UNDER REVIEW

RESCIENCE C

Replication / ML Reproducibility Challenge 2022

[Re] Reproducibility study of Joint Multisided Exposure Fairness for Recommendation

Alessia Hu[1], Oline Ranum[1], Chrysoula Pozrikidou[1], and Miranda Zhou[1]

[1]University of Amsterdam, Amsterdam, Netherlands

Edited by
(Editor)

Received
–

Published
–

DOI
–

Reproducibility Summary

Scope of Reproducibility — This study focuses on investigating the reproducibility of *Joint Multisided Exposure Fairness (JME) for Recommendation* by Wu et al[1]. Our objective is to verify the following claims suggested by the paper: (i) each of the proposed exposure fairness metrics quantifies a different notion of unfairness, (ii) for each of the proposed metrics there exists a disparity-relevance trade-off, and (iii) recommender systems can be optimized towards multiple fairness goals using JME-fairness measures.

Methodology — We modify and extend upon the open-source implementation of the pipeline, published by the authors on GitHub [2]. Our adjustments include restructuring the codebase, adding experimental setup files, and removing several bugs. We run the experiments on a RTX 3070 GPU, at a reproducibility cost of 44.5 GPU hours.

Results — We successfully reproduce the major trends of the core results, although some numerical deviations occur. We are able of providing support to two out of three claims. However, due to insufficient documentation and resources, we were unable to verify the paper's third claim. We conclude that in order to determine the fairness of a recommender system, considering different fairness dimensions with a multi-stakeholder perspective is essential.

What was easy — The JME-fairness metrics proposed in the paper are well-explained and fairly intuitive. Even without a background in fairness in AI and recommender systems, we were able to follow the pipeline and the main ideas presented.

What was difficult — Details regarding the setup of the experiments are missing from the original codebase, and documentation is limited. In addition, for the reproduction of their third claim, familiarity with topics not analyzed in the paper is required.

Communication with original authors — Per request by email, the authors provided some clarifications regarding experimental setups and calculations performed in the experiments of the original paper. We received a response that answered part of our questions, and a reference to a GitHub repository [3] which is potentially suitable for demonstrating optimization with a JME-fairness loss.

Copyright © 2023 A. Hu et al., released under a Creative Commons Attribution 4.0 International license.
Correspondence should be addressed to Alessia Hu (alessia.hu@student.uva.nl)
The authors have declared that no competing interests exist.
Code is available at https://github.com/OlineRanum/FACT.git. – SWH swh:1:dir:ddaee99fffaa5becad67496efe000ca6c47341c7.
Open peer review is available at https://openreview.net/forum?id=A0Sjs3IJWb-.

1 Introduction

Information retrieval (IR) systems, such as search and recommendation algorithms, mediate exposure and consumption of online information. Traditional IR systems are built on ranking algorithms that maximize utility. However, as recent literature [4] [5] [6] suggests, the fairness and biases of ranking algorithms should also be jointly considered.

Metrics and methodologies for optimizations are increasingly developed in order to measure and improve fairness in algorithms [7] [8] [1]. Precursory research has predominantly focused on the fairness of exposure to users [4]. However, as has been argued [7] [8] group attributes on both consumer and producer sides should be jointly considered to accommodate fairness for all stakeholders. Hence, Wu et al.[1] introduced a family of six fairness metrics to accommodate a Joint Multisided Exposure fairness perspective.

2 Scope of reproducibility

In this report, we evaluate the reproducibility of the proposed JME-fairness metric properties. The metric family extends upon the concept of expected exposure as the expectation of a user browsing model provided a stochastic ranking policy [4]. If the reader is unfamiliar with the aforementioned concepts we invite them to consult section 3.1. In our understanding, the main claims of the original paper can be summarized as follows:

1. Stochasticity impacts the JME-fairness metrics and their corresponding disparity and relevance components, in the sense that there exists a disparity-relevance trade-off for each JME-fairness metric.
2. Each JME-fairness metric quantifies a different notion of unfairness: a system that performs well on one fairness dimension can be suboptimal for another.
3. Recommender systems can be optimized towards a specific fairness goal, based on different combinations of JME-fairness metrics

3 Methodology

In the subsequent section, we provide fundamental concepts and definitions that were adopted by Wu et al. [1] and information on the design of our experiments.

3.1 Browsing models, stochastic ranking and exposure

User browsing models estimate the probability of exposure of an item d in a retrieved ranked list of items σ. The browsing model used in this paper is the rank-biased precision (RBP) metric [9] which presumes that the probability of the exposure event ϵ for d depends only on its rank $\rho_{d,\sigma}$ in a retrieved ranked list σ and decreases exponentially as:

$$p(\epsilon|d, \pi_u) = \gamma^{(\rho_{d,\sigma}-1)}$$

The patience factor γ determines how far down the ranking the user is likely to explore.

A stochastic ranking policy π is a probability distribution that covers all item arrangements in the collection [4]. Given such a policy π_u conditioned on user $u \in \mathcal{U}$, the expected value of the probability that an item $d \in \mathcal{D}$ is exposed to the user is:

$$p(\epsilon|d, \pi u) = \mathbb{E}_{\sigma\sim\pi_u}[p(\epsilon|d, \sigma)] \tag{1}$$

Metric
II-F = $\frac{1}{\lvert D\rvert}\frac{1}{\lvert U\rvert}\sum_{j=1}^{\lvert D\rvert}\sum_{i=1}^{\lvert U\rvert}(E_{ij}-E^*_{ij})^2$
IG-F = $\frac{1}{\lvert G_d\rvert}\frac{1}{\lvert U\rvert}\sum_{D\in G_d}\sum_{i=1}^{\lvert U\rvert}(\sum_{j=1}^{\lvert D\rvert}p(D_j\lvert D)(E_{ij}-E^*_{ij}))^2$
GI-F = $\frac{1}{\lvert D\rvert}\frac{1}{\lvert G_u\rvert}\sum_{j=1}^{\lvert D\rvert}\sum_{U\in G_u}(\sum_{i=1}^{\lvert U\rvert}p(U_i\lvert U)(E_{ij}-E^*_{ij}))^2$
GG-F = $\frac{1}{\lvert G_d\rvert}\frac{1}{\lvert G_u\rvert}\sum_{D\in G_d}\sum_{U\in G_u}(\sum_{j=1}^{\lvert D\rvert}\sum_{i=1}^{\lvert U\rvert}p(D_j\lvert D)p(U_i\lvert U)(E_{ij}-E^*_{ij}))^2$
AI-F = $\sum_{j=1}^{\lvert D\rvert}(\sum_{i=1}^{\lvert U\rvert}p(U_i)(E_{ij}-E^*_{ij}))^2$
AG-F = $\frac{1}{\lvert G_d\rvert}\sum_{D\in G_d}(\sum_{j=1}^{\lvert D\rvert}\sum_{i=1}^{\lvert U\rvert}p(D_j\lvert D)p(U_i)(E_{ij}-E^*_{ij}))^2$

Table 1. The formal mathematical definition of the JME-fairness metrics

The authors refer to $\mathrm{E} \in \mathbb{R}^{|\mathcal{U}|x|\mathcal{D}|}$ as the expected exposure matrix, such that $\mathrm{E}_{ij} = p(\epsilon|\mathcal{D}_j, \pi_{\mathcal{U}_i})$. The three expected exposure matrices E defined by the authors are *system exposure E*, the expected exposure corresponding to a stochastic ranking policy π as determined by a retrieval system; *target exposure* E^*, the expected exposure corresponding to an ideal stochastic ranking policy π^* (e.g., the equal expected exposure principle [4]); *random exposure* $E^{\sim}$, the expected exposure corresponding to a stochastic ranking policy $\pi^{\sim}$ defined by a uniformly random distribution over all item permutations.

3.2 Metric definitions, decomposition and method for metric analysis

The formal mathematical definitions for the metrics proposed in the paper can be seen in Table 1 [1]. Each of the metrics represents a different fairness concern measuring the deviation between system exposure and target (ideal) exposure. The first dimension of each metric refers to the users while the second refers to the items. The letters "I", "G" and "A" correspond respectively to individual users or items, groups of individuals or items, and all individuals. For all six metrics, to have a fairer recommendation system, lower values are more desirable.
Each proposed metric can be decomposed into three components: a disparity and a relevance component and a system-independent constant. This allows one to study the trade-off between disparity and relevance, while different degrees of stochasticity are introduced into the model. The use of a static ranking model with all relevant items at the top maximizes relevance and a fully stochastic model minimizes disparity. The decomposition for the JME-fairness metrics can be seen in Appendix A.

A method to generate stochastic ranking policies with varying levels of stochasticity is introduced, in order to examine their ability to distribute item exposure. The Plackett-Luce (PL) model [10] [11] is utilized to produce multiple rankings by sampling from the estimated relevance scores of items for a user, given a deterministic ranker. The model is based on Luce's axiom that the probability of choosing one item over another does not depend on the set of items from which the choice is made [10] [12] and creates a ranking by repeatedly selecting items without replacement from the collection with probability distribution defined as:

$$p(d|u) = \frac{exp(Y_{d,u}/\beta)}{\sum_{d'\in D} exp(Y_{d',u}/\beta)}$$

The parameter β controls the level of stochasticity of the ranking model. A higher value corresponds to more stochasticity, while a lower value indicates a more deterministic

ranking policy. $Y_{d,u}$ represents the relevance score for item d for user u, estimated by the deterministic ranker.

3.3 Datasets

The original work conducts its primary experiments on the MovieLens1M [13] dataset comprised of movie reviews. In addition, the MovieLens100k dataset is considered for the optimization task. However, the current work disregards the MovieLens100k dataset as we were not able to reproduce the optimization experiment. To conduct further experimentation regarding the claims within the original paper, we also consider the LibraryThing [14][15] dataset comprised of book reviews.

MovieLens1M (ml-1m) consists of 1,000,209 numerical ratings ranging from 1-5. The reviews are supplied by 6,040 MovieLens users on 3,706 movies. The available metadata for the users includes 2 gender groups, 7 age groups, 21 occupation groups, and zip codes. The metadata for the movies comprises 19 genre groups (including unknown). Each user in the dataset has written at least 20 reviews.

LibraryThing (lt) Due to computational limitations, we extract a subset of the LibraryThing dataset comprised of 702,522 reviews provided by 12,976 LibraryThing users on 325,075 books. The selection is made so that each user has written at least 20 reviews and each item is reviewed at least 3 times.

3.4 Hyperparameters

In order to reproduce the results of the original paper we set the hyperparameters equal to the default parameters found in the original codebase. The values used to perform metric analysis on both datasets are listed in Table 2.

Dataset	γ	User Group Attributes	Item Group Attribute
MovieLens1M	0.8	gender, age, occupation	movie genre
LibraryThing	0.8	helpfulness of users' ratings	engagement rate

Table 2. Hyperparameter values for the metric analysis experiments

In particular, we introduce 8 degrees of stochasticity with β ={8,4,3,2,1,0.5,0.25,0.125,ST}, where ST refers to a fully static deterministic model. While the MovieLens dataset supplies group attributes such as age, gender, and occupation, the LibraryThing dataset only provides an evaluation rate of the helpfulness of each user's ratings and a comment per movie. As such, we construct an 'engagement' group attribute for each item derived as the average length of its received comments. Furthermore, the user attribute is built by calculating the average number of helpfulness votes received by each user.

For the Bert4Rec model, we initialize the embedding dimension for the hidden layers to 128, with the length of BERT embeddings being at a maximum of 100 items. The learning rate is set to 0.001.

3.5 Experimental setup and code

Code — An open-source implementation of code associated with the original paper was published by the authors on GitHub. Unfortunately, we experienced several conflicts when we attempted to deploy the code. In particular, no code was provided to run the baseline experiments of the original paper, the hyperparameters used in each experiment were not specified and dependencies were not always clarified. Due to insufficient compartmentalization and documentation, the code was generally hard to understand.

A full account of the issues we encountered is provided in Appendix **D**. In order to ease future research and extensions utilizing the code, we provide a restructured codebase with a higher degree of documentation, supplied with additional postprocessing and configuration tools for the experiments.

Pipeline — Identical to the original pipeline we employ the RBP user browsing model. Stochastic ranking policies with stochasticity β are generated over a set of trained deterministic ranking models for the MovieLens1M and the LibraryThing dataset. The pre-trained models are supplied by Valcarce et al. [16] and were made available through GitHub. The fairness metrics are then computed utilizing the expected exposure of equation 1, the formal fairness definitions of section 3.2, and their decomposition as described in section 3.2.

Claim 1 — In order to verify claim 1 we estimate the fairness metrics across 8 different levels of stochasticity β and reproduce Figures 6 and 7 as well as Table 6, included from the original paper. While the authors mentioned that they applied min-max normalization, they did however not state how the normalization was applied. As such, we made the following assumptions based on the min-max values of the original figures: to obtain Figure 6, we perform a min-max normalization over the estimated metrics for each metric dimension and for each component separately, across all stochasticity values. We perform the same analysis across all 21 pre-trained ranking models available in the repository. To obtain Figure 2 we perform a normalization per metric component across the models *BPRMF, LDA, PureSVD, SLIM* and *WRMF*. With regards to Table 6, the authors do not clearly state how the area under the curve (AUC) is calculated. Initially, we did this using the trapezoidal rule. However, after consulting the authors it was made clear that the AUC was calculated for each curve only up until the disparity value which corresponds to the smallest stochastic value across all models. We reproduce Table 3 to quantify the disparity-relevance tradeoff across all five models and calculate the average relative difference between our results and their results.

Claim 2 — In order to verify claim 2 we compute the Kendall rank correlation [17] between the different metrics and their relevance and disparity components across all six proposed fairness dimensions on the pre-trained models. The authors of the original paper state that they evaluate 15 different pre-trained deterministic rankers on the MovieLens1M dataset, for 8 different levels of stochasticity. However, as they do not specify which models are used and the codebase provides 21 pre-trained models, we perform the analysis over all of them. The analysis is performed using gender and age as user-side group attributes.

Claim 3 — The authors provided no codebase for the third claim, but they did provide a general algorithm and proposition for a JME-fairness loss that could be used in a training situation. From there, we attempted to train a BPRMF model with the proposed loss. As we were not able to reproduce their results, its methodology is not elaborated on in this report. However, we advise the curious reader to consult section 6 of the original paper for further information.

Beyond the paper: Experiment I — NN-based approaches to IR have witnessed an explosive growth in recent years [18] [19] [20], yielding state-of-the-art ranking methodologies such as the Bert4Rec model [21]. As such, it is becoming increasingly important to develop scalable fairness frameworks that can mediate multisided exposure fairness for all stakeholders. We extend upon the work by Wu et al. by demonstrating the application of the JME-fairness metric to a neural model and evaluating if their claims are still valid.

We train a Bert4Rec model from scratch and apply it to both the ml-1m and the lt datasets. Bert4Rec is a recommender that is a combination of two SOTA models: the BERT language model and SASRec recommender system. Using the bidirectional ability of BERT on top of a sequential recommender allows the model to capture more complex user behaviors. For the BERT4Rec model [21], we adopt a PyTorch implementation available from the TorchRec library [22]. Both datasets are split into a training, validation, and test set based on leave-one-out, where the second-last and last reviewed items of a user are reserved for validation and testing respectively. We train 10 epochs on the Movielens dataset and run the JME metrics on the obtained predictions as in the original paper, including different levels of stochasticity. However, even after our modification, the LibraryThing dataset still exhibits high sparsity. Therefore, achieving 10 epochs on this particular dataset is not computationally feasible with our current resources.

Beyond the paper: Experiment II — In the original experiments, age and gender are used as group attributes on the user side, we perform the analysis with the Kendall rank correlation with a different attribute: *occupation*. The attributes used originally present very few groups which could be the cause for the high correlation between group-related metrics (GG-F AG-F) in Figure 3. We perform this analysis to investigate whether the same trends are consistent when considering more groups.

Beyond the paper: Experiment III — To assess the validity of the author's choice for the hyperparameter value $gamma = (0.8)$, we conducted an additional parametersweep through assigning different values for $\gamma = [0.01, 0.2, 0.4, 0.6, 0.8, 0.9]$ and analyzing the behavior of the JME fairness metric on a BPRMF model.
All experiments presented in this document can be reproduced by using the code provided in this GitHub repository. In the README section we provide instructions on how to obtain the results for the different experiments.

3.6 Computational requirements

The experiments are conducted on a local machine with the GPU model GeForce RTX 3070 equipped. The total calculated computational costs for running the metrics, with and without a stochastic policy, can be seen in Appendix **G**.

4 Results

In this section, we present the results of our reproducibility study and of our extended research. We include the results of the original paper in Appendix **E** for the readers' convenience and completion.

4.1 Results reproducing original paper

Result 1 — Our reproduction of the JME-fairness stochasticity impact study is presented in Figure 1, the disparity-relevance trade-off is presented in Figure 2 and the AUC estimates in Table 3. Larger stochasticity implies a more fair system, governed by small relevance and disparity. At a qualitative level, we successfully reproduced the trends observed in Figure 6 of the original paper. However, when considering Figure 2 and Table 3 we observe that the relationships between the performance of the models deviate from the relationships found in the original paper. For instance, across the II-dimension, the trends of SLIM and LDA are flipped with respect to Wu et al. When considering the AUC we observe the same deviations in trends from the original paper and find that on average, across all models, our results deviate with $25 \pm 6\%$.

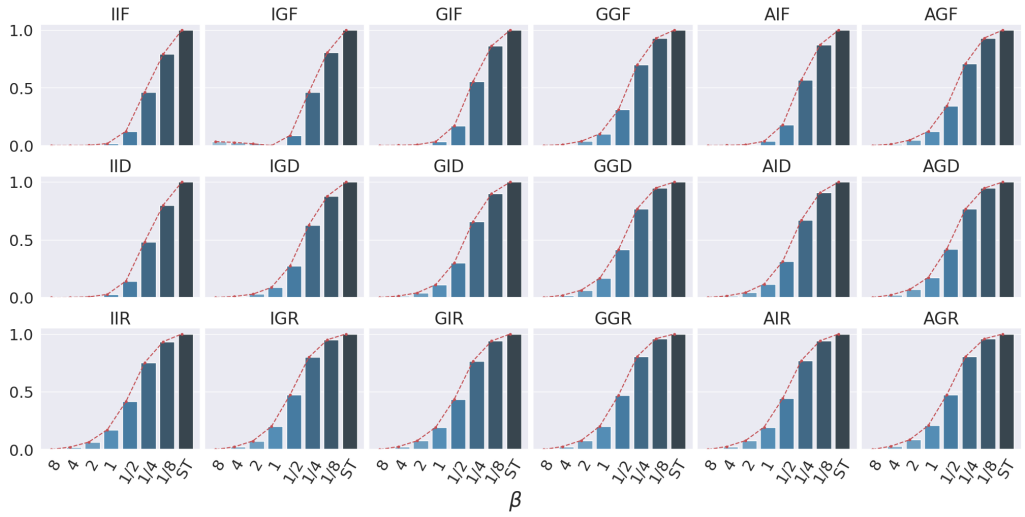


Figure 1. Behaviour of JME-fairness metrics for a stochastic ranking policy. Bars are produced by randomizing the BPRMF model using the Plackett-Luce model on MovieLens1M dataset. First, second and third rows correspond respectively to the impact of different stochasticity on the overall fairness, disparity, and relevance components. The x-axis shows the values of β, where a larger value indicates more randomization

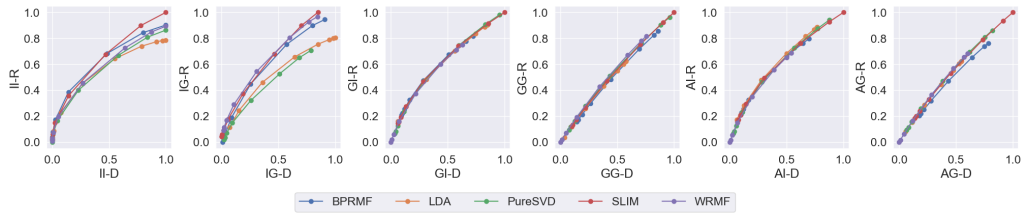


Figure 2. Disparity-relevance trade-off curves across six fairness dimensions with 8 levels of stochasticity for 5 recommendation models

.

Result 2 – To verify claim 2 we reproduce the Kendall rank correlations and present them in Figure 3. We observe that the trends of our results follow similar patterns as the ones observed by Wu et al. Our cross-metric analysis demonstrates a relatively low correlation between II-F and all other metrics. Additionally, for both group attributes it can be seen that there is a high correlation between the GG and AG metric, due to the low number of groups.

Result 3 – In our attempts to optimize a BPRMF model with the fairness loss we encounter vanishing gradient problems, and we are not currently able to make the loss converge. As such, we are not able to reproduce claim 3.

Model	II	IG	GI	GG	AI	AG	Avg. rel. diff.
BPRMF	0.6336	0.437	0.1949	0.2487	0.1879	0.0869	0.26
LDA	0.5567	0.2448	0.2305	0.1877	0.2007	0.1799	0.28
PureSVD	0.5691	0.3239	0.2615	0.2896	0.0712	0.0909	0.30
SLIM	**0.6540**	0.3757	0.2591	0.2739	0.0799	0.1145	0.27
WRMF	0.5901	**0.5171**	**0.2826**	**0.334**	**0.2194**	**0.2219**	0.13
Bert4Rec	0.7325	0.6331	0.7495	0.6229	0.7626	0.6026	

Table 3. The AUC of the disparity-relevance trade-off curves for 5 pre-trained models and Bert4Rec across 6 fairness dimensions. Highest and second highest value of each column is marked bold and underlined. Rightmost column shows the average relative difference with respect to the AUC calculated with our method.

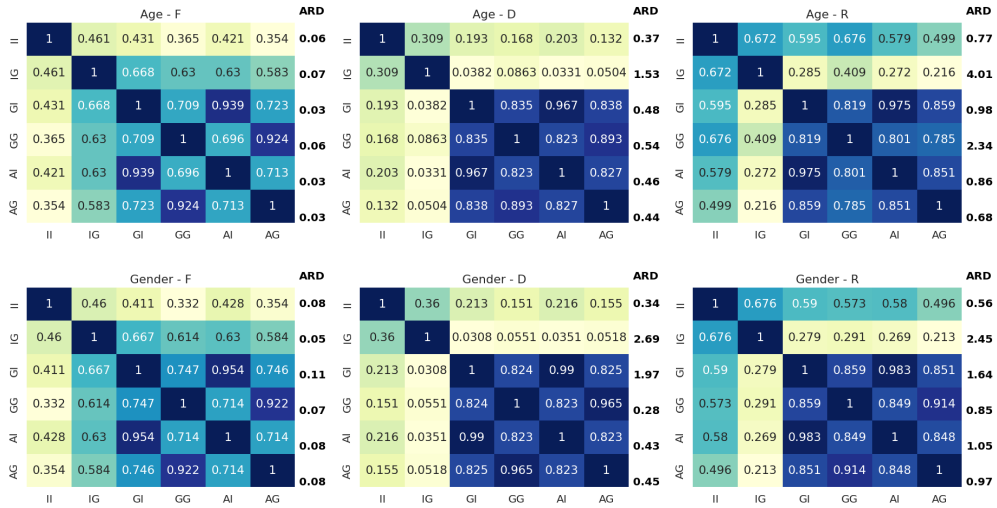


Figure 3. The Kendall rank correlation between the six JMEF metrics and their two components. The first, and second rows correspond to age, and gender user attributes. From the left to the right column, we respectively find the heatmaps for the fairness, disparity, and relevance components.

4.2 Results beyond original paper

Additional Result 1 — The behavioral results of the JME metrics on a neural model can be observed in Figure 4, and the AUC of the ml-1m dataset is presented in Table 3. For the ml-1m dataset, we observe that the metrics generally decrease as β increases. From this, we confirm the presence of a disparity-relevance trade-off, since the stochastic ranking policy still fails to achieve low disparity and high relevance concurrently. In addition, the AUC indicates that the trade-off is generally better than the ones of the deterministic models. Noticeably, the static ranking policy tends to show poor values compared to its stochastic variant when we consider user groupings. As aforementioned, we did not succeed in producing a complete model trained on the LibraryThing dataset due to computational and time limitations. Nonetheless, we chose to include a preliminary evaluation on the 5th epoch in Figure 4 as a demonstration of the metrics on a weak recommender system.

Additional Result 2 — The Kendall rank correlation for the occupation groups can be observed in Appendix C. Although the calculation is computed across a larger number of groups, the correlation between GG-F and AG-F is still high, leading to the conclusion that even the considered 21 occupational groups are too few to lessen their correlation. The results are presented in the Appendix as they are trivial.

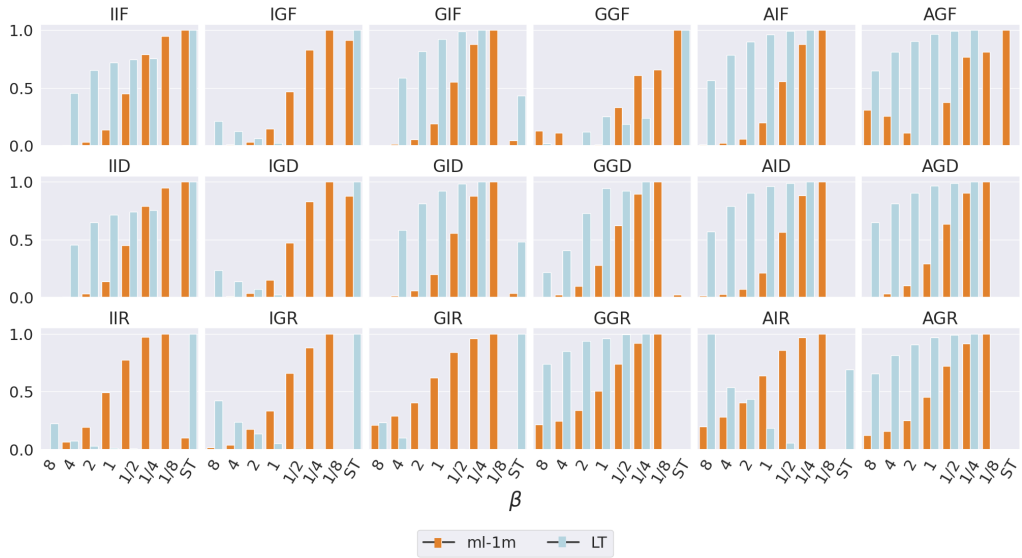


Figure 4. Comparison of Bert4Rec metrics' and disparity and relevance components on MovieLens1M and LibraryThing

Additional Result 3 — The results of our parametersweep can be found in the Appendix **F**. Considering the definition of γ as the patience factor, it describes the probability of an user progressing to the next item in the ranked list [9]. In our case, a low γ value would indicate a smaller set of items being exposed to users, resulting in lower fairness due to limited diversification. From our results, the behavior of the JME fairness coincides with this assumption, assigning higher fairnesses to higher patience factors, whereas a lower values are scored lower. As for why 0.8 is specifically chosen by the authors can be derived from the work of A. Moffat and J. Zobel(2008)[9], where values between 0.5 and 0.8 have been assessed to be the most stable for rank-biased precision.

5 Discussion

In this work, we found that our reproduction study could provide support for two out of three claims, with minor numerical discrepancies. The inconsistencies could be due to the use of different hyperparameters, for instance, the seed as these were not explicitly clarified in the original paper. In Figure 3 we observed varying degrees of numerical deviations. A plausible reason is that we utilized different data in taking into consideration all 21 models. Regarding the experiments on the neural model, we observed behavior across several dimensions deviating from the expected trend posed by the deterministic rankers. This is presumably due to insufficient training but does not exclude the possibility of flaws in the model's inherent properties, which we recommend to be further investigated. It would demonstrate that even if our model achieves favorable results on standard evaluation scores, additional analyses are to be considered with regard to the model's task.

5.1 What was easy

The explanation of each of the proposed JME-fairness metrics was very well-written and easy to follow even for someone with no previous background in the field of fairness in AI. Additionally, the relationship between the metrics and their decomposition was clearly explained and accompanied by examples allowing for comprehension even without very advanced mathematical knowledge.

5.2 What was difficult

The code provided in the original repository was complex and hard to interpret as it was lacking adequate documentation. Large parts of the code required to recreate the original experiments were missing. Additionally, no environment was provided thus, we had to make some assumptions about the requirements to run the experiments and create one. Finally, as there was no publicly available code to optimize for JME-fairness, deeper knowledge regarding the mathematical concepts behind the algorithm was required to implement it from scratch.

5.3 Communication with original authors

We reached out to the authors to get clarifications on several matters. We asked for clarifications regarding the calculation of the AUC of Table 3, as the methods that seemed intuitively correct produced different results from the ones in the original paper. Furthermore, we requested more details regarding the setup for the experiments for claim 3 as they voluntarily did not include their implementation in their repository. They provided us with their way to calculate the AUC, with some indications on the pipeline used to support claim 3 and with a possibly helpful external paper.

References

1. H. Wu, B. Mitra, C. Ma, F. Diaz, and X. Liu. **Joint Multisided Exposure Fairness for Recommendation**. 2022. DOI: 10.48550/ARXIV.2205.00048. URL: https://arxiv.org/abs/2205.00048.
2. H. Wu. **JMEFairness**. https://github.com/haolun-wu/jmefairness. 2022.
3. username: latataro. **BPRMF**. Version 2.0.4. Jan. 2022. URL: https://github.com/jchanxtarov/bprmf.
4. F. Diaz, B. Mitra, M. D. Ekstrand, A. J. Biega, and B. Carterette. "Evaluating Stochastic Rankings with Expected Exposure." In: **CoRR** abs/2004.13157 (2020). arXiv:2004.13157. URL: https://arxiv.org/abs/2004.13157.
5. M. Kay, C. Matuszek, and S. Munson. "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations." In: (Apr. 2015). DOI: 10.1145/2702123.2702520.
6. A. Singh and T. Joachims. "Fairness of Exposure in Rankings." In: (June 2018). DOI: 10.1145/3219819.3220088. URL: https://doi.org/10.1145%5C%2F3219819.3220088.
7. R. Burke. "Multisided Fairness for Recommendation." In: **CoRR** abs/1707.00093 (2017). arXiv:1707.00093. URL: http://arxiv.org/abs/1707.00093.
8. M. D. Ekstrand, A. Das, R. Burke, and F. Diaz. "Fairness and Discrimination in Information Access Systems." In: **CoRR** abs/2105.05779 (2021). arXiv:2105.05779. URL: https://arxiv.org/abs/2105.05779.
9. A. Moffat and J. Zobel. "Rank-Biased Precision for Measurement of Retrieval Effectiveness." In: **ACM Trans. Inf. Syst.** 27.1 (Dec. 2008). DOI: 10.1145/1416950.1416952. URL: https://doi.org/10.1145/1416950.1416952.
10. R. D. Luce. **Individual Choice Behavior: A Theoretical analysis**. New York, NY, USA: Wiley, 1959.
11. R. L. Plackett. "The Analysis of Permutations." In: **Journal of the Royal Statistical Society. Series C (Applied Statistics)** 24.2 (1975), pp. 193–202. URL: http://www.jstor.org/stable/2346567 (visited on 01/26/2023).
12. R. D. Luce. "The Choice Axiom after Twenty Years." In: **Journal of Mathematical Psychology** 15 (1977), pp. 215–233.
13. F. M. Harper and J. A. Konstan. "The MovieLens Datasets: History and Context." In: **ACM Trans. Interact. Intell. Syst.** 5.4 (Dec. 2015). DOI: 10.1145/2827872. URL: https://doi.org/10.1145/2827872.
14. C. Cai, R. He, and J. McAuley. "SPMC: Socially-Aware Personalized Markov Chains for Sparse Sequential Recommendation." In: Aug. 2017, pp. 1476–1482. DOI: 10.24963/ijcai.2017/204.
15. T. Zhao, J. McAuley, and I. King. "Improving Latent Factor Models via Personalized Feature Projection for One Class Recommendation." In: Oct. 2015, pp. 821–830. DOI: 10.1145/2806416.2806511.
16. D. Valcarce, A. Bellogín, J. Parapar, and P. Castells. "On the Robustness and Discriminative Power of Information Retrieval Metrics for Top-N Recommendation." In: **Proceedings of the 12th ACM Conference on Recommender Systems**. RecSys '18. Vancouver, British Columbia, Canada: Association for Computing Machinery, 2018, pp. 260–268. DOI: 10.1145/3240323.3240347. URL: https://doi.org/10.1145/3240323.3240347.
17. M. G. Kendall. "Rank Correlation Methods." In: (1949).
18. T. Kenter, A. Borisov, C. V. Gysel, M. Dehghani, M. de Rijke, and B. Mitra. "Neural Networks for Information Retrieval." In: **CoRR** abs/1801.02178 (2018). arXiv:1801.02178. URL: http://arxiv.org/abs/1801.02178.
19. B. Mitra and N. Craswell. **Neural Models for Information Retrieval**. May 2017. URL: https://www.microsoft.com/en-us/research/publication/neural-models-information-retrieval/.
20. K. D. Onal et al. "Neural information retrieval: at the end of the early years." In: **Information Retrieval Journal** 21 (2018), pp. 111–182.
21. F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, and P. Jiang. "BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer." In: **CoRR** abs/1904.06690 (2019). arXiv:1904.06690. URL: http://arxiv.org/abs/1904.06690.
22. F. Sun. **TorchRec Bert4Rec**. https://github.com/pytorch/torchrec/tree/main/examples/bert4rec. 2022.

6 Appendix

A Table of decomposed metrics

Disparity		
II-D	=	$\frac{1}{\lvert\mathcal{D}\rvert}\frac{1}{\lvert\mathcal{U}\rvert}\sum_{j=1}^{\lvert\mathcal{D}\rvert}\sum_{i=1}^{\lvert\mathcal{U}\rvert}\mathrm{E}_{ij}^{\delta_{ij}^2}$
IG-D	=	$\frac{1}{\lvert\mathcal{G}_d\rvert}\frac{1}{\lvert\mathcal{U}\rvert}\sum_{D\in\mathcal{G}_d}\sum_{i=1}^{\lvert\mathcal{U}\rvert}\left(\sum_{j=1}^{\lvert D\rvert}p\left(D_j \mid D\right)\mathrm{E}^{\delta}{}_{ij}\right)^2$
GI-D	=	$\frac{1}{\lvert\mathcal{D}\rvert}\frac{1}{\lvert\mathcal{G}_u\rvert}\sum_{j=1}^{\lvert\mathcal{D}\rvert}\sum_{U\in\mathcal{G}_u}\left(\sum_{i=1}^{\lvert U\rvert}p\left(U_i \mid U\right)\mathbf{E}^{\delta}{}_{ij}\right)^2$
GG-D	=	$\frac{1}{\lvert\mathcal{G}_d\rvert}\frac{1}{\lvert\mathcal{G}_u\rvert}\sum_{D\in\mathcal{G}_d}\sum_{U\in\mathcal{G}_u}\left(\sum_{j=1}^{\lvert D\rvert}\sum_{i=1}^{\lvert U\rvert}p\left(D_j \mid D\right)p\left(U_i \mid U\right)\mathbf{E}^{\delta}{}_{ij}\right)^2$
AI-D	=	$\sum_{j=1}^{\lvert\mathcal{D}\rvert}\left(\sum_{i=1}^{\lvert\mathcal{U}\rvert}p\left(\mathcal{U}_i\right)\mathrm{E}^{\delta}{}_{ij}\right)^2$
AG-D	=	$\frac{1}{\lvert\mathcal{G}_d\rvert}\sum_{D\in\mathcal{G}_d}\left(\sum_{j=1}^{\lvert D\rvert}\sum_{i=1}^{\lvert\mathcal{U}\rvert}p\left(D_j \mid D\right)p\left(\mathcal{U}_i\right)\mathrm{E}^{\delta}{}_{ij}\right)^2$
Relevance		
II-R	=	$\frac{1}{\lvert\mathcal{D}\rvert}\frac{1}{\lvert\mathcal{U}\rvert}\sum_{j=1}^{\lvert\mathcal{D}\rvert}\sum_{i=1}^{\lvert\mathcal{U}\rvert}2\mathrm{E}^{\delta}{}_{ij}\mathrm{E}^{\Delta}{}_{ij}$
IG-R	=	$\frac{1}{\lvert\mathcal{G}_d\rvert}\frac{1}{\lvert\mathcal{U}\rvert}\sum_{D\in\mathcal{G}_d}\sum_{i=1}^{\lvert\mathcal{U}\rvert}\left(\sum_{j=1}^{\lvert D\rvert}2p\left(D_j \mid D\right)\mathrm{E}^{\delta}{}_{ij}\mathrm{E}^{\Delta}{}_{ij}\right)^2$
GI-R	=	$\frac{1}{\lvert\mathcal{D}\rvert}\frac{1}{\lvert\mathcal{G}_u\rvert}\sum_{j=1}^{\lvert\mathcal{D}\rvert}\sum_{U\in\mathcal{G}_u}\left(\sum_{i=1}^{\lvert U\rvert}2p\left(U_i \mid U\right)\mathrm{E}^{\delta}{}_{ij}\mathrm{E}^{\Delta}{}_{ij}\right)^2$
GG-R	=	$\frac{1}{\lvert\mathcal{G}_d\rvert}\frac{1}{\lvert\mathcal{G}_u\rvert}\sum_{D\in\mathcal{G}_d}\sum_{U\in\mathcal{G}_u}\left(\sum_{j=1}^{\lvert D\rvert}\sum_{i=1}^{\lvert U\rvert}2p\left(D_j \mid D\right)p\left(U_i \mid U\right)\mathrm{E}^{\delta}{}_{ij}\mathrm{E}^{\Delta}{}_{ij}\right)^2$
AI-R	=	$\sum_{j=1}^{\lvert\mathcal{D}\rvert}\left(\sum_{i=1}^{\lvert\mathcal{U}\rvert}2p\left(\mathcal{U}_i\right)\mathrm{E}^{\delta}{}_{ij}\mathrm{E}^{\Delta}{}_{ij}\right)^2$
AG-R	=	$\frac{1}{\lvert\mathcal{G}_d\rvert}\sum_{D\in\mathcal{G}_d}\left(\sum_{j=1}^{\lvert D\rvert}\sum_{i=1}^{\lvert\mathcal{U}\rvert}2p\left(D_j \mid D\right)p\left(\mathcal{U}_i\right)\mathrm{E}^{\delta}{}_{ij}\mathrm{E}^{\Delta}{}_{ij}\right)^2$

Table 4. Decomposition of each JME-fairness metric into their disparity and relevance components, where $\mathrm{E}^{\delta}{}_{ij} = \mathrm{E}_{ij} - \mathrm{E}_{ij}^{\sim}$ and $\mathrm{E}^{\Delta}{}_{ij} = \mathrm{E}_{ij}^{*} - \mathrm{E}_{ij}^{\sim}$

B AUC table: Trapezoidal Rule

Model	II	IG	GI	GG	AI	AG
BPRMF	<u>0.6336</u>	**0.5378**	0.3555	0.3790	0.3324	0.3051
LDA	0.5567	0.509	0.4671	0.1877	0.4148	0.1799
PureSVD	0.5691	0.3239	<u>0.5733</u>	<u>0.5048</u>	<u>0.4995</u>	<u>0.3853</u>
SLIM	**0.6540**	0.5129	**0.6141**	**0.5338**	**0.6141**	**0.5383**
WRMF	0.5901	<u>0.5171</u>	0.2826	0.334	0.2194	0.2219

Table 5. The AUC of the disparity-relevance trade-off curves for different models across the six different fairness dimensions calculated with the trapezoidal rule. Bold text indicates the highest value in each column, and the second highest value is underlined.

C Occupation as group attribute: Kendall rank correlation

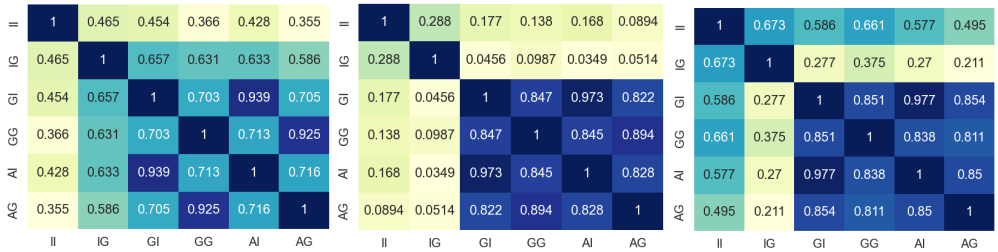


Figure 5. The Kendall rank correlation between the six JMEF metrics and their two components with occupation as user attributes. From the left to the right column, we respectively find the heatmaps for the fairness, disparity, and relevance components.

D Table of code issues

Issue	Provided solution
The folder in which the results from running the main program are saved was missing and was not created during runtime.	Restructured codebase with the required missing folder.
No environment to run the experiments was provided.	Creation of an environment that is available in our repository.
Bug in the code collocating the users into age groups was not correctly allocating them to their corresponding groups.	Reformed code for the corresponding function.
Missing code for the calculation of the AUC used in one of the metric decomposition experiments.	Inclusion of a new function for this purpose.
The code used to plot the figures shown in the paper, as well as the heatmaps, was missing.	Creation of new functions to plot and save the results of the experiments in an image format.
The code was largely uncommented and thus difficult to follow and understand.	We added comprehensive comments, separated large functions into smaller ones, and reformed the structure of the codebase to make it easier to interpret and debug.

E Figures of the results in the original paper

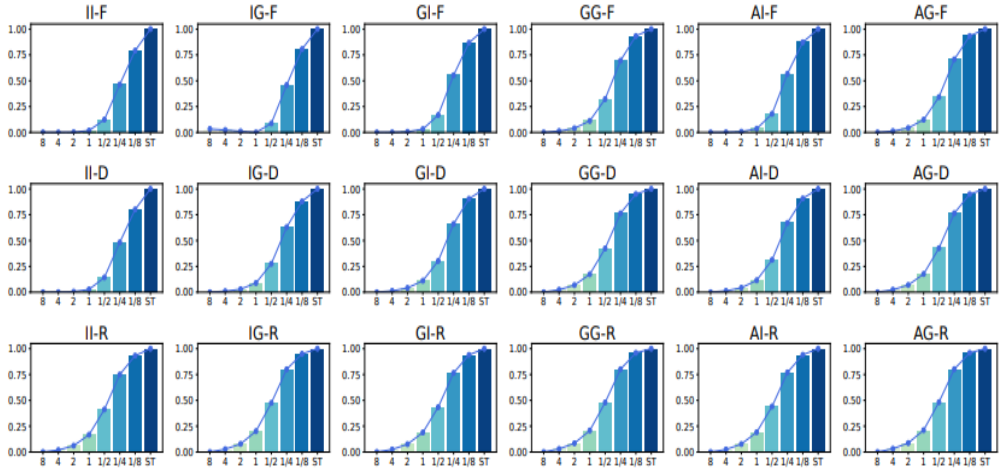


Figure 6. Results for the Behavior of JME-fairness metrics for a stochastic ranking policy, generated by randomizing the BPRMF model using Placket-Luce on the MovieLens1M dataset, as shown in the paper. It corresponds to Figure 1 of this document.

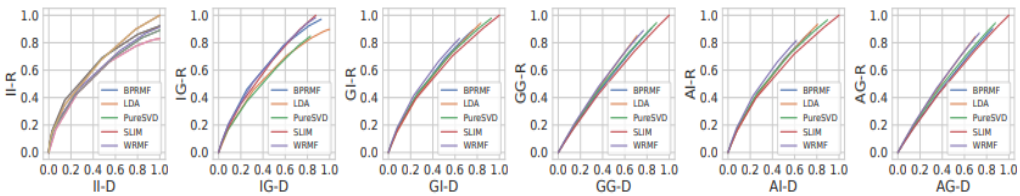


Figure 7. Curves for disparity-relevance trade-off across the six different fairness dimensions, introducing different levels of stochasticity on top of the static rankings from six recommendation models, as shown in the paper. It corresponds to Figure 2 of this document.

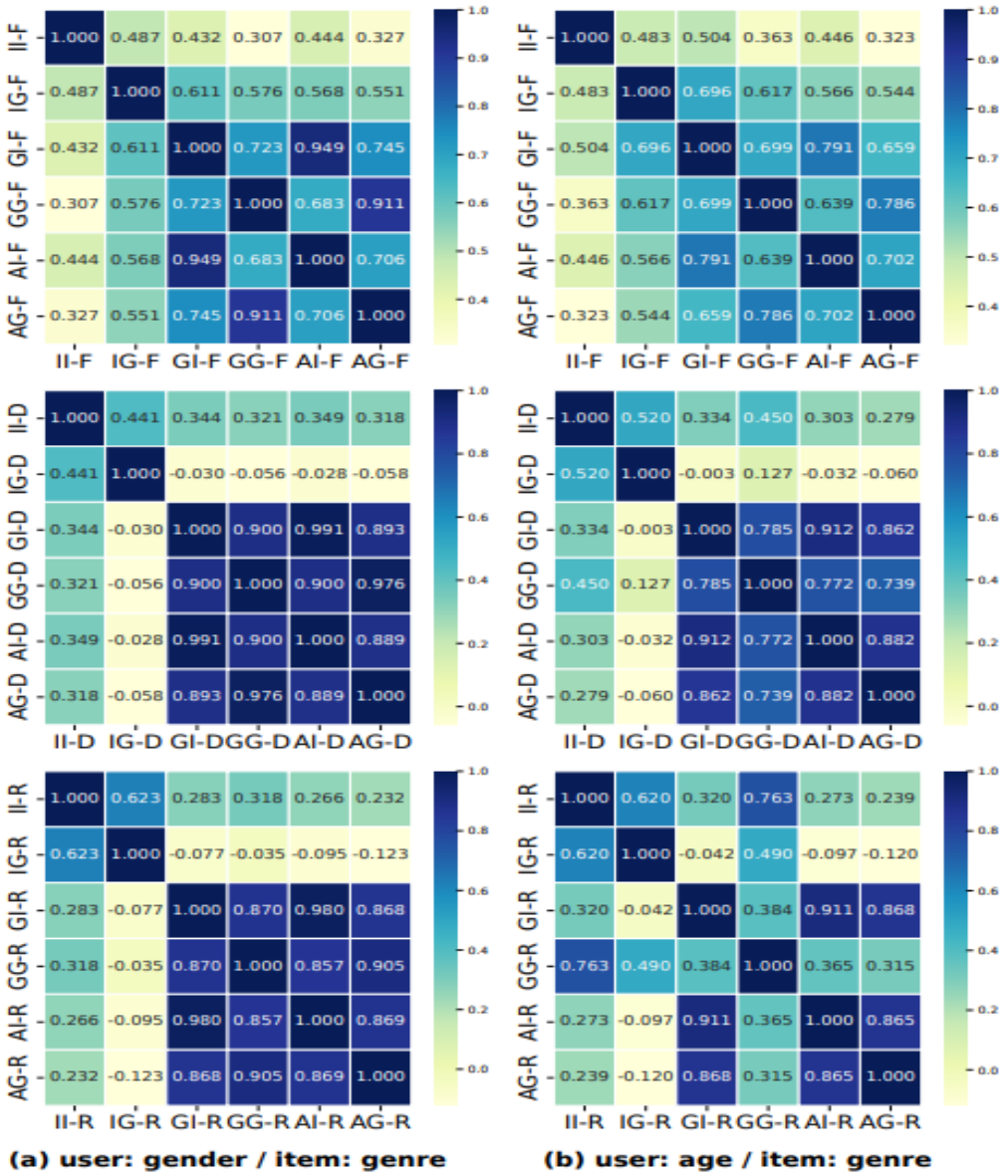
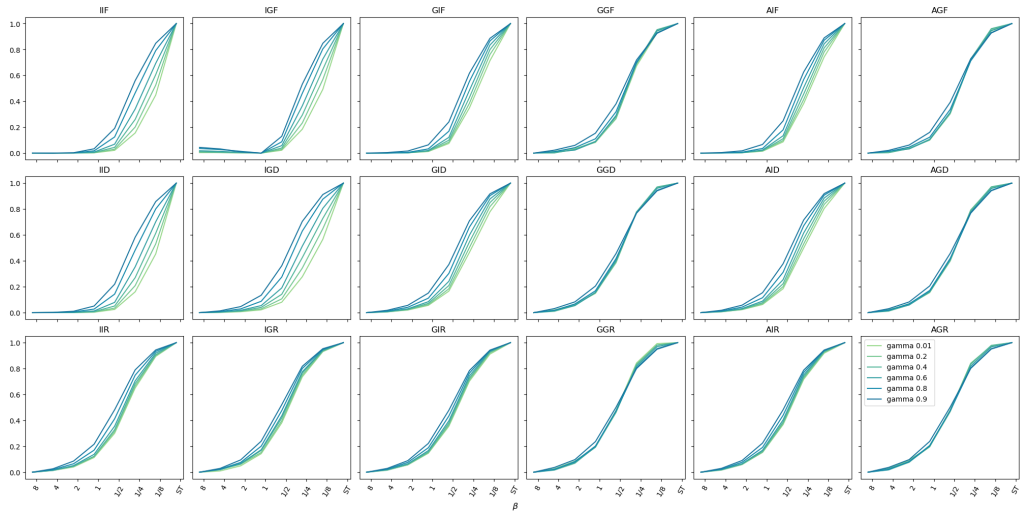


Figure 8. The Kendall rank correlation between different metrics and their disparity and relevance across six different fairness dimensions, as shown in the paper. It corresponds to Figure 3 of this document.

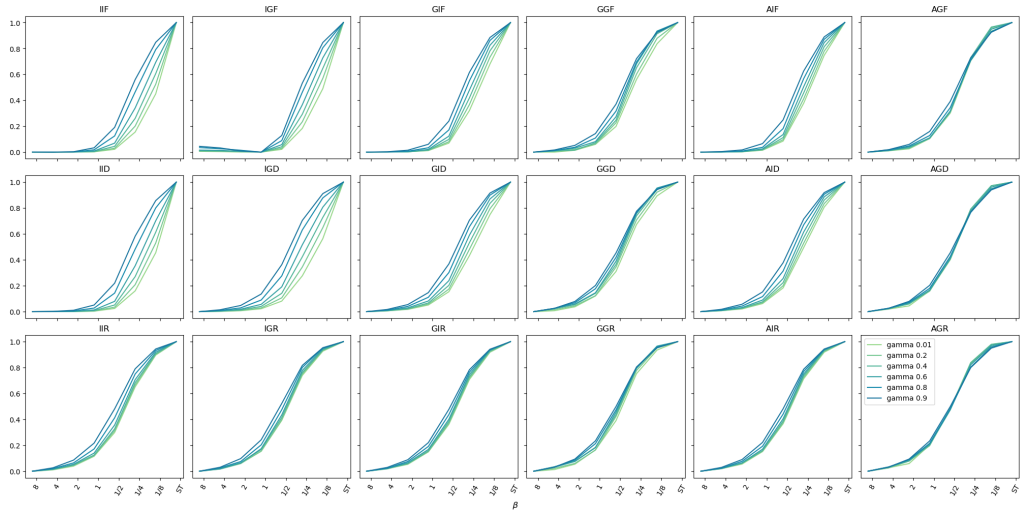
Model	II	IG	GI	GG	AI	AG
BPRMF	<u>0.6331</u>	**0.4774**	0.2904	0.2953	0.2712	0.2814
LDA	0.5664	0.4088	0.2837	<u>0.3164</u>	0.2687	<u>0.3115</u>
PureSVD	0.5830	0.4102	<u>0.2921</u>	0.3030	<u>0.2755</u>	0.2942
SLIM	**0.6408**	0.4654	0.2776	0.2851	0.2605	0.2752
WRMF	0.5996	<u>0.4769</u>	**0.3135**	**0.3186**	**0.2957**	**0.3139**

Table 6. The AUC of the disparity-relevance trade-off curves for different models across the six different fairness dimensions, as shown in the paper. Bold text indicates the highest value in each column, and the second highest value is underlined.

F Hyperparameter sweep

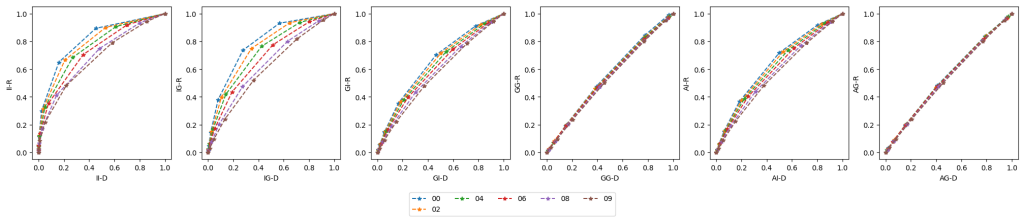


(a) Group: Gender

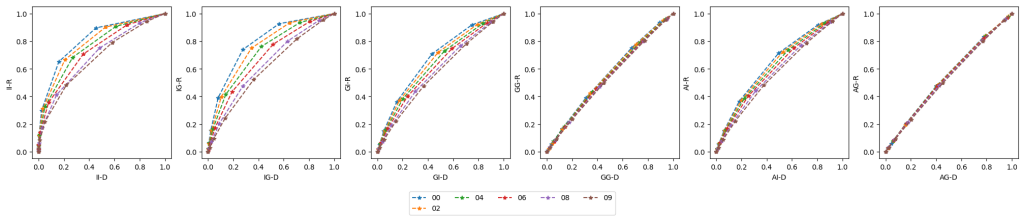


(b) Group: Age

Figure 9. Results for the Behavior of JME-fairness metrics for a stochastic ranking policy, using different values for gamma (patience factor). The same randomized BPRMF model is utilized as in figure 6.



(a) Group: Gender



(b) Group: Age

Figure 10. Curves for disparity-relevance trade-off across the six different fairness dimensions for group 'Age', introducing different levels of stochasticity on top of the static rankings from the BPRMF model, initialized with six different gamma values.

G Computational Costs

Table 7. Computational costs for running JME metrics

Model	Dataset	User Group Attr.	Conduct	Time (s)
BPRMF	ML-1M	Gender	stochastic	725.0684
BPRMF	ML-1M	Gender	static	154.5441
BPRMF	ML-1M	Age	stochastic	957.8651
BPRMF	ML-1M	Age	static	180.7152
LDA	ML-1M	Gender	stochastic	864.2498
LDA	ML-1M	Gender	static	153.9465
LDA	ML-1M	Age	stochastic	1089.0162
LDA	ML-1M	Age	static	186.1006
PureSVD	ML-1M	Gender	stochastic	715.7262
PureSVD	ML-1M	Gender	static	151.7859
PureSVD	ML-1M	Age	stochastic	950.048
PureSVD	ML-1M	Age	static	188.2004
SLIM	ML-1M	Gender	stochastic	719.0509
SLIM	ML-1M	Gender	static	151.7716
SLIM	ML-1M	Age	stochastic	956.7448
SLIM	ML-1M	Age	static	193.7132
WRMF	ML-1M	Gender	stochastic	741.1262
WRMF	ML-1M	Gender	static	161.2652
WRMF	ML-1M	Age	stochastic	993.4329
WRMF	ML-1M	Age	static	198.609
CHI2	ML-1M	Gender	stochastic	683.553
CHI2	ML-1M	Gender	static	161.7956
CHI2	ML-1M	Age	stochastic	935.251
CHI2	ML-1M	Age	static	197.9363
HT	ML-1M	Gender	stochastic	710.8722
HT	ML-1M	Gender	static	166.7169
HT	ML-1M	Age	stochastic	929.8073

Continued on next page

Table 7 – continued from previous page

Model	Dataset	User Group Attr.	Conduct	Time (s)
HT	ML-1M	Age	static	199.2845
KLD	ML-1M	Gender	stochastic	692.9546
KLD	ML-1M	Gender	static	158.0465
KLD	ML-1M	Age	stochastic	932.6938
KLD	ML-1M	Age	static	192.4868
LMWI	ML-1M	Gender	stochastic	688.2506
LMWI	ML-1M	Gender	static	158.2718
LMWI	ML-1M	Age	stochastic	933.3253
LMWI	ML-1M	Age	static	192.1429
LMWU	ML-1M	Gender	stochastic	684.1796
LMWU	ML-1M	Gender	static	159.5841
LMWU	ML-1M	Age	stochastic	921.215
LMWU	ML-1M	Age	static	199.5242
SVD	ML-1M	Gender	stochastic	696.7073
SVD	ML-1M	Gender	static	164.047
SVD	ML-1M	Age	stochastic	942.2578
SVD	ML-1M	Age	static	198.7106
NNI	ML-1M	Gender	stochastic	686.0589
NNI	ML-1M	Gender	static	165.4972
NNI	ML-1M	Age	stochastic	932.6179
NNI	ML-1M	Age	static	199.6226
NNU	ML-1M	Gender	stochastic	688.285
NNU	ML-1M	Gender	static	162.959
NNU	ML-1M	Age	stochastic	938.2951
NNU	ML-1M	Age	static	197.7788
PLSA	ML-1M	Gender	stochastic	694.9885
PLSA	ML-1M	Gender	static	160.9721
PLSA	ML-1M	Age	stochastic	933.0947
PLSA	ML-1M	Age	static	198.9573
Random	ML-1M	Gender	stochastic	697.7134
Random	ML-1M	Gender	static	164.7631
Random	ML-1M	Age	stochastic	949.5797
Random	ML-1M	Age	static	200.4768
RM1	ML-1M	Gender	stochastic	690.9269
RM1	ML-1M	Gender	static	159.6506
RM1	ML-1M	Age	stochastic	943.9163
RM1	ML-1M	Age	static	199.4095
RM2	ML-1M	Gender	stochastic	686.7696
RM2	ML-1M	Gender	static	160.0639
RM2	ML-1M	Age	stochastic	929.7664
RM2	ML-1M	Age	static	194.2399
RSV	ML-1M	Gender	stochastic	686.5787
RSV	ML-1M	Gender	static	154.8014
RSV	ML-1M	Age	stochastic	924.8023
RSV	ML-1M	Age	static	190.9936
RW	ML-1M	Gender	stochastic	726.8647
RW	ML-1M	Gender	static	155.593
RW	ML-1M	Age	stochastic	986.0415
RW	ML-1M	Age	static	189.6771
UIR	ML-1M	Gender	stochastic	678.447
UIR	ML-1M	Gender	static	148.8541

Continued on next page

Table 7 – continued from previous page

Model	Dataset	User Group Attr.	Conduct	Time (s)
UIR	ML-1M	Age	stochastic	935.432
UIR	ML-1M	Age	static	186.4008
Bert4Rec	ML-1M	Gender	static	261.659
Bert4Rec	ML-1M	Gender	stochastic	1779.0463
Bert4Rec	ML-1M	Age	static	330.8377
Bert4Rec	ML-1M	Age	stochastic	2256.4323