Experiments on Generalizability of User-Oriented Fairness in Recommender Systems

Hossein A. Rahmani
University College London
h.rahmani@ucl.ac.uk

Mahdi Dehghan
Shahid Beheshti University
mahdi.dehghan551@gmail.com

Mohammadmehdi Naghiaei
University of Southern California
naghiaei@usc.edu

Mohammad Aliannejadi
University of Amsterdam
m.aliannejadi@uva.nl

ABSTRACT

Recent work in recommender systems mainly focuses on fairness in recommendations as an important aspect of measuring recommendation quality. A fairness-aware recommender system aims to treat different user groups similarly. Relevant work on user-oriented fairness highlights the discriminative behavior of fairness-unaware recommendation algorithms towards a certain user group, defined based on users’ activity level. Typical solutions include proposing a user-centered fairness re-ranking framework applied on top of a base ranking model to mitigate its unfair behavior towards a certain user group i.e., disadvantaged group. In this paper, we re-produce a user-oriented fairness study and provide extensive experiments to analyze the dependency of their proposed method on various fairness and recommendation aspects, including the recommendation domain, nature of the base ranking model, and user grouping method. Moreover, we evaluate the final recommendations provided by the re-ranking framework from both user- (e.g., NDCG, user-fairness) and item-side (e.g., novelty, item-fairness) metrics. We discover interesting trends and trade-offs between the model’s performance in terms of different evaluation metrics. For instance, we see that the definition of the advanced/disadvantaged user groups plays a crucial role in the effectiveness of the fairness algorithm and how it improves the performance of specific base ranking models. Finally, we highlight some important open challenges and future directions in this field. We release the data, evaluation pipeline, and the trained models publicly on https://github.com/rahmanidashti/FairRecSys.

CCS CONCEPTS

• Information systems → Recommender systems.

KEYWORDS

Fairness, Algorithmic Fairness, Re-ranking, Recommender Systems
of a utility-optimized ranking [6, 9, 22]. Post-processing fairness ranking approaches have attracted lots of interest in the community because they provide a tool that can transform various state-of-the-art black-box ranking methods to a fair ranking. Moreover, such techniques are practical choices in many use cases since they do not require re-training in case the definition of fairness or protected groups changes. This is a critical feature, especially when group and fairness definitions are dynamic and re-training a model comes at a high cost.

As mentioned earlier, various bias and fairness aspects (e.g., item and user) can influence how one interprets a fair ranking. Moreover, pre- and post-processing fairness approaches introduce a new dimension, i.e., the base ranking model. We observe that various research studies provide post-processing solutions, applied to a limited number of ranking models, on a limited number of domains (i.e., datasets). Moreover, in the majority of the current studies, only a limited number of fairness aspects are considered. For instance, Ferraro et al. [13] propose a re-ranking algorithm for user-oriented fairness, where they only evaluate the performance of their method on one type of bias and domain i.e., gender and music, respectively. Moreover, they conduct their experiments on a limited range of base ranking models including collaborative filtering approaches and the most famous baselines such as MostPop. Liu et al. [26] also propose a re-ranking fairness algorithm on a dataset generated from Kiva.org. Although they consider fairness in two-sided markets, they only apply their proposed method to one domain. Moreover, they do not test the performance of their method when applied on top of a wide range of base ranking models (e.g., deep models are ignored in their experiments). Experimenting with post-processing methods against multiple base ranking models is crucial, especially when the fairness algorithm does not assume having access to the relevance judgments, since the distribution of top-ranked items and the model’s learned estimate of relevance can influence the performances to a great extent.

To this aim, in this paper, we re-implement and reproduce the results of a state-of-the-art user fairness post-processing framework on eight recommendation datasets, consisting of six domains. In particular, we provide an in-depth experimental analysis of Li et al. [22], based on several reproducibility aspects, namely, (i) domain, (ii) base recommendation algorithm, (iii) user group assumptions, and (iv) interplays in other effectiveness metrics including item fairness. In summary, our reproducibility study concerns the following aspects:

- We conduct extensive experiments on various domains with different feedback types (i.e., explicit and implicit) and characteristics, namely, MovieLens, Epinions, BookCrossing, AmazonToy, AmazonOffice, Gowalla, Foursquare, and Last.fm. We investigate whether the user-oriented fairness re-ranking (UFR) method can be applied to diverse domains and whether underlying data characteristics (e.g., sparsity and average user interactions) can explain the degree of improvement in fairness.
- We examine the effectiveness of UFR on six new models, including MostPop [11], BPR [38], PF [14], WMF [19], NeuMF [18], and VAECF [24]. We aim to analyze how different baseline algorithms influence the fairness performance in the proposed model. In particular, we investigate whether some baseline algorithms (e.g., shallow vs. deep learning-based models) are less prone to unfair treatment between user groups.
- We further analyze the impact of improving user-oriented fairness in two-sided markets. In particular, we investigate whether improving fairness from the user’s perspective affects the producer’s visibility, system catalog coverage, and existing popularity bias, i.e., over-emphasizing popular items (i.e., short-head items) with limited visibility to less popular items (i.e., long-tail items).
- In addition, we investigate the underlying relation between user-oriented fairness and user-oriented evaluation of popularity bias, known as popularity bias calibration fairness [2, 32], i.e., the degree to which popularity bias impacts users with different levels of interest in popular items in their profiles.
- Finally, we investigate the effectiveness of UFR under different user groups based on their activity level and interaction with popular items. We analyze the user grouping effect on the improvement of fairness. We further study whether dividing users into groups based on a cutoff (e.g., top 5%) can create two categories of (dis)advantaged users in a meaningful way across various domains.

Our reproducibility experiments on UFR lead to critical observations and implications on this problem. For instance, our experiments on different domains indicate that the effectiveness of UFR highly depends on the domain. Also, we found a high dependency of fairness performance on the way advantaged/disadvantaged users are defined. This study provides important guidelines on how to evaluate fairness methods more effectively, particularly when proposing a fairness algorithm that involves only one of multiple stakeholders (i.e., only the user or the vendor) since it can affect the exposure of other stakeholders. Therefore, it is crucial for a wide range of fairness algorithms to be evaluated thoroughly, following our proposed guidelines.

2 REPRODUCIBILITY METHODOLOGY

In this section, we first describe the approach that we reproduce. Then, we provide the details of the datasets and the recommendation algorithms used in the fairness re-ranking algorithm. Finally, we give a detailed description on how we define the advantaged and disadvantaged users groups.

2.1 UFR Implementation

Li et al. [22] consider user unfairness across different user groups based on the level of activity in the platform (the more or the less active). They adopt a constrained form of re-ranking integer programming, as opposed to Naghiaei et al. [33] that used an unconstrained version. The user-oriented fairness re-ranking algorithm (UFR) objective function is to select (or re-rank) the recommendation list of each user provided by the baseline algorithms to maximize the total score (accuracy) while constrained by keeping the difference in average recommendation performance between the groups of users in a given range. They use F1@10 and NDCG@10 to assess recommendation performance quality and the difference of NDCG between the two user groups as the constraint, also called user-oriented group fairness (UGF). Their evaluation on Amazon datasets shows that their re-ranking approach leads to improved fairness.
between the two user groups significantly, while at the same time improving the overall recommendation utility. The original code is available in a public GitHub repository, however, it contains only the fairness-aware post-processing re-ranking method and lacks the training and testing pipeline of the baseline models used in the study. Moreover, they make use of Gurobi, a commercial optimization solver to be able to tune the codes.

Here, we implement the post-processing pipeline and the integer programming optimization code. In contrast to Li et al. [22], we use Cornac, a Python-based open-source recommendation toolkit, to provide various state-of-the-art recommendation models, straightforward evaluation pipeline, and to foster the reproducibility of the study. Furthermore, we implement a fairness-aware re-ranking method using MIP, a Python tool for the modeling and optimization, on top of Groubi, enabling us to run the code with a license obtained for academic use only. With this setup, we examine 96 experimental cases on eight well-known recommendation datasets, six state-of-the-art recommendation algorithms, and two user grouping unfairness criteria i.e., 8 (datasets)×6 (CF baselines)×2 (fairness constraints type).

2.2 Datasets
We use eight public datasets from different domains such as Movie, Point-of-Interest (POI), Music, eCommerce, Book, and Opinion with different types of feedback (i.e., explicit and implicit) and characteristics, including MovieLens [16], Epinions [29], BookCrossing [48], AmazonToy [17], AmazonOffice [30], Gowalla [27], Last.fm [8], and Foursquare [27]. We apply k-core pre-processing (i.e., each training user/item has at least k ratings) on the datasets to make sure each user/item has sufficient feedback. Table 1 shows the statistics of all the datasets.

2.3 Base Ranking Models
We compare the performance of UFR on several recommendation approaches, from traditional to deep recommendation models, as suggested by Dacrema et al. [11]. Therefore, we include two traditional methods (PF and WMF), as well as two deep recommendation models (NeuMF and VAECF). We also include two baselines approaches, MostPop and BPR, for further investigation and comparison of the results. Below, we explain the methods in more detail:

- **MostPop** [11]: This method is a non-personalized method that recommends the relevant items to each user. Popularity is measured by the number of interactions of items.
- **BPR** [38]: This is a state-of-the-art method for personalized ranking, particularly on implicit feedback datasets. BPR uses a pairwise ranking loss whose goal is to optimize personalized ranking.
- **PF** [14]: This method is a variant of probabilistic matrix factorization where the weights of each user and item latent features are positive and modeled using the Poisson distribution.
- **WMF** [19]: This method is a weighted matrix factorization with l2-norm regularization that considers the latent features of two items independently.
- **NeuMF** [18]: This algorithm learns user and item features using the multi-layer perceptron (MLP) on one side and matrix factorization (MF) from another side, then applies non-linear activation functions to train the mapping between users and items features that are concatenated from MLP and MF layers.
- **VAECF** [24]: This method is based on variational autoencoders which introduce a generative model with multinomial likelihood and use Bayesian inference for parameter estimation.

Note that we use the implementation of all these methods publicly available in the Cornac library. We compare the fair re-ranking algorithm applied on top of the baselines to show how the re-ranking method can achieve the desirable performance on fairness metrics and overall recommendation performance (accuracy and beyond-accuracy).

**Hyperparameter Setting.** We adopt the baseline algorithms with the default parameter settings, suggested in their original paper. We set the embedding size for users and items to 50 for all baseline algorithms. For NeuMF, we set the size of the MLP with 32, 16, 8 and we apply the hyperbolic tangent (TanH) non-linear activation function between layers. We set the learning rate to 0.001. We apply Adam [20] as the optimization algorithm to update the model parameters.

2.4 Fairness Assumption and Evaluation
In group fairness, the protected groups are treated similarly to the advantaged group [36]. The group of users can be divided under different requirements for different tasks [2, 22]. In this study, we examine two different user grouping methods according to the level of activity of users, as well as consumption of popular items in the user profiles. We label the more active users as an advantaged group, while the remaining users as a disadvantaged group as follows:

**Definition 1.** The level of activity (G1): The study of Li et al. [22] groups users into advantaged and disadvantaged groups according to their level of activity i.e., they select the top 5% of users in the training dataset ranked by the number of interactions and label them as the advantaged group and the remaining users are labeled as the disadvantaged group. We apply this grouping strategy to our study to highlight and evaluate the same strategy as the study of Li et al. [22].

**Definition 2.** The consumption of popular items (G2): Following [2], we also divide users into advantaged and disadvantaged groups based on the number of popular items in their profiles. We define an item to be popular if its popularity value (i.e., the number of interactions) falls within the top 20% of item popularity. We then select the top 20% of users in the training dataset ranked by the number of popular items and label them as the advantaged group and the remaining users are labeled as the disadvantaged group.

The reason for exploring these two methods is that we believe the difference between user interactions and popular consumption will reflect their different activity level in a reasonable manner. In the paper, we refer to the first grouping method as the G1 and the second grouping method as the G2.
We evaluate the popularity bias of different user groups (i.e., advantaged vs. disadvantaged) using the delta Group Average Popularity (ΔGAP) metric proposed by Abdollahpouri et al. [2]. For each recommendation algorithm and user group, ΔGAP measures the difference between the popularity of the recommended items from the expected popularity of the items in the user profiles as $\Delta G A P = GAP(g_1) - GAP(g_2)$, where $g$ is a given user group, $GAP(g_1)$ and $GAP(g_2)$ represent the GAP value for user profiles and recommendations, respectively, and it is defined as $GAP(g) = \sum_{t \in \text{top}(g)} \phi(t) \frac{p_t}{|P_u|}$, where $\phi(t)$ is the popularity of item $t$, which is the number of times it is rated and $p_t$ is the list of items in the profile of user $u$. The value of $\Delta GAP = 0$ indicates the recommendation is fair, that is the average popularity of recommended items matches the average popularity of the user profile.

### 2.5 Relevance Estimation Assumption

Various post-processing fairness algorithms assume having access to the relevance judgments [22, 43] in the post-processing step. Therefore, the fairness algorithm is aware of the relevance of the items in the post-processing step and can leverage this information in the re-ranking process. However, in this work, we conduct our experiments with the assumption that the relevance labels are not available to the post-processing algorithm and are estimated based on the training data. Not only is this a more realistic assumption, but it also leads to a more concrete reproducibility study. In particular, studying the effect of the base ranking model is justified mainly in cases where the fairness algorithm takes as input the base ranking and estimates the relevance of the items based on the model’s predictions (rather than their true labels).

### 2.6 Code and Data

We release the code of the complete pipeline, including training and inference of the base ranking models, as well as the fairness-aware re-ranking stage, together with the pre-processed data used to produce the results in this paper. Moreover, we provide ready-to-run Jupyter Notebooks, tested on Google Colab that can be executed to obtain the results and plots reported in the paper. All the related resources can be accessed on our website: https://github.com/rahmanidashti/FairRecSys

### 3 REPRODUCIBILITY ASPECTS

In this section, we explain the experiments we conduct on various reproducibility aspects and discuss the obtained results and our observations. Note that, in the interest of space, we only report the detailed results on two datasets, but provide the results on all the datasets on our repository.

#### 3.1 Domain

In this experiment, we aim to study the generalizability of UFR on different domains. We study the robustness of UFR, applied on eight different datasets, consisting of six domains. Moreover, we explore the effectiveness of different data characteristics including average interactions per user/item, feedback type, sparsity, etc. on the capability of UFR to mitigate bias towards a certain user group.

**Effect on UGF.** In Figure 1, we depict UFR’s performance in terms of UGF improvement (Δ%UGF) on each domain for both user grouping methods. Figure 1a shows that certain domains exhibit different patterns. It is clear from this figure that mitigating discrimination against the disadvantaged user group can be a challenging task on specific domains. For example, we see that eCommerce domains (Amazon datasets) indicate a lower UGF improvement. We also observe a relatively better performance for POI recommendation datasets (Gowalla and Foursquare). Surprisingly, in Figure 1b we observe a totally different behavior on some datasets such as Epinions, Gowalla, and BookCrossing depending on the user grouping (G1 vs. G2). Moreover, eCommerce domains have considerably similar behavior on both G1 and G2. In general, according to Figures 1a and 1b, we observe a wider range of variance in UGF improvement on implicit-feedback datasets compared to explicit-feedback datasets. We can argue for this particular observation from two specific points of view, including (1) user grouping methods (interactions vs. popular consumption) (2) the characteristics of datasets. Next, we analyze the second point of view while we investigate the first one in Section 3.3.

**Data characteristics impact on UGF improvement.** In Figure 2, for each grouping method we depict a scatter plot in which the horizontal axis (x-axis) is average UGF improvement over all base ranking models on a specific dataset and the vertical axis (y-axis) can be either $|P_f| / |P|$ or $|P_f| / |P_u|$ (see Table 1). In these plots, blue lines

Table 1: Statistics of the datasets: $|U|$ is the number of users, $|I|$ is the number of items, $|P|$ is the number of interactions, $|ST|$ is the number of short-head items, $|L|$ is the number of long-tail items, $|AU|$ and $|DU|$ are the number of advantaged and disadvantaged users, respectively, in G1 and G2 (in brackets are the # of relevant items).

| Dataset       | $|U|$   | $|I|$   | $|P|$ | $|P_f|$   | $%$Sparsity | Domain | $|ST|$   | $|L|$   | $|AU|$ | $|DU|$ | $G1$ | $|AU|$ | $|DU|$ | $G2$ |
|---------------|--------|--------|------|--------|-----------|--------|--------|--------|--------|--------|-------|--------|--------|-------|
| Epinions      | 2,677  | 2,060  | 103,567 | 38.6 | 50.2 | 98.12% | Opinion | 412    | 1,648 | 134 (12220) | 2542 (60276) | 535 (28653) | 2141 (43843) |
| MovieLens     | 943    | 1,349  | 99,287  | 105.2 | 73.6 | 92.19% | Movie   | 269    | 1,080 | 47 (12698)    | 895 (56802)    | 188 (34829)   | 754 (34671)   |
| BookCrossing  | 1,136  | 1,019  | 20,522  | 18.0 | 20.1 | 98.22% | Book    | 203    | 816 | 57 (1831)        | 1078 (12533)   | 227 (4904)     | 908 (9460)     |
| AmazonToy     | 2,170  | 1,733  | 32,852  | 15.1 | 18.9 | 99.12% | eComm.  | 346    | 1,387 | 108 (3867)      | 2061 (19128)   | 434 (9577)     | 1735 (13418)   |
| AmazonOffice  | 2,448  | 1,596  | 36,841  | 15.0 | 23.0 | 99.05% | eComm.  | 319    | 1,276 | 122 (3945)      | 2325 (21842)   | 489 (10392)    | 1958 (15395)   |
| Gowalla       | 1,130  | 1,189  | 66,245  | 58.6 | 55.7 | 95.06% | POI     | 237    | 952 | 56 (6452)         | 1073 (39919)   | 226 (19267)    | 903 (27104)    |
| Foursquare    | 1,568  | 1,461  | 42,678  | 27.2 | 29.2 | 98.13% | POI     | 292    | 1169 | 78 (2899)        | 1489 (26975)   | 313 (10027)    | 1254 (19847)   |
| Last.fm       | 1,797  | 1,130  | 66,245  | 58.6 | 55.7 | 95.06% | Music   | 301    | 1,206 | 90 (2447)        | 1766 (41215)   | 359 (11789)    | 1437 (31873)   |

GAP

We release the code of the complete pipeline, including training and inference of the base ranking models, as well as the fairness-aware re-ranking stage, together with the pre-processed data used to produce the results in this paper. Moreover, we provide ready-to-run Jupyter Notebooks, tested on Google Colab that can be executed to obtain the results and plots reported in the paper. All the related resources can be accessed on our website: https://github.com/rahmanidashti/FairRecSys
are concerned with relation between $\frac{|P|}{|I|}$ and average UGF improvement. The correlation coefficients and p-values of these two variables for interactions and popular consumption grouping methods are (0.72, 0.0431) and (0.84, 0.0076), respectively suggesting a statistically significant correlation. In other words, the higher number of interactions per user can lead to better performance of fairness model. Furthermore, in these plots, orange lines represent relation between $\frac{|P|}{|I|}$ and average UGF improvement. Interestingly, the same behavior exists when investigating the correlation between these two variables. That is, we observe the correlation coefficient and p-values of (0.8076, 0.0153) and (0.8756, 0.0044) for G1 and G2, respectively. Therefore, as the contributions of items in interactions increases, the average improvement of UGF increases. It is worth mentioning that Gowalla and MovieLens have the best average improvement of UGF among implicit and explicit datasets, respectively, since they have the highest values of $\frac{|P|}{|I|}$ and $\frac{|P|}{|T|}$ (see Table 1). Moreover, as Figure 1 shows, eCommerce domains (Amazon datasets) have the lowest improvement in terms of UGF for both G1 and G2. The reason can be rooted in some data characteristics. For instance, these domains have the lowest values of $\frac{|P|}{|I|}$ and $\frac{|P|}{|T|}$. Also, it is worth mentioning that we observe the highest sparsity among all the datasets, suggesting that it can also play a role in our observation.

3.2 Base Recommendation Algorithms

In this experiment, we intend to examine how different types of base ranking models (e.g., deep vs. shallow) treat different user groups before and after applying UFR.

Base ranking fairness analysis. Table 2 and Table 3 show the detailed results of different base ranking models in which Org. and Fair rows indicate results before and after applying UFR, respectively. Here, we intend to investigate which base ranking models are less prone to unfairness before applying UFR. According to these tables, almost all base ranking models suffer from the user-fairness issue. MostPop appears to be particularly prone to fairness issues since, usually, they interact with more long-tail items in their profile [37]. On the other hand, PF considers the tastes of both user groups better than other models. As one can see in Org. rows of Tables 2 and 3, PF puts more emphasis on long-tail items, compared to other base ranking models, leading to less effectiveness on the disadvantaged user group.

Although the treatment of different base ranking models can be attributed to their characteristics, our observations suggest that the
user grouping assumption also affects the sensitivity of a model to fairness. We refer to Table 3 where in some cases, base ranking models provide the disadvantaged group with the same or better recommendations. In this case, the base ranking models do not discriminate against the disadvantaged group on the Last.fm dataset, and seem not to favor the advantaged group. We observe similar behavior of base ranking models on the Foursquare dataset. We bring a comprehensive analysis of the effect of different user grouping methods on the behavior of base ranking models before and after applying UFR in Section 3.3.

Impact of UFR on base ranking models. Figure 4 demonstrates the performance of different models in terms of UGF improvement after applying UFR for both grouping methods. We see that post-processing the output of different models using UFR exhibits different patterns. Interestingly, Figure 4a shows that UFR manages to mitigate unfairness on the two baseline approaches (i.e., MostPop and BPR) much more effectively than VAECF which is a deep recommendation model. Also, they provide comparable results in terms of UGF improvement in comparison to another deep recommendation model (i.e., NeuMF). Moreover, MostPop and BPR have a moderate range of variance indicating the fact that they can be considered as robust models when it comes to dealing with different domains. Also, UFR applied to WMF demonstrates the most robust performance in mitigating user-oriented unfairness, exhibiting the least variance, as well as the best average improvement in terms of UGF (i.e., 68.54%). On the other hand, Figure 4b demonstrates the performance of different base ranking models in terms of UGF improvement for G2.

Comparing Figures 4a and 4b, we observe two considerably different behaviors of different base ranking models when it comes to the variance and UGF improvement. Specifically, we see that the two baseline approaches (i.e., MostPop and BPR) have the worst performance in terms of UGF improvement. Moreover, VAECF model has two opposite treatments on G1 and G2. This model has the minimum and maximum average improvement of UGF (i.e., 41.9% and 55.52%) for G1 and G2, respectively. Also, there is a wider range of variance in VAECF model for G2. However, WMF model seems to be the most reliable base ranking model since it has a reasonable variance range and also a considerable average improvement of UGF for both G1 and G2. To sum up, we see that the way advantaged/disadvantaged users are defined (i.e., user grouping method) can lead to drastically different results, reiterating the importance of a comprehensive evaluation that considers multiple assumptions.

3.3 User Groups Assumptions

In this section, we examine the effect of the user grouping methods on the performance of the UFR. In Section 2.4, we introduce two common settings that are used by previous studies [2, 22] to group users according to their level of activity and the consumption of popular items. Therefore, herein, the main question is to what extent the user grouping assumption could impact the performance or fairness between different user groups.

Effect of user grouping method on performance of UFR. In this work, we aim to study the performance of UFR, under other user group assumptions that are not studied in the original paper, enabling us to compare the improvements with other works in the literature and examine the effectiveness of the model when different assumptions are taken into account. To this end, we consider the following user grouping techniques, namely, G1 and G2 (see Section 2.4). Figure 1 depicts UFR’s performance across different datasets evaluated by Δ%UGF for G1 and G2. UFR enhances recommendation fairness on both G1 and G2 across different datasets. However, we see apparent different performances on G1 and G2. For instance, UFR achieves higher performance in Epinions on G1 compared to G2. Thus, the main question is the reason for different treatment of UFR when grouping users differently. Comparing Figures 1a and 1b, we observe a worse performance on G2 compared to G1. This indicates that the higher the number of users of the advantaged group, the more challenging the task would be (see Table 1). Moreover, we see that the box plots show a wider range of variance in terms of the performance of different models, suggesting that the tested models are less robust when it comes to dealing with more users. Interestingly, we see a very large drop in the performance in Epinions. Looking at Table 2, we observe that UFR’s UGF improvement shows a high correlation with the performance difference between the two groups in the original ranking, that is, the more unfair an original ranking is, the more UGF improvement is achieved after applying UFR (Pearson’s $r = 0.36185, p < 0.05$). This shows a clear dependence of the achieved improvement on how users are grouped.

Moreover, we see a reverse behavior in Last.fm, where the performance with G2 is better than the performance with G1. This finding is in line with the study of Kowald et al. [21] and the reason could be related to the fact that the users on Last.fm have relatively more average interactions, compared to other datasets. Therefore,
Table 2: The recommendation performance of all, advantaged, and disadvantaged users of UFR and corresponding baselines on Epinions datasets for G1 and G2. All re-ranking results here are obtained under the fairness constraint on NDCG. The evaluation metrics here are calculated based on the top-10 predictions in the test set.

| Model     | Type | User Relevance (NDCG) | Item Exposure |
|-----------|------|------------------------|---------------|
|           |      | All | Adv. | Disadv. | UGF ↓ | Δ% | Nov. | Cov. | Short. | Long. | ΔGAP_A↓ | ΔGAP_D↓ |
| MostPop   | Org. | 0.0302 | 0.0566 | 0.0288 | 0.0278 | – | 3.7133 | 0.49 | 26770.0 | 0.0 | 255.307 | 204.721 |
|           | Fair | 0.0301 | 0.0374 | 0.0298 | 0.0077 | 72.3 | 3.7123 | 2.43 | 26770.0 | 0.0 | 287.94 | 205.039 |
| BPR       | Org. | 0.0297 | 0.055 | 0.0283 | 0.0267 | – | 3.703 | 0.63 | 26770.0 | 0.0 | 257.095 | 206.271 |
|           | Fair | 0.0316 | 0.0384 | 0.0312 | 0.0071 | 73.41 | 3.6658 | 1.55 | 26770.0 | 0.0 | 291.545 | 217.278 |
| PF        | Org. | 0.0321 | 0.0751 | 0.0298 | 0.0453 | – | 4.9602 | 50.73 | 22169.0 | 4601.0 | 83.502 | 61.209 |
|           | Fair | 0.0346 | 0.0504 | 0.0356 | 0.0148 | 67.33 | 5.016 | 51.12 | 21872.0 | 4898.0 | 90.315 | 53.907 |
| WMF       | Org. | 0.0235 | 0.0577 | 0.0217 | 0.036 | – | 4.9483 | 29.13 | 25225.0 | 1545.0 | 72.469 | 41.703 |
|           | Fair | 0.0255 | 0.0417 | 0.0246 | 0.0171 | 52.5 | 4.9588 | 27.09 | 25335.0 | 1435.0 | 74.201 | 39.559 |
| NeuMF     | Org. | 0.0451 | 0.0865 | 0.0429 | 0.0436 | – | 3.8923 | 12.14 | 26533.0 | 237.0 | 195.495 | 196.681 |
|           | Fair | 0.0488 | 0.0548 | 0.0485 | 0.0062 | 85.78 | 3.9259 | 12.62 | 26531.0 | 239.0 | 191.713 | 188.855 |
| VAECF     | Org. | 0.0444 | 0.0801 | 0.0425 | 0.0376 | – | 4.061 | 31.07 | 24602.0 | 2168.0 | 123.222 | 123.408 |
|           | Fair | 0.0461 | 0.0848 | 0.0459 | 0.0029 | 92.29 | 4.4212 | 31.12 | 24550.0 | 2220.0 | 124.084 | 120.874 |

Considering only top-5% of users with the largest profile size as an advantaged group can not represent the unfairness treatment of fairness-unaware recommendation algorithms. As Table 3 shows, the recommendations provided by most base ranking models fairly represent the tastes of both groups of users. On the other hand, as can be seen in Table 3, G2 can reveal the unfairness issue on the user side on Last.fm. This observation can determine that G2 is not influenced by user profile size as it differentiates users according to the number of popular items in their profiles, irrespective of their profile size. Consequently, based on our experiments, G2 appears to be a relatively more generalizable grouping method than G1.

To sum up, the instability of results on different datasets can be rooted in the fact that classifying users based on a cutoff (e.g., top 5% users based on the number of interactions) is not an accurate way of classifying advantaged vs. disadvantaged user groups. Thus, we suggest to group users using approaches that classify users based on multiple features such as gender, age, race, and interactions count simultaneously. In general, finding a more stable and reasonable way of classifying users such as K-means (or other clustering techniques) is a potential topic for future work.

**Impact of user grouping on Δ%UGF across base ranking models.** When comparing Figures 4a and 4b, we see that various base ranking models show inconsistent behavior in terms of UGF improvement for different grouping methods. Interestingly, we see a large drop in the performance of the MostPop and BPR models with the popular consumption grouping method, G2, in comparison to G1. Previous studies (e.g., Mansoury et al. [28]) show that BPR favors popular items and gives them a high rank in the recommendation list. Clearly, the same holds for MostPop algorithm. We also see a drop, to a lower extent, in the performance of WMF, and NeuMF in Δ%UGF for G2. On the other hand, PF and VAECF models show a more robust behavior with changes in the grouping assumption while the latter (VAECF), achieves slightly better performance (see Figure 4). Our experiments show that this variability is explainable by the extent different algorithms are prone to popularity bias, i.e., base ranking models that mostly include popular items in the recommended list are less stable in the improvement of user-oriented fairness (Δ%UGF) with changes in the grouping method. For example, the average number of long-tail items in the original top-10 recommendation list in BPR, MostPop, PF, WMF, NeuMF, and VAECF across all datasets are 0, 0, 5083, 6211.87, 157,
and 3102.75, respectively. This shows that with the exception of WMF, all algorithms with a small number of long-tail items in the original recommendation list suffer from this instability in the improvement of fairness.

### 3.4 Fairness vs. Effectiveness Metrics

The main objective of this aspect is to mine the underlying trends and trade-offs between various evaluation metrics in recommender systems [37]. In particular, we are interested in analyzing the potential correlations and trade-offs between UGF improvement, i.e., mitigating consumer-oriented fairness, and improvement of other effectiveness metrics such as overall accuracy (NDCG), providers’ (items’) exposure, the novelty of recommendation, and popularity bias among user groups. Figure 3 shows the regression plot of overall datasets and models with G1 and G2. Each point represents one run of UFR with a baseline model in one domain. The horizontal axis (x-axis) shows the UGF improvement (i.e., Δ%UGF) of that run while the vertical axis (y-axis) denotes the improvement in terms of another effectiveness metric, namely, NDCG, Novelty, and ΔGAP. In what follows we discuss these correlations separately.

**Correlation between Δ%UGF and Δ%NDCG.** Surprisingly, as can be seen in Figures 3a and 3d, there exists a positive correlation between improvement in terms of UGF (i.e., Δ%UGF) and improvement in the overall accuracy of the recommender system measured by NDCG. The Pearson’s r correlation coefficient and p-values are (0.55036, 5.06e-05), (0.60236, 5.9e-06) respectively for G1 and G2, suggesting a moderate uphill (positive) relationship. In other words, improving user-oriented fairness by applying UFR does not necessarily sacrifice the overall accuracy of the system. The reason lies behind the fact that the total NDCG of the recommender system is closely dependent on NDCG of disadvantaged users as they are the majority and UFR tries to achieve fairness by giving more attention to disadvantaged users leading to an increase in total accuracy. Across all our experiments, we have seen an average 41.73% improvement over UGF and 4.446% over NDCG.

**Correlation between Δ%UGF and Δ%Nov.** Figures 3b and 3e show the correlation between Δ%UGF and the novelty of the recommender system (Nov). The Pearson’s r correlation coefficients and p-values are (0.512, 0.0001994) and (0.30056, 0.037921) for G1 and G2, respectively. As can be seen, there is a weak uphill correlation (+0.3 correlation) between improvement in fairness and overall novelty of recommender systems and among all experiments, we have seen an average improvement of novelty by 0.304. That is due to the fact that UFR mainly improves user-oriented fairness.
of recommendation by including less popular items on top of the ranking list. Therefore, as the results in Tables 2 and 3 suggest, we naturally observe more recommended long-tail items, causing an improvement in novelty in the fair ranking.

**Correlations between Δ%UGF and Δ%ΔGAP.** The correlation between these two metrics is illustrated in Figures 3c and 3f. The Pearson’s $r$ coefficients and $p$-values for G1 and G2 are (0.10307, 0.4857423) and (0.2131, 0.145885) respectively. This suggests that these two evaluation metrics are not meaningfully correlated. The variation in the Δ%UGF cannot explain the observed variability in Δ%ΔGAP, and UFR does not affect the existing popularity bias in the dataset. Moreover, our experiment indicates the average values of 97.06 and 91.01 for ΔGAP for G1 and G2, respectively, indicating that the average popularity of items in the recommended list is significantly higher than expected according to user profiles after running UFR. Therefore, it is vital to investigate how this model can be extended to incorporate mitigation of popularity bias as well as improvement in UGF.

**4 CONCLUSIONS, IMPLICATIONS, AND FUTURE WORK**

In this paper, we conducted an extensive reproducibility study of the work by Li et al. [22]. The original work focuses on the fairness in recommendation algorithms from the user perspective, evaluating three eCommerce datasets. In this work, we re-implemented the original work and studied its performance under various conditions, namely, fairness assumption, base ranking model, datasets, and evaluation aspects. We provided a deep analysis of how user-oriented fairness can be affected by different user grouping assumptions and data distributions. Moreover, in this work, we aimed at studying the effect of user-oriented fairness on item exposure. Li et al. [22] treat recommendation in the eCommerce domain like a one-sided market; however, item exposure is obviously as important as user fairness, as it translates to vendors’ revenue. As such, we studied the implications of conducting user-oriented fairness on item exposure and showed how mitigating item fairness can affect item exposure. Our experiments on various reproducibility aspects led to interesting new observations and possible future directions in this area, summarized below.

**Domain dependency.** We evaluated the performance of user-oriented fairness on six domains. Our main findings indicate that the effectiveness of UFR highly depends on the domain. For instance, we found a significantly higher improvement in user fairness on POI domain than in eCommerce. Furthermore, our experiment shows that this variability in performance can be explained using some data characteristics. In particular, we found that there exists a strong positive correlation between improvement in fairness with both average interaction per user ($\frac{|P|}{|I|}$) and average interaction per item ($\frac{|P|}{|U|}$). That is, domains with high value of $\frac{|P|}{|I|}$ or $\frac{|P|}{|U|}$ can be improved in terms of fairness to a greater extent. We conclude that it is crucial to evaluate the fairness models on various domains, since data in different domains can have different types of bias and distributions, such as distinct feedback types.

**Baseline model dependency.** Our experiment on reproducing UFR using six new base recommendation algorithms shows that UFR is algorithm agnostic. In other words, it will improve fairness regardless of the baseline recommendation algorithm. However, the extent of this improvement significantly varies across different algorithms. We conclude that, especially for fairness algorithms that do not assume having access to the relevance judgments, it is useful to evaluate the model on different base ranking models.

**User grouping assumptions.** User fairness arises mainly when we can demonstrate that a particular recommendation algorithm does not truly represent the tastes of one user group while recommendations are provided for other groups that are consistent with their preferences. It is, therefore, crucial to determine if a grouping criterion can capture this unfair treatment properly. For instance, we observed that the interactions cutoff (top-5% users based on the interactions) method proposed [22] does not work on Last.fm dataset, and no unfair behavior is observed on this dataset when grouping is done by this method. Based on the findings of this study, we found that the user grouping method is one of the most important aspects of the user fairness algorithm since it directly affects our interpretation of an algorithm’s fair behavior. Furthermore, our result shows that grouping users according to a cutoff (e.g., top 5% users based on the number of interactions) is a domain-dependent method with poor generalizability. Hence, a new research line would be to define a domain-independent method of grouping users on the basis of multiple features simultaneously. In addition, exploring other user grouping strategies that can show less dependency on the data distribution and domain properties should be explored in this area.

Based on our experiments, we envisage using clustering algorithms (such as K-means) can be a promising future research direction in group fairness problems in recommendation systems.

**Fairness evaluation by multi-sided metrics.** In general, through extensive experiments, we conclude that a well-tuned UFR model can improve user-oriented fairness among groups considerably by achieving higher overall accuracy. This achievement is without sacrificing producers’ exposure or general novelty. However, UFR does not have a statistically significant effect on mitigating popularity bias among users with different degrees of interest toward popular items, i.e., calibration fairness in recommendation system. It is worthwhile to explore methodologies to add calibration fairness and provider aspect to UFR as an extra objective or to add them as additional fairness constraints, thus accounting for the shortcoming and taking into account the needs for a multi-sided marketplace.

Overall, our extensive reproducibility experiments on user-oriented post-processing fairness algorithms revealed several novel observations, suggesting various possible directions in this area, especially on the evaluation of such techniques. In the future, we aim to extend this study to multiple fairness algorithms, each of which focuses on a different aspect of fairness or bias. Our goal is to evaluate the extent of generalizability of our findings in this study and propose a general set of guidelines for a fair evaluation of fairness algorithms. We believe that our experiments and shared resources open opportunities for both reproducibility and evaluation in this area.
