words."

Table 7.9: Effectiveness (Eff) and fairness (Fair) scores at $k = 10$ for intersectional groups (Grp.) and individuals (Ind.) of LLMRecs with non-sensitive (NS) and sensitive (S) prompts. All Fair scores are computed with P. All measures range in [0,1], except the Grp. measures below them. The best Eff/Fair scores are bolded. Darker green marks scores closer to the best Eff/Fair per measure. ↑ / ↓ means the higher/lower the better.

	LLMRec	GLM-4-9B		Llama-3.1-8B		Ministral-8B		Qwen2.5-7B	
	prompt type	NS	S	NS	S	NS	S	NS	S
	ML-1M								
Eff	↑ HR	**0.377**	0.358	0.260	0.269	0.342	0.340	0.363	0.371
	↑ MRR	**0.189**	0.174	0.101	0.113	0.159	0.173	0.165	0.180
	↑ P	**0.082**	0.074	0.046	0.051	0.067	0.067	0.074	0.080
Fair (Grp.)	↑ Min	0.036	**0.049**	0.030	0.034	0.022	0.040	0.024	0.035
	↓ Range	**0.075**	0.088	**0.075**	0.088	0.089	0.125	0.175	0.125
	↓ SD	0.024	0.023	**0.019**	0.020	0.027	0.028	0.044	0.033
	↓ MAD	0.028	0.028	**0.022**	0.023	0.031	0.031	0.051	0.040
	↓ Gini	0.169	**0.166**	0.201	0.188	0.229	0.196	0.301	0.228
	↓ Atk	**0.018**	0.019	0.031	0.021	0.032	0.037	0.057	0.041
	↓ CV	0.323	**0.299**	0.388	0.358	0.444	0.391	0.577	0.416
	↓ FStat	0.504	0.450	0.472	**0.369**	0.625	0.580	1.586	0.973
	↓ KL	0.836	1.211	1.328	1.349	**0.793**	1.292	1.380	1.150
	↓ GCE	0.106	**0.051**	0.068	0.140	659.834	659.704	0.387	0.273
Fair (Ind.)	↓ SD	0.141	0.129	**0.096**	0.109	0.124	0.123	0.136	0.141
	↓ Gini	**0.752**	0.760	0.818	0.819	0.777	0.774	0.764	0.761
	↓ Atk	**0.658**	0.674	0.760	0.755	0.693	0.692	0.675	0.667
	JobRec								
	LLMRec	GLM-4-9B		Llama-3.1-8B		Ministral-8B		Qwen2.5-7B	
	prompt type	NS	S	NS	S	NS	S	NS	S
Eff	↑ HR	0.054	0.033	0.044	0.021	0.057	0.044	**0.065**	0.057
	↑ MRR	0.037	0.023	0.017	0.009	**0.048**	0.033	0.046	0.045
	↑ P	0.006	0.003	0.005	0.002	0.006	0.004	**0.007**	**0.007**
Fair (Grp.)	↑ Min	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
	↓ Range	0.050	**0.010**	0.050	0.050	0.033	0.021	0.033	0.033
	↓ SD	0.010	**0.003**	0.010	0.008	0.008	0.005	0.009	0.008
	↓ MAD	0.008	**0.003**	0.008	0.004	0.007	0.005	0.009	0.007
	↓ Gini	0.800	0.800	0.827	0.911	0.742	0.791	**0.725**	0.769
	↓ Atk	0.509	0.480	0.641	0.758	0.498	0.508	0.466	**0.455**
	↓ CV	2.079	1.875	2.262	3.634	1.629	1.850	**1.508**	1.770
	↓ FStat	0.848	**0.447**	1.193	1.007	0.871	0.696	0.927	0.642
	↓ KL	2.706	**1.330**	3.195	4.677	1.716	1.818	2.004	1.926
	↓ GCE	1685.893	1979.080	1685.948	2052.428	1612.569	1759.183	**1539.280**	1759.168
Fair (Ind.)	↓ SD	0.024	0.018	0.022	**0.014**	0.026	0.021	0.028	0.028
	↓ Gini	0.948	0.967	0.958	0.979	0.946	0.956	**0.940**	0.948
	↓ Atk	0.947	0.967	0.956	0.979	0.943	0.956	**0.936**	0.944
	LFM-1B								
	LLMRec	GLM-4-9B		Llama-3.1-8B		Ministral-8B		Qwen2.5-7B	
	prompt type	NS	S	NS	S	NS	S	NS	S
Eff	↑ HR	0.658	**0.661**	0.609	0.618	0.451	0.317	0.375	0.317
	↑ MRR	**0.409**	0.408	0.347	0.357	0.266	0.174	0.199	0.160
	↑ P	0.203	**0.206**	0.172	0.181	0.118	0.080	0.106	0.085
Fair (Grp.)	↑ Min	0.093	**0.098**	0.097	0.092	0.037	0.019	0.038	0.018
	↓ Range	0.313	0.300	0.500	0.600	0.600	0.200	0.200	**0.136**
	↓ SD	0.076	0.068	0.089	0.095	0.102	0.042	0.048	**0.040**
	↓ MAD	0.081	0.075	0.082	0.081	0.079	**0.045**	0.051	0.046
	↓ Gini	0.212	**0.197**	0.229	0.234	0.339	0.316	0.251	0.313
	↓ Atk	**0.004**	**0.004**	0.005	0.006	0.008	0.016	0.008	0.023
	↓ CV	0.414	**0.372**	0.513	0.566	0.903	0.600	0.484	0.564
	↓ FStat	3.344	3.108	3.679	4.152	2.887	3.444	**1.895**	4.933
	↓ KL	2.735	2.716	3.501	3.235	3.513	2.729	2.734	**2.194**
	↓ GCE	338.856	225.920	**113.018**	113.024	451.857	451.864	451.815	564.777
Fair (Ind.)	↓ SD	0.226	0.228	0.204	0.210	0.183	**0.157**	0.187	0.170
	↓ Gini	0.586	**0.583**	0.610	0.603	0.718	0.803	0.769	0.807
	↓ Atk	0.415	**0.411**	0.454	0.445	0.599	0.719	0.669	0.722

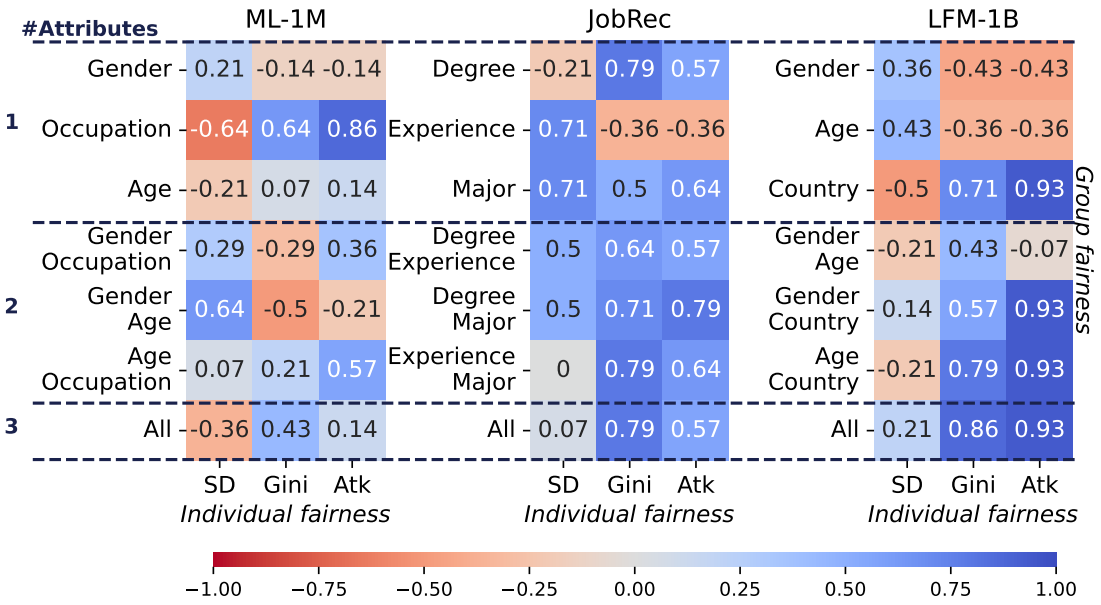


Figure 7.4: Agreement (Kendall's τ) between the same family of measure in ranking LLMRecs for NDCG-based group fairness (y-axis) and individual fairness (x-axis). Group fairness is computed for each combination of users' sensitive attributes.

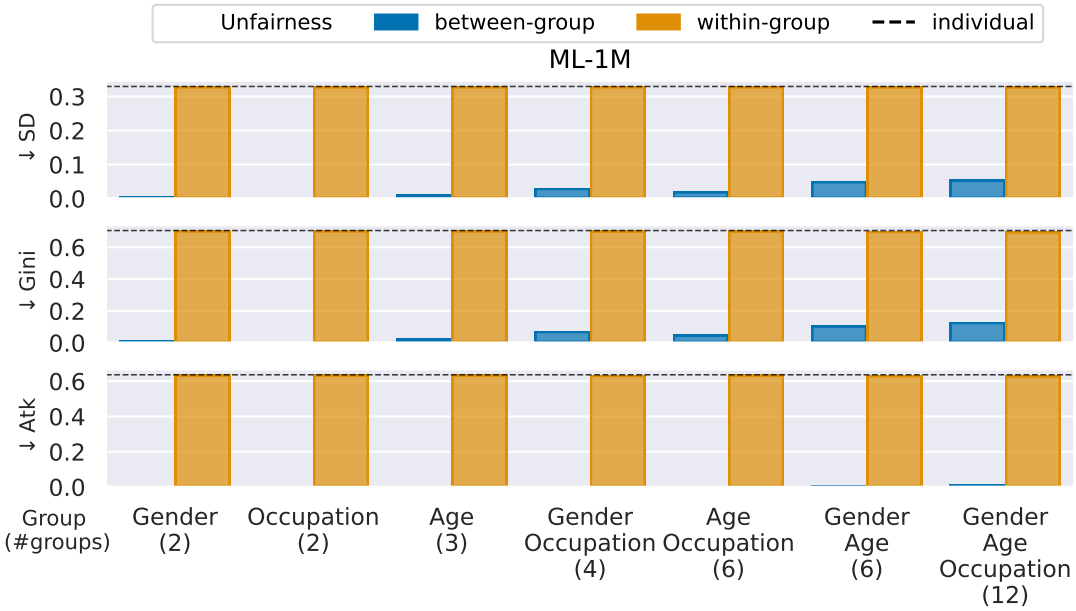


Figure 7.5: NDCG-based individual, between- and within-group unfairness of GLM-4-9B (NS-prompt) for all ways of grouping users in ML-1M.

Errata

A summary of corrections for non-substantive errors that do not change the conclusions of the papers are provided below.

Summary of corrections to non-substantive errors.

	Original Text	Correction	Explanation
Chapter 3	$e_{\text{inv}}(u, i) = 1/z(u, i)$	$e_{\text{inv}}(p) = 1/p$	The correct formulation of the inverse examination function depends directly on the position p of the recommendation slot.
Chapter 3	We use NDCG and Jain as they are more sensitive to changes than HR and QF.	We use NDCG and Jain as they tend to be stricter than some other single-aspect measures used in this work.	The text has been changed for clarity.
Chapter 3	IAA/HD/II-F strongly disagrees with IFD, $\tau \in [-0.71, -0.5]$	IAA/HD/II-F strongly disagrees with IFD, $\tau \in [-0.93, -0.5]$	The lower bound of τ has been corrected. This does not change the conclusion.
Chapter 5	... averaging fails to reach the same conclusion as DPFR almost half the time	... averaging fails to reach the same conclusion as DPFR almost half the time for the rank-based Rel measures	The phrase "for the rank-based Rel measures" has been added for clarity.
Chapter 5	DPFR demonstrates distinct benefits in mitigating false conclusions by up to 50%...	DPFR demonstrates distinct benefits in mitigating false conclusions by up to 58%...	The performance of DPFR (our proposed approach) is better than previously reported.

Bibliography

[1] Syed Wajid Aalam, Abdul Basit Ahanger, Muzafar Rasool Bhat, and Assif Assad. Evaluation of Fairness in Recommender Systems: A Review. In Balas Valentina E., G R. Sinha, Agarwal Basant, Sharma Tarun Kumar, Dadheech Pankaj, and Mahrishi Mehul, editors, *Emerging Technologies in Computer Engineering: Cognitive Computing and Intelligent IoT*, pages 456–465, Cham, 2022. Springer International Publishing. ISBN 978-3-031-07012-9.

[2] Panagiotis Adamopoulos and Alexander Tuzhilin. On over-specialization and concentration bias of recommendations: probabilistic neighborhood selection in collaborative filtering systems. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 153–160, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450326681. doi: 10.1145/2645710.2645752. URL `https://doi.org/10.1145/2645710.2645752`.

[3] Gediminas Adomavicius, Bamshad Mobasher, Francesco Ricci, and Alexander Tuzhilin. Context-aware recommender systems. *AI Magazine*, 32(3):67–80, Oct. 2011. doi: 10.1609/aimag.v32i3.2364. URL `https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2364`.

[4] Charu C. Aggarwal. *Recommender Systems: The Textbook*. Springer Publishing Company, Incorporated, 1st edition, 2016. ISBN 3319296574.

[5] Ameer Albahem, Damiano Spina, Falk Scholer, Alistair Moffat, and Lawrence Cavedon. Desirable properties for diversity and truncated effectiveness metrics. In *Proceedings of the 23rd Australasian Document Computing Symposium*, ADCS '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450365499. doi: 10.1145/3291992.3291996. URL `https://doi.org/10.1145/3291992.3291996`.

[6] Bushra Alhijawi, Arafat Awajan, and Salam Fraihat. Survey on the objectives of recommender systems: Measures, solutions, evaluation methodology, and

new perspectives. *ACM Comput. Surv.*, 55(5), December 2022. ISSN 0360-0300. doi: 10.1145/3527449. URL https://doi.org/10.1145/3527449.

[7] Paul D Allison. Measures of Inequality. *American Sociological Review*, 43(6): 865–880, 12 1978.

[8] Enrique Amigó, Damiano Spina, and Jorge Carrillo-de Albornoz. An axiomatic analysis of diversity evaluation metrics: Introducing the rank-biased utility metric. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 625–634, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210024. URL https://doi.org/10.1145/3209978.3210024.

[9] Enrique Amigó, Yashar Deldjoo, Stefano Mizzaro, and Alejandro Bellogín. A unifying and general account of fairness measurement in recommender systems. *Information Processing & Management*, 60(1):103115, 1 2023. ISSN 0306-4573. doi: 10.1016/J.IPM.2022.103115.

[10] Anthony B Atkinson. On the measurement of inequality. *Journal of Economic Theory*, 2(3):244–263, 1970. ISSN 0022-0531. doi: https://doi.org/10.1016/0022-0531(70)90039-6. URL https://www.sciencedirect.com/science/article/pii/0022053170900396.

[11] Charles Audet, Jean Bigeon, Dominique Cartier, Sébastien Le Digabel, and Ludovic Salomon. Performance indicators in multiobjective optimization. *European Journal of Operational Research*, 292(2):397–422, 2020. doi: 10.1016/j.ejor.2020.11.016. URL https://hal.science/hal-03048871.

[12] Max Bachmann. rapidfuzz/rapidfuzz: Release 3.8.1, April 2024. URL https://doi.org/10.5281/zenodo.10938887.

[13] Krisztian Balog, Donald Metzler, and Zhen Qin. Rankers, judges, and assistants: Towards understanding the interplay of llms in information retrieval evaluation, 2025. URL https://arxiv.org/abs/2503.19092.

[14] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[15] Christine Bauer, Alan Said, and Eva Zangerle. Evaluation Perspectives of Recommender Systems: Driving Research and Education (Dagstuhl Seminar

24211). *Dagstuhl Reports*, 14(5):58–172, 2024. ISSN 2192-5283. doi: 10.4230/DagRep.14.5.58. URL https://drops.dagstuhl.de/entities/document/10.4230/DagRep.14.5.58.

[16] Joeran Beel and Stefan Langer. A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. In Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla, editors, *Proceedings of the 19th International Conference on Theory and Practice of Digital Libraries (TPDL)*, volume 9316 of *Lecture Notes in Computer Science*, pages 153–168, 2015. doi: 10.1007/978-3-319-24592-8_12.

[17] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1 1995. ISSN 2517-6161. doi: 10.1111/J.2517-6161.1995.TB02031.X.

[18] 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 & Data Mining*, KDD '19, page 2212–2220, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330745. URL https://doi.org/10.1145/3292500.3330745.

[19] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, volume 18, pages 405–414. Association for Computing Machinery, Inc, 6 2018. ISBN 9781450356572. doi: 10.1145/3209978.3210063. URL https://doi.org/10.1145/3209978.3210063.

[20] Arpita Biswas, Gourab K Patro, Niloy Ganguly, Krishna P Gummadi, and Abhijnan Chakraborty. Toward Fair Recommendation in Two-sided Platforms. *ACM Trans. Web*, 16(2), 12 2021. ISSN 1559-1131. doi: 10.1145/3503624. URL https://doi.org/10.1145/3503624.

[21] Charles Blackorby, Walter Bossert, and David Donaldson. *Income Inequality Measurement: The Normative Approach*, pages 133–161. Springer Netherlands, Dordrecht, 1999. ISBN 978-94-011-4413-1. doi: 10.1007/978-94-011-4413-1_4. URL https://doi.org/10.1007/978-94-011-4413-1_4.

[22] Simone Borg Bruun, Maria Maistro, and Christina Lioma. Learning Recommendations from User Actions in the Item-poor Insurance Domain. In *RecSys 2022 - Proceedings of the 16th ACM Conference on Recommender Systems*, pages 113–123. Association for Computing Machinery, Inc, 9 2022. ISBN 9781450392785. doi: 10.1145/3523227.3546775. URL https://dl.acm.org/doi/10.1145/3523227.3546775.

[23] 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*, New York, NY, USA, 2019. ACM. ISBN 9781450362382. doi: 10.1145/3297662. URL https://doi.org/10.1145/3297662.3365798.

[24] Francois Bourguignon. Decomposable Income Inequality Measures. *Econometrica*, 47(4):901–920, 1979.

[25] Amanda Bower, Hamid Eftekhari, Mikhail Yurochkin, and Yuekai Sun. Individually Fair Ranking. In *ICLR 2021 - 9th International Conference on Learning Representations*. OpenReview.net, 2021. URL https://openreview.net/forum?id=71zCSP_HuBN.

[26] Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, page 33–40, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132263. doi: 10.1145/345508.345543. URL https://doi.org/10.1145/345508.345543.

[27] Robin Burke. Multisided fairness for recommendation, 2017. URL https://arxiv.org/abs/1707.00093.

[28] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. Balanced Neighborhoods for Multi-sided Fairness in Recommendation, 1 2018. ISSN 2640-3498. URL https://proceedings.mlr.press/v81/burke18a.html.

[29] Robin Burke, Gediminas Adomavicius, Toine Bogers, Tommaso Di Noia, Dominik Kowald, Julia Neidhardt, Özlem Özgöbek, Maria Soledad Pera, Nava Tintarev, and Jürgen Ziegler. De-centering the (Traditional) User: Multistakeholder Evaluation of Recommender Systems, 2025. URL https://arxiv.org/abs/2501.05170.

[30] Maarten Buyl and Tijl De Bie. Inherent Limitations of AI Fairness. *Commun. ACM*, 67(2):48–55, 1 2024. ISSN 0001-0782. doi: 10.1145/3624700. URL `https://doi.org/10.1145/3624700`.

[31] Rocío Cañamares, Pablo Castells, and Alistair Moffat. Offline evaluation options for recommender systems. *Information Retrieval Journal*, 23(4): 387–410, 2020. ISSN 15737659. doi: 10.1007/s10791-020-09371-3. URL `https://doi.org/10.1007/s10791-020-09371-3`.

[32] Iván Cantador, Peter Brusilovsky, and Tsvi Kuflik. 2nd Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011). In *Proceedings of the 5th ACM conference on Recommender systems*, RecSys 2011, New York, NY, USA, 2011. ACM.

[33] Pablo Castells and Alistair Moffat. Offline recommender system evaluation: Challenges and new directions. *AI Magazine*, 43(2):225–238, 2022. doi: https://doi.org/10.1002/aaai.12051. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.12051`.

[34] Pablo Castells, Neil J. Hurley, and Saul Vargas. *Novelty and Diversity in Recommender Systems*, pages 881–918. Springer US, Boston, MA, 2015. ISBN 978-1-4899-7637-6. doi: 10.1007/978-1-4899-7637-6_26. URL `https://doi.org/10.1007/978-1-4899-7637-6_26`.

[35] O Celma. *Music Recommendation and Discovery in the Long Tail*. Springer, 2010.

[36] Yair Censor. Pareto optimality in multiobjective problems. *Applied Mathematics & Optimization*, 4(1):41–59, 3 1977. ISSN 14320606. doi: 10.1007/BF01442131/METRICS. URL `https://link.springer.com/article/10.1007/BF01442131`.

[37] Lidia Ceriani and Paolo Verme. The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *J Econ Inequal*, 10:421–443, 2012. doi: 10.1007/s10888-011-9188-x. URL `http://www.umass.edu/wsp/statistics/tales/gini.html`.

[38] Wei Chen, Yiqing Wu, Zhao Zhang, Fuzhen Zhuang, Zhongshi He, Ruobing Xie, and Feng Xia. FairGap: Fairness-Aware Recommendation via Generating Counterfactual Graph. *ACM Trans. Inf. Syst.*, 42(4), 2 2024. ISSN 1046-8188. doi: 10.1145/3638352. URL `https://doi.org/10.1145/3638352`.

[39] Weixin Chen, Li Chen, and Yuhan Zhao. Investigating user-side fairness in outcome and process for multi-type sensitive attributes in recommendations. *ACM Trans. Recomm. Syst.*, April 2025. doi: 10.1145/3731568. URL `https://doi.org/10.1145/3731568`. Just Accepted.

[40] Sachin Pathiyan Cherumanal, Damiano Spina, Falk Scholer, and W. Bruce Croft. Evaluating Fairness in Argument Retrieval. In *International Conference on Information and Knowledge Management, Proceedings*, pages 3363–3367. Association for Computing Machinery, 10 2021. ISBN 9781450384469. doi: 10.1145/3459637.3482099. URL `https://dl.acm.org/doi/10.1145/3459637.3482099`.

[41] Konstantina Christakopoulou, Alex Beutel, Rui Li, Sagar Jain, and Ed H Chi. Q&R: A Two-Stage Approach toward Interactive Recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 139–148, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219894. URL `https://doi.org/10.1145/3219819.3219894`.

[42] Team Cohere, :, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, Neeral Beladia, Walter Beller-Morales, Alexandre Bérard, Andrew Berneshawi, Anna Bialas, Phil Blunsom, Matt Bobkin, Adi Bongale, Sam Braun, Maxime Brunet, Samuel Cahyawijaya, David Cairuz, Jon Ander Campos, Cassie Cao, Kris Cao, Roman Castagné, Julián Cendrero, Leila Chan Currie, Yash Chandak, Diane Chang, Giannis Chatziveroglou, Hongyu Chen, Claire Cheng, Alexis Chevalier, Justin T. Chiu, Eugene Cho, Eugene Choi, Eujeong Choi, Tim Chung, Volkan Cirik, Ana Cismaru, Pierre Clavier, Henry Conklin, Lucas Crawhall-Stein, Devon Crouse, Andres Felipe Cruz-Salinas, Ben Cyrus, Daniel D'souza, Hugo Dalla-Torre, John Dang, William Darling, Omar Darwiche Domingues, Saurabh Dash, Antoine Debugne, Théo Dehaze, Shaan Desai, Joan Devassy, Rishit Dholakia, Kyle Duffy, Ali Edalati, Ace Eldeib, Abdullah Elkady, Sarah Elsharkawy, Irem Ergün, Beyza Ermis, Marzieh Fadaee, Boyu Fan, Lucas Fayoux, Yannis Flet-Berliac, Nick Frosst, Matthias Gallé, Wojciech Galuba, Utsav Garg, Matthieu Geist, Mohammad Gheshlaghi Azar, Ellen Gilsenan-McMahon, Seraphina Goldfarb-Tarrant, Tomas Goldsack, Aidan Gomez, Victor Machado Gonzaga, Nithya Govindarajan, Manoj Govindassamy, Nathan

Grinsztajn, Nikolas Gritsch, Patrick Gu, Shangmin Guo, Kilian Haefeli, Rod Hajjar, Tim Hawes, Jingyi He, Sebastian Hofstätter, Sungjin Hong, Sara Hooker, Tom Hosking, Stephanie Howe, Eric Hu, Renjie Huang, Hemant Jain, Ritika Jain, Nick Jakobi, Madeline Jenkins, JJ Jordan, Dhruti Joshi, Jason Jung, Trushant Kalyanpur, Siddhartha Rao Kamalakara, Julia Kedrzycki, Gokce Keskin, Edward Kim, Joon Kim, Wei-Yin Ko, Tom Kocmi, Michael Kozakov, Wojciech Kryściński, Arnav Kumar Jain, Komal Kumar Teru, Sander Land, Michael Lasby, Olivia Lasche, Justin Lee, Patrick Lewis, Jeffrey Li, Jonathan Li, Hangyu Lin, Acyr Locatelli, Kevin Luong, Raymond Ma, Lukáš Mach, Marina Machado, Joanne Magbitang, Brenda Malacara Lopez, Aryan Mann, Kelly Marchisio, Olivia Markham, Alexandre Matton, Alex McKinney, Dominic McLoughlin, Jozef Mokry, Adrien Morisot, Autumn Moulder, Harry Moynehan, Maximilian Mozes, Vivek Muppalla, Lidiya Murakhovska, Hemangani Nagarajan, Alekhya Nandula, Hisham Nasir, Shauna Nehra, Josh Netto-Rosen, Daniel Ohashi, James Owers-Bardsley, Jason Ozuzu, Dennis Padilla, Gloria Park, Sam Passaglia, Jeremy Pekmez, Laura Penstone, Aleksandra Piktus, Case Ploeg, Andrew Poulton, Youran Qi, Shubha Raghvendra, Miguel Ramos, Ekagra Ranjan, Pierre Richemond, Cécile Robert-Michon, Aurélien Rodriguez, Sudip Roy, Sebastian Ruder, Laura Ruis, Louise Rust, Anubhav Sachan, Alejandro Salamanca, Kailash Karthik Saravanakumar, Isha Satyakam, Alice Schoenauer Sebag, Priyanka Sen, Sholeh Sepehri, Preethi Seshadri, Ye Shen, Tom Sherborne, Sylvie Shang Shi, Sanal Shivaprasad, Vladyslav Shmyhlo, Anirudh Shrinivason, Inna Shteinbuk, Amir Shukayev, Mathieu Simard, Ella Snyder, Ava Spataru, Victoria Spooner, Trisha Starostina, Florian Strub, Yixuan Su, Jimin Sun, Dwarak Talupuru, Eugene Tarassov, Elena Tommasone, Jennifer Tracey, Billy Trend, Evren Tumer, Ahmet Üstün, Bharat Venkitesh, David Venuto, Pat Verga, Maxime Voisin, Alex Wang, Donglu Wang, Shijian Wang, Edmond Wen, Naomi White, Jesse Willman, Marysia Winkels, Chen Xia, Jessica Xie, Minjie Xu, Bowen Yang, Tan Yi-Chern, Ivan Zhang, Zhenyu Zhao, and Zhoujie Zhao. Command a: An enterprise-ready large language model, 2025. URL `https://arxiv.org/abs/2504.00698`.

[43] 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`.

[44] Kimberle Crenshaw. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, 43(6):1241–1299, 1991. ISSN 00389765. URL http://www.jstor.org/stable/1229039.

[45] Maurizio Ferrari Dacrema, Nicolò Felicioni, and Paolo Cremonesi. Offline evaluation of recommender systems in a user interface with multiple carousels. *Frontiers in Big Data*, 5:910030, 6 2022. ISSN 2624909X. doi: 10.3389/FDATA.2022.910030/BIBTEX. URL www.frontiersin.org.

[46] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, and Dasarathi Sampath. The youtube video recommendation system. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, page 293–296, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589060. doi: 10.1145/1864708.1864770. URL https://doi.org/10.1145/1864708.1864770.

[47] Meltem Dayioğlu and Cem Başlevent. Imputed rents and regional income inequality in turkey: A subgroup decomposition of the atkinson index. *Regional Studies*, 40(8):889–905, 2006. doi: 10.1080/00343400600984395. URL https://doi.org/10.1080/00343400600984395.

[48] Fernando G. De Maio. Income inequality measures. *Journal of Epidemiology and Community Health*, 61(10):849, 10 2007. ISSN 0143005X. doi: 10.1136/JECH.2006.052969. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2652960/.

[49] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li,

Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL `https://arxiv.org/abs/2501.12948`.

[50] Yashar Deldjoo and Tommaso Di Noia. Cfairllm: Consumer fairness evaluation in large-language model recommender system. *ACM Trans. Intell. Syst. Technol.*, March 2025. ISSN 2157-6904. doi: 10.1145/3725853. URL `https://doi.org/10.1145/3725853`. Just Accepted.

[51] Yashar Deldjoo and Fatemeh Nazary. A normative framework for benchmarking consumer fairness in large language model recommender system, 2024. URL `https://arxiv.org/abs/2405.02219`.

[52] Yashar Deldjoo, Vito Walter Anelli, Hamed Zamani, Alejandro Bellogín, and Tommaso Di Noia. Recommender Systems Fairness Evaluation via Generalized Cross Entropy. In *Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019)*. CEUR-WS, 2019.

[53] Yashar Deldjoo, Vito Walter Anelli, Hamed Zamani, Alejandro Bellogín, • Tommaso, and Di Noia. A flexible framework for evaluating user and item fairness in recommender systems. *User Modeling and User-Adapted Interaction*, 31:457–511, 2021. doi: 10.1007/s11257-020-09285-1. URL https://doi.org/10.1007/s11257-020-09285-1.

[54] 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*, 34(1): 59–108, 2024. ISSN 1573-1391. doi: 10.1007/s11257-023-09364-z. URL https://doi.org/10.1007/s11257-023-09364-z.

[55] Mukund Deshpande and George Karypis. Item-based top-N recommendation algorithms. *ACM Transactions on Information Systems*, 22(1):143–177, 1 2004. ISSN 10468188. doi: 10.1145/963770.963776. URL https://dl.acm.org/doi/10.1145/963770.963776.

[56] 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 & Knowledge Management*, New York, NY, USA, 2020. ACM. ISBN 9781450368599. doi: 10.1145/3340531. URL https://doi.org/10.1145/3340531.3411962.

[57] Karlijn Dinnissen and Christine Bauer. Amplifying artists' voices: Item provider perspectives on influence and fairness of music streaming platforms. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23, page 238–249, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399326. doi: 10.1145/3565472.3592960. URL https://doi.org/10.1145/3565472.3592960.

[58] Virginie Do. *Fairness in recommender systems: insights from social choice*. Theses, Université Paris sciences et lettres, July 2023. URL https://theses.hal.science/tel-04213955.

[59] Virginie Do and Nicolas Usunier. Optimizing Generalized Gini Indices for Fairness in Rankings. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, volume 1, pages 737–747, New York, NY, USA, 2022. ACM. ISBN 9781450387323. doi: 10.1145/3477495. URL https://doi.org/10.1145/3477495.3532035.

[60] Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. Two-sided fairness in rankings via Lorenz dominance. In *Advances in Neural Information Processing Systems*, volume 34, pages 8596–8608, 2021.

[61] Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. Online Certification of Preference-Based Fairness for Personalized Recommender Systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6): 6532–6540, 6 2022. doi: 10.1609/aaai.v36i6.20606. URL `https://ojs.aaai.org/index.php/AAAI/article/view/20606`.

[62] Yijiang River Dong, Tiancheng Hu, and Nigel Collier. Can LLM be a personalized judge? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10126–10141, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.592. URL `https://aclanthology.org/2024.findings-emnlp.592/`.

[63] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *ITCS 2012 - Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL `https://dl.acm.org/doi/10.1145/2090236.2090255`.

[64] 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, 10 2018. URL `https://proceedings.mlr.press/v81/ekstrand18b.html`.

[65] 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`.

[66] Michael D Ekstrand, Ben Carterette, and Fernando Diaz. Distributionally-Informed Recommender System Evaluation. *ACM Trans. Recomm. Syst.*, 8 2023. doi: 10.1145/3613455. URL `https://doi.org/10.1145/3613455`.

[67] Mehdi Elahi, Himan Abdollahpouri, Masoud Mansoury, and Helma Torkamaan. Beyond algorithmic fairness in recommender systems. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '21, page 41–46, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383677. doi: 10.1145/3450614.3461685. URL https://doi.org/10.1145/3450614.3461685.

[68] Guido Erreygers, Roselinde Kessels, Linkun Chen, and Philip Clarke. Subgroup decomposability of income-related inequality of health, with an application to australia. *Economic Record*, 94(304):39–50, 2018. doi: https://doi.org/10.1111/1475-4932.12373. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-4932.12373.

[69] European Parliament and Council. Regulation (EU) 2024/1689: Artificial Intelligence Act, 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng. Recital 27 discusses principles for trustworthy and ethical AI.

[70] Eurostat. Young people on the labour market - statistics, 2018. URL https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Young_people_on_the_labour_market_-_statistics.

[71] Alessandro Fabris, Gianmaria Silvello, Gian Antonio Susto, and Asia J Biega. Pairwise Fairness in Ranking as a Dissatisfaction Measure. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, pages 931–939, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394079. doi: 10.1145/3539597.3570459. URL https://doi.org/10.1145/3539597.3570459.

[72] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642. doi: 10.1145/2783258.2783311. URL https://doi.org/10.1145/2783258.2783311.

[73] Andres Ferraro, Xavier Serra, and Christine Bauer. Break the loop: Gender imbalance in music recommenders. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR '21, page 249–254, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380553. doi: 10.1145/3406522.3446033. URL https://doi.org/10.1145/3406522.3446033.

[74] Andres Ferraro, Xavier Serra, and Christine Bauer. What is fair? exploring the artists' perspective on the fairness of music streaming platforms. In Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen, editors, *Human-Computer Interaction – INTERACT 2021*, pages 562–584, Cham, 2021. Springer International Publishing. ISBN 978-3-030-85616-8.

[75] Andres Ferraro, Michael D. Ekstrand, and Christine Bauer. It's not you, it's me: The impact of choice models and ranking strategies on gender imbalance in music recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 884–889, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705052. doi: 10.1145/3640457.3688163. URL `https://doi.org/10.1145/3640457.3688163`.

[76] Jill Freyne and Shlomo Berkovsky. *Evaluating Recommender Systems for Supportive Technologies*, pages 195–217. Springer London, London, 2013. ISBN 978-1-4471-4778-7. doi: 10.1007/978-1-4471-4778-7_8. URL `https://doi.org/10.1007/978-1-4471-4778-7_8`.

[77] 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`.

[78] Jingtong Gao, Bo Chen, Xiangyu Zhao, Weiwen Liu, Xiangyang Li, Yichao Wang, Wanyu Wang, Huifeng Guo, and Ruiming Tang. Llm4rerank: Llm-based auto-reranking framework for recommendations. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 228–239, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714922. URL `https://doi.org/10.1145/3696410.3714922`.

[79] Ruoyuan Gao, Yingqiang Ge, and Chirag Shah. FAIR: Fairness-aware information retrieval evaluation. *Journal of the Association for Information Science and Technology*, 2022. ISSN 23301643. doi: 10.1002/ASI.24648.

[80] Florent Garcin, Boi Faltings, Olivier Donatsch, Ayar Alazzawi, Christophe Bruttin, and Amr Huber. Offline and online evaluation of news recommender

systems at swissinfo.ch. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, page 169–176, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450326681. doi: 10.1145/2645710.2645745. URL https://doi.org/10.1145/2645710.2645745.

[81] David García-Soriano and Francesco Bonchi. Maxmin-Fair Ranking: Individual Fairness under Group-Fairness Constraints. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pages 436–446, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467349. URL https://doi.org/10.1145/3447548.3467349.

[82] 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. *WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 445–453, 8 2021. doi: 10.1145/3437963.3441824.

[83] 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 *WSDM 2022 - Proceedings of the 15th ACM International Conference on Web Search and Data Mining*, pages 316–324, Virtual Event, AZ, USA, 2 2022. Association for Computing Machinery, Inc. ISBN 9781450391320. doi: 10.1145/3488560.3498487. URL https://dl.acm.org/doi/10.1145/3488560.3498487.

[84] Avijit Ghosh, Tomo Lazovich, Kristian Lum, and Christo Wilson. Reducing population-level inequality can improve demographic group fairness: a twitter case study, 2024. URL https://arxiv.org/abs/2409.08135.

[85] C. Gini. Variabilità e mutabilità. *Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*, 1912.

[86] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan

Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

[87] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A Constant Time Collaborative Filtering Algorithm. *Information Retrieval*, 4(2):133–151, 7 2001. ISSN 13864564. doi: 10.1023/A:1011419012209/METRICS. URL `https://link.springer.com/article/10.1023/A:1011419012209`.

[88] Elizabeth Gómez, David Contreras, Ludovico Boratto, and Maria Salamo. Ambar: A dataset for assessing multiple beyond-accuracy recommenders. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 137–147, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705052. doi: 10.1145/3640457.3688067. URL `https://doi.org/10.1145/3640457.3688067`.

[89] Carlos A. Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Trans. Manage. Inf. Syst.*, 6 (4), December 2016. ISSN 2158-656X. doi: 10.1145/2843948. URL `https://doi.org/10.1145/2843948`.

[90] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert,

Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie

Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rit-

tner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

[91] Asela Gunawardana, Guy Shani, and Sivan Yogev. *Evaluating Recommender Systems*, pages 547–601. Springer US, New York, NY, 2022. ISBN 978-1-0716-2197-4. doi: 10.1007/978-1-0716-2197-4_15. URL `https://doi.org/10.1007/978-1-0716-2197-4_15`.

[92] Elizabeth Gómez, David Contreras, Ludovico Boratto, and Maria Salamó. Enhancing recommender systems with provider fairness through preference distribution awareness. *International Journal of Information Management Data Insights*, 5(1):100311, 2025. ISSN 2667-0968. doi: https://doi.org/10.1016/j.jjimei.2024.100311. URL `https://www.sciencedirect.com/science/article/pii/S2667096824001009`.

[93] Ben Hamner, Road Warrior, and Wojciech Krupa. Job Recommendation Challenge. https://kaggle.com/competitions/job-recommendation, 2012.

[94] F Maxwell Harper and Joseph A Konstan. The MovieLens datasets: History and context. *ACM Trans. Interact. Intell. Syst. 5, 4, Article*, 19, 2015. doi: 10.1145/2827872. URL `http://dx.doi.org/10.1145/2827872`.

[95] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat Seng Chua. Neural collaborative filtering. In *26th International World Wide Web Conference, WWW 2017*, pages 173–182. International World Wide Web Conferences Steering Committee, 2017. ISBN 9781450349130. doi: 10.1145/3038912.3052569. URL `https://dl.acm.org/doi/10.1145/3038912.3052569`.

[96] Hellinger distance. Encyclopedia of mathematics. URL `https://encyclopediaofmath.org/wiki/Hellinger_distance`.

[97] 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, pages 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`.

[98] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, January 2004. ISSN 1046-8188. doi: 10.1145/963770.963772. URL `https://doi.org/10.1145/963770.963772`.

[99] S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 1979. doi: 10.2307/4615733.

[100] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large Language Models are Zero-Shot Rankers for Recommender Systems. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 14609 LNCS, pages 364–381. Springer-Verlag, Berlin, Heidelberg, 2024. ISBN 9783031560590. URL `https://link.springer.com/10.1007/978-3-031-56060-6_24`.

[101] Rajendra K Jain, Dah-Ming W Chiu, William R Hawe, and others. A Quantitative Measure Of Fairness And Discrimination For Resource Allocation In Shared Computer Systems, 1998. URL `http://arxiv.org/abs/cs/9809099`.

[102] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 10 2002. ISSN 10468188. doi: 10.1145/582415.582418. URL https://dl.acm.org/doi/10.1145/582415.582418.

[103] Olivier Jeunen and Bart Goethals. Top-K Contextual Bandits with Equity of Exposure. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, pages 310–320, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384582. doi: 10.1145/3460231.3474248. URL https://doi.org/10.1145/3460231.3474248.

[104] Yitong Ji, Aixin Sun, Jie Zhang, and Chenliang Li. A re-visit of the popularity baseline in recommender systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1749–1752, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164. doi: 10.1145/3397271.3401233. URL https://doi.org/10.1145/3397271.3401233.

[105] Chumeng Jiang, Jiayin Wang, Weizhi Ma, Charles L. A. Clarke, Shuai Wang, Chuhan Wu, and Min Zhang. Beyond utility: Evaluating llm as recommender. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 3850–3862, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714759. URL https://doi.org/10.1145/3696410.3714759.

[106] JiYitong, SunAixin, ZhangJie, and LiChenliang. A Critical Study on Data Leakage in Recommender System Offline Evaluation. *ACM Transactions on Information Systems*, 41(3):1–27, 2 2023. ISSN 1046-8188. doi: 10.1145/3569930. URL https://dl.acm.org/doi/10.1145/3569930.

[107] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. *SIGIR 2005 - Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 154–161, 2005. doi: 10.1145/1076034.1076063. URL https://dl.acm.org/doi/10.1145/1076034.1076063.

[108] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972. ISSN 00220418. doi: 10.1108/EB026526/FULL/PDF.

[109] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. News recommender systems – Survey and roads ahead. *Information Processing & Management*, 54 (6):1203–1227, 11 2018. ISSN 0306-4573. doi: 10.1016/J.IPM.2018.04.008.

[110] Saeedeh Karimi, Hossein A. Rahmani, Mohammadmehdi Naghiaei, and Leila Safari. Provider fairness and beyond-accuracy trade-offs in recommender systems. In *The 6th Workshop on Responsible Recommendation (FAccTRec 2023) at RecSys 2023*, 2023. URL `https://arxiv.org/abs/2309.04250`.

[111] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 100–109, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287592. URL `https://doi.org/10.1145/3287560.3287592`.

[112] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2): 81–93, 06 1938. ISSN 0006-3444. doi: 10.1093/biomet/30.1-2.81. URL `https://doi.org/10.1093/biomet/30.1-2.81`.

[113] M. G. Kendall. The Treatment of Ties in Ranking Problems. *Biometrika*, 33 (3):239–251, 11 1945. ISSN 0006-3444. doi: 10.1093/BIOMET/33.3.239. URL `https://dx.doi.org/10.1093/biomet/33.3.239`.

[114] Naime Ranjbar Kermany, Weiliang Zhao, Jian Yang, and Jia Wu. ReInCre: Enhancing Collaborative Filtering Recommendations by Incorporating User Rating Credibility. In *Web Information Systems Engineering*, pages 64–72, Singapore, 2020. Springer Singapore. ISBN 978-981-15-3281-8.

[115] Naime Ranjbar Kermany, Weiliang Zhao, Tseesuren Batsuuri, Jian Yang, and Jia Wu. Incorporating user rating credibility in recommender systems. *Future Generation Computer Systems*, 147:30–43, 2023. ISSN 0167-739X. doi: https://doi.org/10.1016/j.future.2023.04.029. URL `https://www.sciencedirect.com/science/article/pii/S0167739X2300170X`.

[116] Tahsin Alamgir Kheya, Mohamed Reda Bouadjenek, and Sunil Aryal. Unmasking gender bias in recommendation systems and enhancing category-aware fairness. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 5127–5138, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714528. URL `https://doi.org/10.1145/3696410.3714528`.

[117] Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 12 2014. doi: 10.48550/arxiv.1412.6980. URL https://arxiv.org/abs/1412.6980v9.

[118] Peter Knees, Julia Neidhardt, and Irina Nalis. *Recommender Systems: Techniques, Effects, and Measures Toward Pluralism and Fairness*, pages 417–434. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-45304-5. doi: 10.1007/978-3-031-45304-5_27. URL https://doi.org/10.1007/978-3-031-45304-5_27.

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

[120] Walid Krichene and Steffen Rendle. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1748–1757, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403226. URL https://doi.org/10.1145/3394486.3403226.

[121] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4069–4079, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[122] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E Gonzalez, Hao Zhang, and Ion Stoica. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[123] Maciej Laszczyk and Paweł B. Myszkowski. Survey of quality measures for multi-objective optimization: Construction of complementary set of multi-objective quality measures. *Swarm and Evolutionary Computation*, 48:109–133, 8 2019. ISSN 2210-6502. doi: 10.1016/J.SWEVO.2019.04.001.

[124] Tomo Lazovich, Luca Belli, Aaron Gonzales, Amanda Bower, Uthaipon Tantipongpipat, Kristian Lum, Ferenc Huszár, and Rumman Chowdhury. Mea-

suring disparate outcomes of content recommendation algorithms with distributional inequality metrics. *Patterns*, 3(8), 8 2022. ISSN 2666-3899. doi: 10.1016/j.patter.2022.100568. URL https://doi.org/10.1016/j.patter.2022.100568.

[125] Joon Ho Lee. Analyses of multiple evidence combination. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–276, Philadelphia, 7 1997. Association for Computing Machinery (ACM). doi: 10.1145/258525.258587. URL https://dl.acm.org/doi/10.1145/258525.258587.

[126] Jurek Leonhardt, Avishek Anand, and Megha Khosla. User Fairness in Recommender Systems. *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, pages 101–102, 4 2018. doi: 10.1145/3184558.3186949.

[127] Lukas Lerche, Dietmar Jannach, and Malte Ludewig. On the value of reminders within e-commerce recommendations. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, UMAP '16, page 27–35, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450343688. doi: 10.1145/2930238.2930244. URL https://doi.org/10.1145/2930238.2930244.

[128] Jie Li, Yongli Ren, Mark Sanderson, and Ke Deng. Explaining Recommendation Fairness from a User/Item Perspective. *ACM Trans. Inf. Syst.*, 10 2024. ISSN 1046-8188. doi: 10.1145/3698877. URL https://doi.org/10.1145/3698877.

[129] Roger Zhe Li, Julián Urbano, and Alan Hanjalic. Leave No User Behind: Towards Improving the Utility of Recommender Systems for Non-mainstream Users. *WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 103–111, 8 2021. doi: 10.1145/3437963.3441769.

[130] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. User-oriented fairness in recommendation. *The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021*, pages 624–632, 4 2021. doi: 10.1145/3442381.3449866.

[131] 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.

[132] Yunqi Li, Yingqiang Ge, and Yongfeng Zhang. Tutorial on fairness of machine learning in recommender systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2654–2657, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462814. URL https://doi.org/10.1145/3404835.3462814.

[133] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. Fairness in Recommendation: Foundations, Methods and Applications. *ACM Transactions on Intelligent Systems and Technology*, 1 2023. ISSN 2157-6904. doi: 10.1145/3610302. URL https://dl.acm.org/doi/10.1145/3610302.

[134] Jianxun Lian, Yuxuan Lei, Xu Huang, Jing Yao, Wei Xu, and Xing Xie. RecAI: Leveraging Large Language Models for Next-Generation Recommender Systems. *arXiv preprint arXiv:2403.06465*, 2024.

[135] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*, volume 10, pages 689–698. Association for Computing Machinery, Inc, 4 2018. ISBN 9781450356398. doi: 10.1145/3178876.3186150. URL https://doi.org/10.1145/3178876.3186150.

[136] Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. LLaRA: Large Language-Recommendation Assistant. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 1785–1795, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657690. URL https://doi.org/10.1145/3626772.3657690.

[137] Lucas Chaves Lima. *Semantic Classification and Evaluation*. PhD thesis, University of Copenhagen, 2021. URL https://di.ku.dk/english/research/phd/phd-theses/2021/Lucas_Chaves_Lima_Phd_Thesis.pdf.

[138] Xiao Lin, Hongjie Chen, Changhua Pei, Fei Sun, Xuanji Xiao, Hanxiao Sun, Yongfeng Zhang, Wenwu Ou, and Peng Jiang. A pareto-eficient algorithm for multiple objective optimization in e-commerce recommendation. In *RecSys 2019 - 13th ACM Conference on Recommender Systems*, pages 20–28, Copenhagen, Denmark, 9 2019. Association for Computing Machinery, Inc. ISBN 9781450362436. doi: 10.1145/3298689.3346998. URL https://dl.acm.org/doi/10.1145/3298689.3346998.

[139] Zihan Lin, Changxin Tian, Yupeng Hou, and Wayne Xin Zhao. Improving Graph Collaborative Filtering with Neighborhood-enriched Contrastive Learning. In *WWW 2022 - Proceedings of the ACM Web Conference 2022*, pages 2320–2329, Virtual Event, Lyon, France, 4 2022. Association for Computing Machinery, Inc. ISBN 9781450390965. doi: 10.1145/3485447.3512104. URL https://dl.acm.org/doi/10.1145/3485447.3512104.

[140] Christina Lioma, Jakob Grue Simonsen, and Birger Larsen. Evaluation measures for relevance and credibility in ranked lists. In *ICTIR 2017 - Proceedings of the 2017 ACM SIGIR International Conference on the Theory of Information Retrieval*, pages 91–98, Amsterdam, The Netherlands, 10 2017. Association for Computing Machinery, Inc. ISBN 9781450344906. doi: 10.1145/3121050.3121072. URL https://dl.acm.org/doi/10.1145/3121050.3121072.

[141] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. Towards Conversational Recommendation over Multi-Type Dialogs. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049, Online, 7 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.98. URL https://aclanthology.org/2020.acl-main.98/.

[142] Zhiqiang Liu, Xiaoxiao Xu, Jiaqi Yu, Han Xu, Lantao Hu, Han Li, and Kun Gai. A self-adaptive fairness constraint framework for industrial recommender system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 4726–4733, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3680099. URL https://doi.org/10.1145/3627673.3680099.

[143] Maria Maistro, Lucas Chaves Lima, Jakob Grue Simonsen, and Christina Lioma. Principled Multi-Aspect Evaluation Measures of Rankings. In *Inter-*

national Conference on Information and Knowledge Management, Proceedings, pages 1232–1242, Virtual Event, Queensland, Australia, 10 2021. Association for Computing Machinery. ISBN 9781450384469. doi: 10.1145/3459637.3482287. URL https://dl.acm.org/doi/10.1145/3459637.3482287.

[144] Suvodeep Majumder, Joymallya Chakraborty, Gina R Bai, Kathryn T Stolee, and Tim Menzies. Fair Enough: Searching for Sufficient Measures of Fairness. *ACM Trans. Softw. Eng. Methodol.*, 32(6), 9 2023. ISSN 1049-331X. doi: 10.1145/3585006. URL https://doi.org/10.1145/3585006.

[145] Elizaveta Makhneva, Anna Sverkunova, Oleg Lashinin, Marina Ananyeva, and Sergey Kolesnikov. Make your next item recommendation model time sensitive. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23 Adjunct, pages 191–195, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450398916. doi: 10.1145/3563359.3596980. URL https://doi.org/10.1145/3563359.3596980.

[146] Brian Manolovitz and Mitsunori Ogihara. Practical evaluation of repeated recommendations in personalized music discovery. In *ISMIR*, pages 633–639, 2020.

[147] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. FairMatch: A Graph-based Approach for Improving Aggregate Diversity in Recommender Systems. In *UMAP 2020 - Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, volume 20, pages 154–162. ACM, 7 2020. ISBN 9781450368612. doi: 10.1145/3340631.3394860. URL https://doi.org/10.1145/3340631.3394860.

[148] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. A graph-based approach for mitigating multi-sided exposure bias in recommender systems. *ACM Trans. Inf. Syst.*, 40 (2), November 2021. ISSN 1046-8188. doi: 10.1145/3470948. URL https://doi.org/10.1145/3470948.

[149] M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *J. ACM*, 7(3):216–244, July 1960. ISSN 0004-5411. doi: 10.1145/321033.321035. URL https://doi.org/10.1145/321033.321035.

[150] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 2243–2251, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3272027. URL https://doi.org/10.1145/3269206.3272027.

[151] Zaiqiao Meng, Richard McCreadie, Craig MacDonald, and Iadh Ounis. Exploring Data Splitting Strategies for the Evaluation of Recommendation Models. In *RecSys 2020 - 14th ACM Conference on Recommender Systems*, pages 681–686, Virtual Event, Brazil, 9 2020. Association for Computing Machinery, Inc. ISBN 9781450375832. doi: 10.1145/3383313.3418479. URL https://dl.acm.org/doi/10.1145/3383313.3418479.

[152] Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Remi Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 117–123, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375832. URL https://doi.org/10.1145/3375627.3375832.

[153] Alistair Moffat. Seven Numeric Properties of Effectiveness Metrics. In Rafael E Banchs, Fabrizio Silvestri, Tie-Yan Liu, Min Zhang, Sheng Gao, and Jun Lang, editors, *Information Retrieval Technology*, pages 1–12, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-45068-6.

[154] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1), 12 2008. ISSN 10468188. doi: 10.1145/1416950.1416952. URL https://dl.acm.org/doi/10.1145/1416950.1416952.

[155] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. Controlling Fairness and Bias in Dynamic Learning-to-Rank. In *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 429–438. Association for Computing Machinery, Inc, 7 2020. ISBN 9781450380164. doi: 10.1145/3397271.3401100.

[156] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *EMNLP-IJCNLP*

2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, pages 188–197. Association for Computational Linguistics, 2019. ISBN 9781950737901. doi: 10.18653/V1/D19-1018. URL https://aclanthology.org/D19-1018.

[157] Vahid Partovi Nia, Alireza Ghaffari, Mahdi Zolnouri, and Yvon Savaria. Rethinking Pareto Frontier for Performance Evaluation of Deep Neural Networks, 2022.

[158] Xia Ning and George Karypis. SLIM: Sparse LInear Methods for top-N recommender systems. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 497–506, 2011. ISBN 9780769544083. doi: 10.1109/ICDM.2011.134. URL https://dl.acm.org/doi/10.1109/ICDM.2011.134.

[159] Encyclopedia of Mathematics. Metric. URL https://encyclopediaofmath.org/wiki/Metric.

[160] Office for National Statistics. Age classifications: Census 2021, 2023. URL https://www.ons.gov.uk/census/census2021dictionary/variablesbytopic/demographyvariablescensus2021/age/classifications.

[161] Harrie Oosterhuis. Computationally Efficient Optimization of Plackett-Luce Ranking Models for Relevance and Fairness. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 1023–1032, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462830. URL https://doi.org/10.1145/3404835.3462830.

[162] OpenAI. Learning to reason with llms, 9 2024. URL https://openai.com/index/learning-to-reason-with-llms/.

[163] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

[164] Joao Palotti, Guido Zuccon, and Allan Hanbury. MM: A new framework for multidimensional evaluation of search engines. In *International Conference on Information and Knowledge Management, Proceedings*, pages 1699–1702, Torino, Italy, 10 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3269261. URL https://dl.acm.org/doi/10.1145/3269206.3269261.

[165] Vincenzo Paparella, Vito Walter Anelli, Franco Maria Nardini, Raffaele Perego, and Tommaso Di Noia. Post-hoc selection of pareto-optimal solutions in search and recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 2013–2023, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701245. doi: 10.1145/3583780.3615010. URL https://doi.org/10.1145/3583780.3615010.

[166] Eliana Pastor and Francesco Bonchi. Intersectional fair ranking via subgroup divergence. *Data Min. Knowl. Discov.*, 38(4):2186–2222, May 2024. ISSN 1384-5810. doi: 10.1007/s10618-024-01029-8. URL https://doi.org/10.1007/s10618-024-01029-8.

[167] 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 *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020*, pages 1194–1204. Association for Computing Machinery, Inc, 4 2020. ISBN 9781450370233. doi: 10.1145/3366423.3380196.

[168] Gourab K Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. Fair Ranking: A Critical Review, Challenges, and Future Directions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1929–1942, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533238. URL https://doi.org/10.1145/3531146.3533238.

[169] Michael J. Pazzani and Daniel Billsus. *Content-Based Recommendation Systems*, pages 325–341. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-72079-9. doi: 10.1007/978-3-540-72079-9_10. URL https://doi.org/10.1007/978-3-540-72079-9_10.

[170] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3(none):96 – 146, 2009. doi: 10.1214/09-SS057. URL https://doi.org/10.1214/09-SS057.

[171] Giovanni Pellegrini, Vittorio Maria Faraco, and Yashar Deldjoo. Fairness for all: Investigating harms to within-group individuals in producer fairness re-ranking optimization – a reproducibility study, 2023. URL https://arxiv.org/abs/2309.09277.

[172] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250/.

[173] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. Fairness in rankings and recommendations: an overview. *VLDB Journal*, 31(3):431–458, 5 2022. ISSN 0949877X. doi: 10.1007/S00778-021-00697-Y/TABLES/3. URL https://link.springer.com/article/10.1007/s00778-021-00697-y.

[174] Flavien Prost, Ben Packer, Jilin Chen, Li Wei, Pierre Kremp, Nicholas Blumm, Susan Wang, Tulsee Doshi, Tonia Osadebe, Lukasz Heldt, Ed H. Chi, and Alex Beutel. Simpson's paradox in recommender fairness: Reconciling differences between per-user and aggregated evaluations, 2022. URL https://arxiv.org/abs/2210.07755.

[175] Amifa Raj and Michael D. Ekstrand. Measuring Fairness in Ranked Results. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 726–736, New York, NY, USA, 7 2022. ACM. ISBN 9781450387323. doi: 10.1145/3477495.3532018. URL https://dl.acm.org/doi/10.1145/3477495.3532018.

[176] Amifa Raj and Michael D. Ekstrand. Towards optimizing ranking in grid-layout for provider-side fairness. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part V*, page 90–105, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-56068-2. doi: 10.1007/978-3-031-56069-9_7. URL https://doi.org/10.1007/978-3-031-56069-9_7.

[177] Theresia Veronika Rampisela, Maria Maistro, Tuukka Ruotsalo, and Christina Lioma. Evaluation measures of individual item fairness for recommender systems: A critical study. *ACM Trans. Recomm. Syst.*, 3(2):1–52, November 2024. doi: 10.1145/3631943. URL https://doi.org/10.1145/3631943.

[178] Theresia Veronika Rampisela, Tuukka Ruotsalo, Maria Maistro, and Christina Lioma. Can We Trust Recommender System Fairness Evaluation? The Role of Fairness and Relevance. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 271–281, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657832. URL https://doi.org/10.1145/3626772.3657832.

[179] Theresia Veronika Rampisela, Maria Maistro, Tuukka Ruotsalo, Falk Scholer, and Christina Lioma. Relevance-aware individual item fairness measures for recommender systems: Limitations and usage guidelines. *ACM Trans. Recomm. Syst.*, September 2025. doi: 10.1145/3765624. URL https://doi.org/10.1145/3765624.

[180] Theresia Veronika Rampisela, Maria Maistro, Tuukka Ruotsalo, Falk Scholer, and Christina Lioma. Stairway to fairness: Connecting group and individual fairness. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, RecSys '25, page 677–683, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713644. doi: 10.1145/3705328.3748031. URL https://doi.org/10.1145/3705328.3748031.

[181] Theresia Veronika Rampisela, Tuukka Ruotsalo, Maria Maistro, and Christina Lioma. Joint Evaluation of Fairness and Relevance in Recommender Systems with Pareto Frontier. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 1548–1566, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714589. URL https://doi.org/10.1145/3696410.3714589.

[182] Al Mamunur Rashid, Istvan Albert, Dan Cosley, Shyong K. Lam, Sean M. McNee, Joseph A. Konstan, and John Riedl. Getting to know you. In *Proceedings of the 7th international conference on Intelligent user interfaces - IUI '02*, page 127, New York, New York, USA, 2002. ACM Press. ISBN 1581133820. doi: 10.1145/502716.502737. URL https://dl.acm.org/doi/10.1145/502716.502737.

[183] Bashir Rastegarpanah, Krishna P. Gummadi, and Mark Crovella. Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. *WSDM 2019 - Proceedings of the 12th ACM International Conference on Web Search and Data Mining*, pages 231–239, 2019. doi: 10.1145/3289600.3291002.

[184] Niklas Reisz, Vito D P Servedio, and Stefan Thurner. Quantifying the impact of homophily and influencer networks on song popularity prediction. *Scientific Reports*, 14(1):8929, 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-58969-w. URL `https://doi.org/10.1038/s41598-024-58969-w`.

[185] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 452–461, Arlington, Virginia, USA, 2009. AUAI Press. ISBN 9780974903958.

[186] 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, pages 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`.

[187] Marco Tulio Ribeiro, Nivio Ziviani, Edleno Silva De Moura, Itamar Hata, Anisio Lacerda, and Adriano Veloso. Multiobjective Pareto-Efficient Approaches for Recommender Systems. *ACM Trans. Intell. Syst. Technol.*, 5(4), 12 2015. ISSN 2157-6904. doi: 10.1145/2629350. URL `https://doi.org/10.1145/2629350`.

[188] Brianna Richardson, Jean Garcia-Gathright, and Samuel F. Way. Towards fairness in practice: A practitioner-oriented rubric for evaluating fair ml toolkits. *Conference on Human Factors in Computing Systems - Proceedings*, 5 2021. doi: 10.1145/3411764.3445604. URL `https://dl.acm.org/doi/10.1145/3411764.3445604`.

[189] Namhee Ryu and Seungjae Min. Multi-objective Optimization with an Adaptive Weight Determination Scheme Using the Concept of Hyperplane: Multi-objective Optimization with an Adaptive Weight. *International Journal for Numerical Methods in Engineering*, 118, 10 2018. doi: 10.1002/nme.6013.

[190] Dimitris Sacharidis. Top-n group recommendations with fairness. *Proceedings of the ACM Symposium on Applied Computing*, Part F147772:1663–1670, 2019. doi: 10.1145/3297280.3297442.

[191] Yuta Saito and Thorsten Joachims. Fair Ranking as Fair Division: Impact-Based Individual Fairness in Ranking. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14-18, 2022, Washington, DC, USA*, volume 1, pages 1514–1524. ACM, 2022. ISBN 9781450393850. doi: 10.1145/3534678.3539353.

[192] Tetsuya Sakai. *Metrics, Statistics, Tests*, pages 116–163. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. ISBN 978-3-642-54798-0. doi: 10.1007/978-3-642-54798-0_6. URL `https://doi.org/10.1007/978-3-642-54798-0_6`.

[193] Tetsuya Sakai and Zhaohao Zeng. Which diversity evaluation measures are "good"? In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 595–604, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361729. doi: 10.1145/3331184.3331215. URL `https://doi.org/10.1145/3331184.3331215`.

[194] Tetsuya Sakai, Jin Young Kim, and Inho Kang. A versatile framework for evaluating ranked lists in terms of group fairness and relevance. *ACM Trans. Inf. Syst.*, 42(1), August 2023. ISSN 1046-8188. doi: 10.1145/3589763. URL `https://doi.org/10.1145/3589763`.

[195] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. *Collaborative Filtering Recommender Systems*, pages 291–324. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-72079-9. doi: 10.1007/978-3-540-72079-9_9. URL `https://doi.org/10.1007/978-3-540-72079-9_9`.

[196] 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, pages 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`.

[197] Markus Schedl, Peter Knees, Brian McFee, Dmitry Bogdanov, and Marius Kaminskas. Music recommender systems. *Recommender systems handbook*, pages 453–492, 2015.

[198] Tobias Schumacher, Marlene Lutz, Sandipan Sikdar, and Markus Strohmaier. Properties of Group Fairness Measures for Rankings. *Trans. Soc. Comput.*, 8 2024. doi: 10.1145/3674883. URL `https://doi.org/10.1145/3674883`.

[199] 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, pages 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`.

[200] C E Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:623–656, 1948.

[201] Pengyang Shao, Le Wu, Kun Zhang, Defu Lian, Richang Hong, Yong Li, and Meng Wang. Average user-side counterfactual fairness for collaborative filtering. *ACM Trans. Inf. Syst.*, 42(5), May 2024. ISSN 1046-8188. doi: 10.1145/3656639. URL `https://doi.org/10.1145/3656639`.

[202] A. F. Shorrocks. The class of additively decomposable inequality measures. *Econometrica*, 48(3):613–625, 1980. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1913126`.

[203] Anthony F. Shorrocks. Inequality decomposition by population subgroups. *Econometrica*, 52(6):1369–1385, 1984. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1913511`.

[204] Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2219–2228. Association for Computing Machinery, 7 2018. ISBN 9781450355520. doi: 10.1145/3219819.3220088. URL `https://doi.org/10.1145/3219819.3220088`.

[205] Ashudeep Singh and Thorsten Joachims. Policy Learning for Fairness in Ranking. In H Wallach, H Larochelle, A Beygelzimer, F d Alché-Buc, E Fox, and R Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

URL https://proceedings.neurips.cc/paper_files/paper/2019/file/9e82757e9a1c12cb710ad680db11f6f1-Paper.pdf.

[206] Zheng Siting, Hong Wenxing, Zhang Ning, and Yang Fan. Job recommender systems: A survey. In *2012 7th International Conference on Computer Science & Education (ICCSE)*, pages 920–924, 2012. doi: 10.1109/ICCSE.2012.6295216.

[207] Jessie Smith, Nasim Sonboli, Casey Fiesler, and Robin Burke. Exploring user opinions of fairness in recommender systems. In *CHI'20 Workshop on Human-Centered Approaches to Fair and Responsible AI*, 2020. URL https://arxiv.org/abs/2003.06461.

[208] Jessie J. Smith and Lex Beattie. Recsys fairness metrics: Many to use but which one to choose?, 2022. URL https://arxiv.org/abs/2209.04011.

[209] Jessie J Smith, Lex Beattie, and Henriette Cramer. Scoping Fairness Objectives and Identifying Fairness Metrics for Recommender Systems: The Practitioners' Perspective. In *Proceedings of the ACM Web Conference 2023*, WWW '23, pages 3648–3659, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394161. doi: 10.1145/3543507.3583204. URL https://doi.org/10.1145/3543507.3583204.

[210] Jessie J. Smith, Aishwarya Satwani, Robin Burke, and Casey Fiesler. Recommend me? designing fairness metrics with providers. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2389–2399, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3659044. URL https://doi.org/10.1145/3630106.3659044.

[211] Nasim Sonboli, Jessie J. Smith, Florencia Cabral Berenfus, Robin Burke, and Casey Fiesler. Fairness and transparency in recommendation: The users' perspective. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '21, page 274–279, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383660. doi: 10.1145/3450613.3456835. URL https://doi.org/10.1145/3450613.3456835.

[212] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness

via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2239–2248, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3220046. URL https://doi.org/10.1145/3219819.3220046.

[213] Mistral AI team. Un ministral, des ministraux. https://mistral.ai/news/ministraux, 2024. Accessed: 2025-04-23.

[214] Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

[215] Antonela Tommasel. Fairness Matters: A look at LLM-generated group recommendations. In *18th ACM Conference on Recommender Systems*, pages 993–998, New York, NY, USA, 10 2024. ACM. ISBN 9798400705052. doi: 10.1145/3640457.3688182. URL https://dl.acm.org/doi/10.1145/3640457.3688182.

[216] Leonidas Tsepenekas, Ivan Brugere, Freddy Lecue, and Daniele Magazzeni. Comparing Apples to Oranges: Learning Similarity Functions for Data Produced by Different Distributions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=zjpjsJeVJZ.

[217] Robin Ungruh, Karlijn Dinnissen, Anja Volk, Maria Soledad Pera, and Hanna Hauptmann. Putting popularity bias mitigation to the test: A user-centric evaluation in music recommenders. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 169–178, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705052. doi: 10.1145/3640457.3688102. URL https://doi.org/10.1145/3640457.3688102.

[218] U.S. Bureau of Labor Statistics. May 2023 National Occupational Employment and Wage Estimates, 2023. URL https://www.bls.gov/oes/current/oes_nat.htm.

[219] U.S. Bureau of Labor Statistics. College Enrollment and Work Activity of Recent High School and College Graduates Summary, 4 2024. URL https://www.bls.gov/news.release/hsgec.nr0.htm.

[220] David A van Veldhuizen. *Multiobjective evolutionary algorithms: classifications, analyses, and new innovations.* PhD thesis, Air Force Institute

of Technology, 1999. URL https://api.semanticscholar.org/CorpusID:61080988.

[221] Bjørnar Vassøy and Helge Langseth. Consumer-side fairness in recommender systems: a systematic survey of methods and evaluation. *Artificial Intelligence Review*, 57(4):101, 2024. ISSN 1573-7462. doi: 10.1007/s10462-023-10663-5. URL https://doi.org/10.1007/s10462-023-10663-5.

[222] Sahil Verma, Ruoyuan Gao, and Chirag Shah. Facets of fairness in search and recommendation. In Ludovico Boratto, Stefano Faralli, Mirko Marras, and Giovanni Stilo, editors, *Bias and Social Aspects in Search and Recommendation*, pages 1–11, Cham, 2020. Springer International Publishing. ISBN 978-3-030-52485-2.

[223] Ellen M Voorhees. Evaluation by Highly Relevant Documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 74–82, New York, NY, USA, 2001. Association for Computing Machinery. ISBN 1581133316. doi: 10.1145/383952.383963. URL https://doi.org/10.1145/383952.383963.

[224] Mengting Wan, Jianmo Ni, Rishabh Misra, and Julian McAuley. Addressing marketing bias in product recommendations. *WSDM 2020 - Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 618–626, 1 2020. doi: 10.1145/3336191.3371855.

[225] Shuai Wang, Shaukat Ali, Tao Yue, Yan Li, and Marius Liaaen. A Practical Guide to Select Quality Indicators for Assessing Pareto-Based Search Algorithms in Search-Based Software Engineering. In *Proceedings of the 38th International Conference on Software Engineering*, ICSE '16, pages 631–642, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450339001. doi: 10.1145/2884781.2884880. URL https://doi.org/10.1145/2884781.2884880.

[226] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat Seng Chua. Neural graph collaborative filtering. In *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 165–174. Association for Computing Machinery, Inc, 7 2019. ISBN 9781450361729. doi: 10.1145/3331184.3331267. URL https://dl.acm.org/doi/10.1145/3331184.3331267.

[227] Xiuling Wang and Wendy Hui Wang. Providing Item-side Individual Fairness for Deep Recommender Systems. In *ACM International Conference Proceeding Series*, volume 22, pages 117–127. Association for Computing Machinery, 6 2022. ISBN 9781450393522. doi: 10.1145/3531146.3533079. URL https://doi.org/10.1145/3531146.3533079.

[228] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A Survey on the Fairness of Recommender Systems. *ACM Trans. Inf. Syst.*, 41(3), 2 2023. ISSN 1046-8188. doi: 10.1145/3547333. URL https://doi.org/10.1145/3547333.

[229] Yifan Wang, Peijie Sun, Weizhi Ma, Min Zhang, Yuan Zhang, Peng Jiang, and Shaoping Ma. Intersectional Two-sided Fairness in Recommendation. In *WWW 2024 - Proceedings of the ACM Web Conference*, pages 3609–3620. Association for Computing Machinery, Inc, 5 2024. ISBN 9798400701719. doi: 10.1145/3589334.3645518. URL https://dl.acm.org/doi/10.1145/3589334.3645518.

[230] William Webber, Alistair Moffat, Justin Zobel, and Tetsuya Sakai. Precision-at-ten considered redundant. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 695–696, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581644. doi: 10.1145/1390334.1390456. URL https://doi.org/10.1145/1390334.1390456.

[231] W. Weibull. A Statistical Distribution Function of Wide Applicability. *Journal of Applied Mechanics*, 1951.

[232] Yan Wen, Chen Gao, Lingling Yi, Liwei Qiu, Yaqing Wang, and Yong Li. Efficient and Joint Hyperparameter and Architecture Search for Collaborative Filtering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pages 2547–2558, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599322. URL https://doi.org/10.1145/3580305.3599322.

[233] Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. Fairness-aware news recommendation with decomposed adversarial learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4462–4469, May 2021. doi: 10.1609/aaai.v35i5.16573. URL https://ojs.aaai.org/index.php/AAAI/article/view/16573.

[234] Haolun Wu, Bhaskar Mitra, Chen Ma, Fernando Diaz, and Xue Liu. Joint Multisided Exposure Fairness for Recommendation. In *SIGIR 2022 - Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 703–714. Association for Computing Machinery, Inc, 7 2022. ISBN 9781450387323. doi: 10.1145/3477495.3532007. URL `https://doi.org/10.1145/3477495.3532007`.

[235] Kun Wu, Jacob Erickson, Wendy Hui Wang, and Yue Ning. Equipping Recommender Systems with Individual Fairness via Second-Order Proximity Embedding. In *Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '22, pages 171–175. IEEE Press, 2023. ISBN 9781665456616. doi: 10.1109/ASONAM55673.2022.10068703. URL `https://doi.org/10.1109/ASONAM55673.2022.10068703`.

[236] Yao Wu, Jian Cao, Guandong Xu, Yudong Tan, Yao Wu Jian Cao, Guandong Xu, Yudong Tan, Yao Wu, and Jian Cao. TFROM: A Two-sided Fairness-Aware Recommendation Model for Both Customers and Providers. In *SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1013–1022. Virtual Event, Canada, 7 2021. ISBN 9781450380379. doi: 10.1145/3404835.3462882. URL `https://doi.org/10.1145/3404835.3462882`.

[237] Yao Wu, Jian Cao, and Guandong Xu. Fairness in Recommender Systems: Evaluation Approaches and Assurance Strategies. *ACM Trans. Knowl. Discov. Data*, 18(1), 8 2023. ISSN 1556-4681. doi: 10.1145/3604558. URL `https://doi.org/10.1145/3604558`.

[238] Lin Xiao, Zhang Min, Zhang Yongfeng, Gu Zhaoquan, Liu Yiqun, and Ma Shaoping. Fairness-Aware Group Recommendation with Pareto-Efficiency. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17, pages 107–115, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346528. doi: 10.1145/3109859.3109887. URL `https://doi.org/10.1145/3109859.3109887`.

[239] 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`.

[240] Chen Xu, Wenjie Wang, Yuxin Li, Liang Pang, Jun Xu, and Tat-Seng Chua. A Study of Implicit Ranking Unfairness in Large Language Models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7957–7970, Miami, Florida, USA, 11 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.findings-emnlp.467`.

[241] Dingqi Yang, Daqing Zhang, Zhiyong Yu, and Zhu Wang. A sentiment-enhanced personalized location recommendation system. In *HT 2013 - Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 119–128, 2013. ISBN 9781450319676. doi: 10.1145/2481492.2481505. URL `https://dl.acm.org/doi/10.1145/2481492.2481505`.

[242] Tao Yang, Zhichao Xu, Zhenduo Wang, and Qingyao Ai. FARA: Future-Aware Ranking Algorithm for Fairness Optimization. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pages 2906–2916, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701245. doi: 10.1145/3583780.3614877. URL `https://doi.org/10.1145/3583780.3614877`.

[243] Tao Yang, Zhichao Xu, Zhenduo Wang, Anh Tran, and Qingyao Ai. Marginal-Certainty-Aware Fair Ranking Algorithm. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, pages 24–32, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394079. doi: 10.1145/3539597.3570474. URL `https://doi.org/10.1145/3539597.3570474`.

[244] Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/e6384711491713d29bc63fc5eeb5ba4f-Paper.pdf`.

[245] Hyunsik Yoo, Zhichen Zeng, Jian Kang, Ruizhong Qiu, David Zhou, Zhining Liu, Fei Wang, Charlie Xu, Eunice Chan, and Hanghang Tong. Ensuring user-side fairness in dynamic recommender systems. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 3667–3678, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645536. URL `https://doi.org/10.1145/3589334.3645536`.

[246] Junliang Yu, Jay Thai Duong Huynh, Shaoyang Fan, Gianluca Demartini, Tong Chen, Hongzhi Yin, and Shazia Sadiq. Biasnavi: Llm-empowered data bias management. In *Companion Proceedings of the 2025 ACM Web Conference (TheWebConf 2025) - Demo Track (in press)*, April 2025.

[247] Guanghu Yuan, Fajie Yuan, Yudong Li, Beibei Kong, Shujie Li, Lei Chen, Min Yang, Chenyun Yu, Bo Hu, Zang Li, Yu Xu, and Xiaohu Qi. Tenrec: a large-scale multipurpose benchmark dataset for recommender systems. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

[248] Eva Zangerle and Christine Bauer. Evaluating recommender systems: Survey and framework. *ACM Comput. Surv.*, 55(8), December 2022. ISSN 0360-0300. doi: 10.1145/3556536. URL `https://doi.org/10.1145/3556536`.

[249] Meike Zehlike and Carlos Castillo. Reducing Disparate Exposure in Ranking: A Learning To Rank Approach. In *Proceedings of The Web Conference 2020*, WWW '20, pages 2849–2855, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366424.3380048. URL `https://doi.org/10.1145/3366424.3380048`.

[250] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. FA∗IR: A fair top-k ranking algorithm. *International Conference on Information and Knowledge Management, Proceedings*, Part F1318:1569–1578, 2017. doi: 10.1145/3132847.3132938.

[251] Meike Zehlike, Tom Sühr, Ricardo Baeza-Yates, Francesco Bonchi, Carlos Castillo, and Sara Hajian. Fair top-k ranking with multiple protected groups. *Information Processing & Management*, 59(1):102707, 2022. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.102707. URL `https://www.sciencedirect.com/science/article/pii/S0306457321001916`.

[252] Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in Ranking, Part II: Learning-to-Rank and Recommender Systems. *ACM Computing Surveys*, 55 (6), 12 2022. ISSN 15577341. doi: 10.1145/3533380. URL `https://dl.acm.org/doi/10.1145/3533380`.

[253] Huimin Zeng, Zhankui He, Zhenrui Yue, Julian McAuley, and Dong Wang. Fair Sequential Recommendation without User Demographics. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in*

Information Retrieval, SIGIR '24, pages 395–404, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657703. URL https://doi.org/10.1145/3626772.3657703.

[254] Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tr0KidwPLc.

[255] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 993–999, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702419. doi: 10.1145/3604915.3608860. URL https://doi.org/10.1145/3604915.3608860.

[256] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Comput. Surv.*, 52 (1), February 2019. ISSN 0360-0300. doi: 10.1145/3285029. URL https://doi.org/10.1145/3285029.

[257] Xiaoyu Zhang, Yishan Li, Jiayin Wang, Bowen Sun, Weizhi Ma, Peijie Sun, and Min Zhang. Large language models as evaluators for recommendation explanations. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 33–42, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705052. doi: 10.1145/3640457.3688075. URL https://doi.org/10.1145/3640457.3688075.

[258] Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. How do large language models capture the ever-changing world knowledge? a review of recent advances. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8289–8311, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.516. URL https://aclanthology.org/2023.emnlp-main.516/.

[259] Kesen Zhao, Shuchang Liu, Qingpeng Cai, Xiangyu Zhao, Ziru Liu, Dong Zheng, Peng Jiang, and Kun Gai. KuaiSim: a comprehensive simulator for recommender systems. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[260] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji Rong Wen. RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms. In *International Conference on Information and Knowledge Management, Proceedings*, pages 4653–4664, New York, NY, USA, 2021. ACM. ISBN 9781450384469. doi: 10.1145/3459637.3482016. URL https://doi.org/10.1145/3459637.3482016.

[261] Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu C. Aggarwal, and Tyler Derr. Fairness and diversity in recommender systems: A survey. *ACM Trans. Intell. Syst. Technol.*, 16(1), January 2025. ISSN 2157-6904. doi: 10.1145/3664928. URL https://doi.org/10.1145/3664928.

[262] Yong Zheng and David (Xuejun) Wang. A survey of recommender systems with multi-objective optimization. *Neurocomputing*, 474:141–153, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.11.041. URL https://www.sciencedirect.com/science/article/pii/S0925231221017185.

[263] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

[264] Qiliang Zhu, Qibo Sun, Zengxiang Li, and Shangguang Wang. FARM: A Fairness-Aware Recommendation Method for High Visibility and Low Visibility Mobile APPs. *IEEE Access*, 8:122747–122756, 2020. ISSN 21693536. doi: 10.1109/ACCESS.2020.3007617.

[265] 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 '18, pages 1153–1162, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3271795. URL https://doi.org/10.1145/3269206.3271795.

[266] Ziwei Zhu, Jianling Wang, and James Caverlee. Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems. In *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 449–458. ACM, 7 2020. ISBN 9781450380164. doi: 10.1145/3397271.3401177. URL https://doi.org/10.1145/3397271.3401177.

[267] Ziwei Zhu, Yun He, Xing Zhao, Yin Zhang, Jianling Wang, and James Caverlee. Popularity-Opportunity Bias in Collaborative Filtering. In *WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 85–93. Association for Computing Machinery, Inc, 8 2021. ISBN 9781450382977. doi: 10.1145/3437963.3441820. URL https://doi.org/10.1145/3437963.3441820.

[268] Ziwei Zhu, Jingu Kim, Trung Nguyen, Aish Fenton, and James Caverlee. Fairness among New Items in Cold Start Recommender Systems. In *SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 767–776. Association for Computing Machinery, Inc, 7 2021. ISBN 9781450380379. doi: 10.1145/3404835.3462948. URL https://doi.org/10.1145/3404835.3462948.

[269] 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, pages 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.