User-oriented Fairness in Recommendation

Yunqi Li
Rutgers University
New Brunswick, NJ, US
yunqi.li@rutgers.edu

Hanxiong Chen
Rutgers University
New Brunswick, NJ, US
hanxiong.chen@rutgers.edu

Zuohui Fu
Rutgers University
New Brunswick, NJ, US
zuohui.fu@rutgers.edu

Yingqiang Ge
Rutgers University
New Brunswick, NJ, US
yingqiang.ge@rutgers.edu

Yongfeng Zhang
Rutgers University
New Brunswick, NJ, US
yongfeng.zhang@rutgers.edu

ABSTRACT

As a highly data-driven application, recommender systems could be affected by data bias, resulting in unfair results for different data groups, which could be a reason that affects the system performance. Therefore, it is important to identify and solve the unfairness issues in recommendation scenarios.

In this paper, we address the unfairness problem in recommender systems from the user perspective. We group users into advantaged and disadvantaged groups according to their level of activity, and conduct experiments to show that current recommender systems will behave unfairly between two groups of users. Specifically, the advantaged users (active) who only account for a small proportion in data enjoy much higher recommendation quality than those disadvantaged users (inactive). Such bias can also affect the overall performance since the disadvantaged users are the majority. To solve this problem, we provide a re-ranking approach to mitigate this unfairness problem by adding constraints over evaluation metrics. The experiments we conducted on several real-world datasets with various recommendation algorithms show that our approach can not only improve group fairness of users in recommender systems, but also achieve better overall recommendation performance.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Computing methodologies** → **Artificial intelligence**.

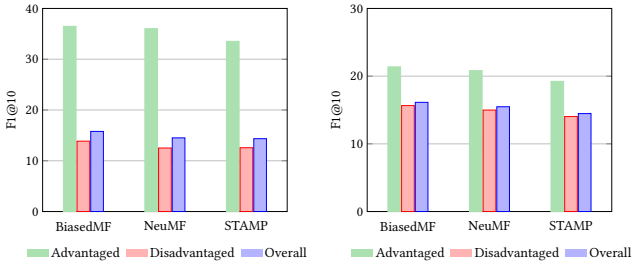
KEYWORDS

Recommendation System; Fairness; Re-ranking; AI Ethics

ACM Reference Format:
Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented Fairness in Recommendation. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3442381.3449866

1 INTRODUCTION

Recently, there has been growing attention on fairness considerations in machine learning community, including classification tasks [12, 31, 40, 44] and ranking tasks [8, 38, 39, 43], etc. Recommendation algorithms can usually be considered as a type of ranking algorithm. However, the ranking problem usually only considers fairness issue from the perspective of items, while the concept of fairness in recommender systems has been extended to multiple stakeholders [9], i.e., the unfairness issue should be considered not only from items or providers side, but also ought to be taken care of from the user side. Comparing with the sufficient work about solving the discrimination from items side in recommendations [2, 3, 5, 21], algorithmic bias existing on the user side has been rarely studied.



(a) Original **(b) Fair Method**

Figure 1: (a) Original results show the significant performance difference between advantaged and disadvantaged groups in F1@10; (b) Fairer performance between two groups and better overall performance achieved from our fair re-ranking method. The results here were obtained by controlling the difference of F1@10 between the two groups by less than a quarter of the original.

In this paper, we consider unfairness issues between different group of users regarding recommendation performance in commercial recommendation scenarios. The unfairness could result from the data imbalance. Researchers have shown that recommender systems may suffer from the item popularity bias, i.e., the popular items can get more exposure than those unpopular ones in recommendations [33]. The underlying mechanism is that popular items will gain more visibility when training the recommendation model due to their sufficient data, and thus the model may be biased to or even dominated by these items.

The data imbalance and algorithmic bias problem that exists on the item side also exists on the user side. Specifically, users who interact with the platform more actively will contribute more sufficient data than those less active users when training the model. Due to the fundamental idea of collaborative filtering in most recommendation algorithms, this would lead to the problem that the trained recommender systems would be biased towards or even

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3449866

dominated by those more active users. As a result, the users with less activity are more likely to receive unsatisfied recommendation results. This will give rise to the unfair treatment between user groups with different activity levels. Here we call such recommendation performance disparity between more and less active user groups as unfairness, since it is caused by the bias that exists in data and the algorithmic bias in some recommendation algorithms. What's more, such unfairness issue can also be a reason of the system's overall performance degradation since those less active users play the majority portion in most cases. Therefore, it is essential to pay attention to the majority of the less active users so that they can enjoy more satisfactory recommendation experience, and thus improve the overall recommendation quality.

To capture users' activity level, we explore three grouping methods through observable information to distinguish users into different activity level, including their number of interactions; total consumption capacity (i.e., the accumulative price the user consumed); and maximum consumption level (i.e., the maximum price of items that the user bought). We label those more active users as advantaged group, while the remaining users as disadvantaged group. The reason for exploring these three methods is that we believe the difference of user interactions and consumption power will reflect their different activity level in a reasonable manner, since usually users who interact with the e-commerce platform more actively will tend to make more purchases, show higher consumption capacity, and have greater consumption budget. We conduct data-driven analysis to explore the performance of some shallow or deep recommender systems on several Amazon datasets under the three grouping methods. Specifically, we discover that the user distribution is more concentrated in the area with fewer interactions or lower consumption, while the advantaged users—who only account for a very small proportion of users in data—enjoy significantly higher recommendation quality than those disadvantaged users, as well as the overall recommendations performance. To solve such unfairness issue, we provide a framework based on re-ranking with fairness constraints to mitigate the performance disparity.

In summary, we aim at mitigating the unfairness of recommendations from the user perspective in this paper. We differentiate users into advantaged and disadvantaged groups in commercial recommendation systems according to their level of activity, and find that disadvantaged users are more likely to receive unsatisfied recommendations because of their insufficient training data compared with advantaged users. To address the unfairness problem above, we provide a re-ranking method with user-oriented group fairness constrained on the recommendation lists generated from any base recommender algorithm. The re-ranking strategy helps to mitigate the recommendation performance bias by taking advantage of making no assumptions of the underlying recommendation model, and offering the model-agnostic flexibility. Our experiments on three Amazon datasets with different types of shallow or deep recommendation algorithms show that our method is not only able to reduce unfairness between two user groups, but also improve the overall recommendation performance. Figure 1(a) shows the significant algorithmic unfairness of the recommendations between advantaged and disadvantaged groups, and Figure 1(b) shows the results generated by our re-ranking method, which provides fairer and best overall recommendation performance.

The key contributions of this paper are as follows:

- We state the importance of concerning the unfairness issue caused by data imbalance between user groups with different activity level in commercial recommendation systems. We explore three methods to capture user activity levels using observable information.
- We provide a fairness constrained re-ranking method and formalize it as a 0-1 integer programming problem to reduce the bias.
- We conduct extensive experiments on three Amazon datasets with four shallow or deep recommendation algorithms to show that our method can shrink not only the fairness disparity between different groups of users, but also improve the overall recommendation quality.

In the following, we review related work in Section 2 and motivate the fairness concerns in Section 3. In Section 4, details of our framework are introduced. Experimental settings and results are provided in Section 5. Finally, we conclude this work in Section 6.

2 RELATED WORK

2.1 Algorithmic Fairness

Fairness is becoming one of the most important topics in machine learning in recent years [20, 36, 38]. There are two basic frameworks adopted in recent studies on algorithmic discrimination: individual fairness and group fairness. Individual fairness requires that each similar individual should be treated similarly, which is hard to define precisely due to the lack of agreement on task-specific similarity metrics for individuals [14]. Group fairness requires that the protected groups should be treated similarly to the advantaged group or the populations as a whole [35]. The group fairness perspective for supervised learning usually implies constraints such as equalized odds and demographic parity. Equalized odds defines the constraint that the false positive rate and true positive rate should be equal for the protected group and advantaged group, which represents the equal opportunity principle [18, 42]. Demographic parity, also called independence or statistical parity, is one of the most well-known criteria for fairness [10]. It requires that decisions should be similar around a sensitive attribute such as gender or nationality. The flaw is that demographic parity will cause a loss in the utility and also infringes individual fairness [14]. Most recent works about fairness concerns have focused on designing algorithms compatible with such fairness constraints on fair classification [40, 44]. The fairness metrics for binary classification problems can be written in terms of rate constraints, which are on the classifier's positive or negative prediction rate for different protected groups [12, 31]. For example, demographic parity posits that the classifier's positive prediction rate is the same across all groups. Such constraints for fairness metrics can be added to the training objective for a binary classifier, and be solved using constrained optimization algorithms or relaxation methods [6, 17]. Here, we address group unfairness in recommender systems from user perspective.

2.2 Fair Ranking

Besides fairness concerns in classification, some recent works have raised the question of fairness in rankings. Recommendation algorithms can usually be considered as a type of ranking algorithm.

However, existing work usually only consider fairness issue from the item perspective in ranking problem, while the concept of fairness in recommender systems is more complicated as the unfair issue will also lie on the user side.

Biega et al. [8] capture unfairness at the level of individual subjects, as such subsume group unfairness. They claim that no single ranking can achieve individual attention fairness, so they propose a new mechanism to quantify and mitigate the position bias, which leads to disproportionately less attention being paid to low-ranked subjects. The results achieve amortized fairness by making the attention accumulated across a series of rankings to be proportional to accumulated relevance. Nowadays, most existing works measure unfairness in ranking at the level of subject groups. The fairness metrics are usually relevant to the exposure of the items belonging to a different protected group. As concluded in [32], the metrics include the supervised criteria which control the average exposure of groups to be proportional to the average relevance of the results of groups to a query [8, 39], and the unsupervised criteria which require that the average exposure at the top of the ranking list is equal for different groups [11, 38, 43]. In a fair ranking problem, some research works directly learn a ranking model from scratch [32, 39, 43], while others consider re-ranking or post-processing algorithms after a ranking has been given [8, 11]. In this paper, we use re-ranking method to reduce discrimination to take advantage of making no assumptions to the underlying recommendation models and offering the flexibility to be applied to different models.

2.3 Fair Recommendation

There has been a small amount of work on fairness in recommendation task, and each work takes very different perspectives. Different from fair ranking and classification, in the field of recommendation systems, the concept of fairness has been extended to multiple stakeholders [9]. The unfair issue can be considered not only from the item or the provider side, but also can be considered from the user side, which makes the problem to be more complex. Both [9] and [1] categorize different types of multi-stakeholder platforms and the different group fairness properties they desired. Mehrotra et al. [29] address the supplier fairness in two-sided marketplace platforms and propose a heuristic strategy to jointly optimize fairness and performance. Patro et al. [34] address individual fairness for both producers and customers, and answer the question of the long-term sustainability of two-sided platforms. Yao and Huang [41] study fairness in collaborative filtering recommender systems, and propose four new metrics that address different forms of unfairness. These fairness metrics can be optimized by adding fairness terms to the learning objective. Lin et al. [27] provide an optimization framework for fairness-aware group recommendation from the perspective of Pareto Efficiency, and further explore the fairness of measure trade-off in recommender systems under a Pareto optimization framework [26]. Beutel et al. [7] show how to measure fairness based on pairwise comparisons from randomized experiments, and offer a regularizer to improve fairness when training recommendation models. Leonhardt et al. [23] quantify the user unfairness caused by the post-processing algorithms which have the original goal of improving diversity in recommendations. Ge et al.[16] explore long-term fairness in recommendation and accomplish the problem through dynamic fairness learning. Fu et al. [15] propose a fairness constrained approach to mitigate the unfairness problem in the context of explainable recommendation over knowledge graphs. They find that performance bias exists between different user groups, and claim that such bias comes from the different distribution of path diversity. Here, we show that such recommendation performance bias also exists in general recommender systems. There are more researches concerning the popularity bias problem in recommendations, i.e., the frequently rated items will get more exposure than those less popular ones. Such researches mainly solve this problem by increasing the number of recommended unpopular items (long-tail items) or otherwise the overall catalog coverage [2, 3, 5, 21]. Abdollahpouri et al. [4] see the problem from the users' perspective with finding how popularity bias causes the recommendations to deviate from what the user expects to get from the recommender system. In this paper, we concern about the unfair issue caused by the bias on the user side, and reasonably divide users into different groups according to their behaviour [24, 25].

Table 1: Percentage of users located at different number of interactions thresholds (as n represents) in the training set of the datasets.

Dataset	$n \geq 5$	$n \geq 10$	$n \geq 20$	$n \geq 30$
Beauty	69.91%	15.40%	3.64%	1.44%
Grocery & Gourmet Food	71.94%	21.22%	6.44%	2.75%
Health & Personal Care	69.04 %	14.96%	3.76%	1.62%

Table 2: Percentage of users located at different total consumption thresholds (as P represents) in the training set of the datasets.

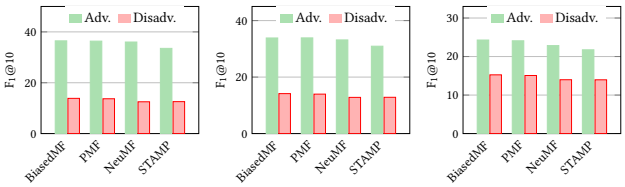
Dataset	$P \geq 50$	$P \geq 100$	$P \geq 200$	$P \geq 400$
Beauty	71.23%	33.49%	10.07%	2.22%
Grocery & Gourmet Food	90.87%	60.11%	20.35%	5.62 %
Health & Personal Care	87.59%	55.15%	20.59%	5.78%

Table 3: Percentage of users located at different maximum price of purchase records thresholds (as P represents) in the training set of the datasets.

Dataset	$P \geq 20$	$P \geq 40$	$P \geq 80$	$P \geq 160$
Beauty	66.43 %	23.20%	6.14%	1.47%
Grocery & Gourmet Food	92.90%	44.38 %	2.40%	0.00%
Health & Personal Care	85.66 %	46.39%	16.91%	5.24%

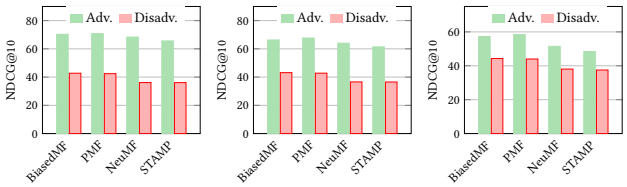
3 MOTIVATING FAIRNESS CONCERNS

In this section, we aim to motivate fairness concerns by conducting data-driven observational analysis to show the unfair performance of current recommender systems. More concretely, we access the imbalanced data distribution on three Amazon Review datasets: **Beauty**, **Grocery & Gourmet Food** (**Grocery**), and **Health & Personal Care** (**Health**), while the details of data are given in Table.4. Furthermore, we show the recommendation performance (F_1@10 and NDCG@10) of several different kinds of traditional fairness-unaware recommendation algorithms on the three datasets



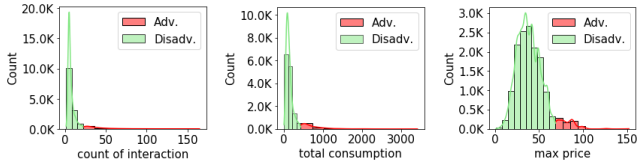
(a) interactions (b) total consumption (c) max price

Figure 2: The difference between advantaged group and disadvantaged group on $F_1@10$ on the Grocery dataset.



(a) interactions (b) total consumption (c) max price

Figure 3: The difference between advantaged group and disadvantaged group on NDCG@10 on the Grocery dataset.



(a) interactions (b) total consumption (c) max price

Figure 4: User distribution of the advantaged and disadvantaged groups under three grouping methods on the Grocery dataset.

to present their unfairness issue, including the shallow models biasedMF [22] and PMF [30], deep model NeuMF [19], and sequential recommendation algorithm STAMP [28].

In this section, we aim to show two facts. The first one is that the majority of users in these datasets only have limited interactions or consumption on the platform, and those users with much more interactions or much higher consumption only account for a small proportion of users. The second fact is that the average recommendation quality on this small group is significantly better than that on the remaining majority of users for all baselines.

Tables 1,2,3 show the users' distribution in three datasets with different number of interactions or consumption. We can see that users are concentrated in areas with less interaction or less consumption. Considering such imbalanced data distribution, we select the top 5% of users in the training dataset ranked by: 1) the number of interactions; 2) total consumption, i.e., the accumulated price of items bought by the user; and 3) the maximum price of items bought by the user, and label them as the advantaged group. The remaining users are labeled as the disadvantaged group. Intuitively, we present the distribution of users in advantaged and disadvantaged groups in **Grocery** as Figure.4, which we can see clearly the difference between the two groups.

Next, we test the recommendation quality of the four baselines on the three datasets. Here we show the recommendation quality of these fairness-unaware recommendation algorithms on **Grocery** in Figure 2 and 3. Similar trends are observed for the other two datasets as well. The details of the experiment results are given in Table 5, Table 6, and Table 7 in later sections. We can see that although the advantaged user group only accounts for a very small proportion of users (5%), they enjoy much higher recommendation quality than those disadvantaged users. This reflects the majority of users are easily disregarded by commercial recommendation engines, which gives rise to unfair recommendations, as well as results in degradation of the overall performance. Therefore, it is important to devise techniques to better serve such users with higher quality recommendations to encourage them to make further interactions with the system, and also to improve the overall recommendation quality since the disadvantaged users are the vast majority.

4 THE FRAMEWORK

In order to address the unfairness concerns presented in Section 3, we provide a framework in this section to generate fair recommendation lists for different user groups, which also has the ability to improve the overall performance through providing more satisfying recommendations to the majority disadvantaged users. We first give the definition of user-oriented group fairness in recommendation systems, and then provide our re-ranking method to formalize the fair recommendation problem under fairness constraints.

In the problem of recommendation, suppose there are user set $\{u_1, u_2, \cdots, u_n\} \in \mathcal{U}$ and item set $\{v_1, v_2, \cdots, v_m\} \in \mathcal{V}$, where $n = |\mathcal{U}|$, $m = |\mathcal{V}|$. Given a recommender system, each user u_i will have a top-N recommendation list $\{v_1, v_2, \cdots, v_N|u_i\}$. As we analyzed in the previous section, recommender systems without considering the user-oriented fairness will generate unfair recommendation lists to different groups of users. To address this issue, we provide a re-ranking framework to generate fair recommendations based on the recommendation lists produced by traditional fairness-unaware recommender systems.

We define binary matrix $\mathbf{W} = \left[\mathbf{W}_{ij}\right]_{n\times N}$ to denote whether an item j is recommended to a user i in fair recommendation lists, where $\mathbf{W}_{ij} \in \{0, 1\}$, and $\{1 \le i \le n\}$ and $\{1 \le j \le N\}$ index users and items respectively. Specifically, if item j is recommended to the user i, then we have $\mathbf{W}_{ij} = 1$, else $\mathbf{W}_{ij} = 0$. We use $\mathbf{W}_{\mathrm{i}} = [\mathbf{W}_{i1}, \mathbf{W}_{i2}, \cdots, \mathbf{W}_{iN}]^{\mathrm{T}}$ to represent the new Top-K recommendation list of user i, where $\sum_{j=1}^{N} \mathbf{W}_{ij} = K, K \le N$. Next we define the user-oriented group fairness in recommender systems following these notations.

4.1 User-oriented Group Fairness

Group fairness requires that the protected groups should be treated similarly to the advantaged group [35]. The group of users can be divided under different requirements for different tasks. In this paper, we consider grouping users as Z_1 and Z_2 so that $Z_1 \cap Z_2 = \varnothing$ in accordance with their different activity level. The notation $\mathcal{M}$ is a metric that can evaluate the recommendation quality such as NDCG@K or F_1 score, and thus we use $\mathcal{M}(\mathrm{W_i})$ to represent the recommendation quality for user i.

The user-oriented group fairness in recommendation is defined as follows:

Definition 1. *User-oriented Group Fairness (UGF)*

$$\mathbb{E}[\mathcal{M}(\mathbf{W})|Z = Z_1] = \mathbb{E}[\mathcal{M}(\mathbf{W})|Z = Z_2] \tag{1}$$

It requires that a fair recommendation algorithm should offer the same recommendation quality for different group of users. Furthermore, we use the difference of average recommendation performance between two groups to measure the user-oriented group unfairness of a recommendation algorithm. We define ε-fairness recommendation algorithm as:

Definition 2. *(ε-fairness) A recommendation algorithm satisfies ε-fairness if:*

$$UGF(Z_1, Z_2, \mathbf{W}) = \left| \frac{1}{|Z_1|} \sum_{i \in Z_1} \mathcal{M}(\mathbf{W_i}) - \frac{1}{|Z_2|} \sum_{i \in Z_2} \mathcal{M}(\mathbf{W_i}) \right| \leq \varepsilon. \quad (2)$$

In this formulation, ϵ represents the strictness of fairness requirements. It also trades off fairness and the recommendation quality of the advantaged group, so that if ϵ approaches to zero, the resulting recommendation will be fairer but will potentially suffer from a huge sacrifice in the recommendation performance of the advantaged group.

4.2 Fairness-aware Algorithm

In this part, we provide a framework which can generate fairness-aware recommendation lists based on a fairness-constrained re-ranking method.

Given a traditional recommender system, each user u_i is recommended with a top-N list $\{v_1, v_2, \cdots, v_N | u_i\}$, each user-item pair is associated with a score $S_{i,j}$ which represents the preference of user i in terms of item j. Here we follow the preference scores calculated by the base recommendation systems. We use these system generated top-N ranking lists as the baseline, and apply re-ranking algorithm to maximize the sum of preference scores under the fairness constraint to generate new fair top-K recommendation lists, where $K \leq N$. Therefore, we can formulate the optimization procedure of the fairness-aware recommendation problem as follows:

$$\begin{aligned} \max_{\mathbf{W}_{ij}} \quad & \sum_{i=1}^{n} \sum_{j=1}^{N} \mathbf{W}_{ij} S_{i,j} \\ \text{s.t.} \quad & \text{UGF}(Z_1, Z_2, \mathbf{W}) < \varepsilon \\ & \sum_{j=1}^{N} \mathbf{W}_{ij} = K, \mathbf{W}_{ij} \in \{0, 1\} \end{aligned} \quad (3)$$

This objective function can be interpreted as that selecting exactly K items out of the baseline top-N list of each user so that the objective function could be maximized. Meanwhile, these selected items need to make the new top-K recommendation list satisfy the constraint defined in Definition 2. The optimization problem here can be solved as a 0−1 integer programming problem. We can find feasible solutions to this NP-complete problem through fast heuristics[1]. Although such methods may not converge to the global optimal solution, our experiments show that we can still obtain satisfactory results in this way. After obtaining the item set which is recommended under fairness constrained, we rank the K items by their original preference score $S_{i,j}$ to construct the final recommendation list.

[1]We use gurobi solver in our experiment. https://www.gurobi.com

Table 4: Statistics of the datasets

Dataset	#Action	#User	#Item	Sparsity
Beauty	198,502	22,363	12,101	99.93%
Grocery & Gourmet Food	151,254	14,681	8,713	99.88%
Health & Personal Care	346,355	38,609	18,534	99.95%

5 EXPERIMENTS

In this section, we first briefly describe the datasets, baselines and experimental setup used for experiments. All source code and dataset of this project has been released publicly[2]. Then we evaluate our proposed fair re-ranking algorithm on top of the baselines to show its desirable performance on both of the fairness metrics and the recommendation performance.

5.1 Experimental Settings

5.1.1 ***Dataset.*** Our experiments are performed on publicly available Amazon 5-core datasets[3], which include user, item, rating information spanning from May 1996 to July 2014 without duplicated interactions. It covers 24 different categories and we take three datasets **Beauty**, **Grocery & Gourmet Food** (**Grocery**), and **Health & Personal Care** (**Health**) to model training and evaluation. The statistics of the datasets are summarized in Table 4. In our experiments, we split each dataset into train (80%), validation (10%) and test sets (10%) and all the baseline models share these datasets for training and evaluation.

For the re-ranking experiment, as stated in Section 3, we select the top 5% users under each grouping type from the training set as the advantaged group and the rest as the disadvantaged group. Then we split the test set based on the two user groups and test the results, respectively.

5.1.2 ***Baselines.*** We take both shallow and deep recommendation models as baselines as suggested in [13]. We compare with two traditional shallow methods (Biased-MF and PMF), one deep model (NeuMF), as well as one sequential model (STAMP). The introduction of baselines are as the following:

- **Biased-MF** [22]: This matrix factorization algorithm takes user, item and global bias terms into consideration.
- **PMF** [30]: This is a probabilistic matrix factorization algorithm by adding Gaussian prior into the user and item latent factors distribution.
- **NeuMF** [19]: This algorithm applies deep neural network with non-linear activation functions to train a user and item matching function.
- **STAMP** [28]: A session-based recommendation model based on attention mechanism, which can capture user's long-term and short-term preferences.

We set the embedding size for users and items to 64 for all the models. For NeuMF, we set the size of multi-layer perceptron (MLP) with 32, 16, 8 as suggested in the paper. The final output layer has only one layer with a dimension of 64. For STAMP, we set the maximum user history length to 30. We apply rectified linear unit (ReLU) non-linear activation function between layers.

[2]Source code available at https://github.com/rutgerswiselab/user-fairness
[3]http://jmcauley.ucsd.edu/data/amazon/

Table 5: The recommendation performance of overall, advantaged, and disadvantaged users of our re-ranking method and corresponding baselines on three Amazon datasets, with the type of grouping users by their number of interactions. The results are reported in percentage (%). All re-ranking results here are obtained under the fairness constraint on F_1. The evaluation metrics here are calculated based on the top-10 predictions in the test set. Our best results are highlighted in bold.

			Beauty				Grocery				Health			
			Overall	Adv.	Disadv.	UGF	Overall	Adv.	Disadv.	UGF	Overall	Adv.	Disadv.	UGF
BiasedMF	F1	Orig.	14.27	30.68	12.77	17.91	15.78	36.48	13.86	22.62	13.92	33.06	12.13	20.93
		Fair	**15.06**	19.18	**14.68**	**4.50**	**16.13**	21.37	**15.65**	**5.72**	**14.54**	19.30	**14.10**	**5.20**
	NDCG	Orig.	43.25	67.79	41.00	26.79	45.08	70.29	42.74	27.55	41.37	66.55	39.01	27.54
		Fair	**43.97**	52.51	**43.19**	**9.32**	**45.75**	55.74	**44.83**	**10.91**	**42.31**	52.26	**41.37**	**10.89**
PMF	F1	Orig.	13.72	30.87	12.15	18.72	15.62	36.37	13.70	22.67	11.34	22.46	10.30	12.16
		Fair	**14.56**	18.86	**14.16**	**4.70**	**16.06**	21.28	**15.58**	**5.70**	**11.42**	14.16	**11.16**	**3.00**
	NDCG	Orig.	41.06	66.74	38.72	28.02	44.85	70.80	42.44	28.36	36.23	52.43	34.72	17.71
		Fair	**41.97**	51.41	**41.10**	**10.31**	**45.74**	57.23	**44.67**	**12.56**	**36.75**	46.39	**35.85**	**10.54**
NeuMF	F1	Orig.	12.44	29.10	10.91	18.19	14.51	36.02	12.51	23.51	12.20	31.84	10.36	21.48
		Fair	**13.81**	17.93	**13.43**	**3.50**	**15.48**	20.81	**14.99**	**5.82**	**12.96**	17.89	**12.49**	**5.40**
	NDCG	Orig.	35.13	61.76	32.69	29.07	38.86	68.34	36.13	32.21	33.55	61.72	30.91	30.81
		Fair	**36.30**	40.89	**35.88**	**5.01**	**40.09**	50.02	**39.17**	**10.85**	**34.30**	42.57	**33.53**	**9.04**
STAMP	F1	Orig.	12.76	27.68	11.38	16.30	14.35	33.52	12.57	20.95	13.15	30.76	11.50	19.26
		Fair	12.76	20.27	**12.07**	**8.20**	**14.47**	19.23	**14.03**	**5.20**	13.15	17.55	**12.74**	**4.81**
	NDCG	Orig.	35.54	58.32	33.45	24.87	38.58	65.61	36.08	29.53	36.53	61.06	34.23	26.83
		Fair	**35.71**	51.02	**34.31**	**16.71**	**39.16**	52.42	**37.93**	**14.49**	**36.69**	46.91	**35.73**	**11.18**

We apply Bayesian Personalized Ranking (BPR) [37] loss for all the baseline models. For each user-item pair in the training dataset, we randomly sample one item that the user has never interacted with as the negative sample in one training epoch. We carefully select the hyper-parameters to tune the models to reach their best performance. The learning rate for training is 0.001, $\ell 2$-regularization coefficient is 0.00001 for all the datasets. The best models are selected based on the performance on the validation set within 100 epochs.

5.1.3 ***Evaluation.*** We use standard metrics F_1@10 score and Normalized Discounted accumulated Gain at rank 10 (NDCG@10) to evaluate the top-10 recommendation quality. The metric $\mathcal{M}$ given in Definition 2 is F_1@10 in all our experiments. For efficiency consideration, we use sampled negative interactions for evaluation instead of computing the user-item pairs scores for each user over the entire item space [45]. For each user, we randomly generate 100 negative samples, which the user has never interacted with, together with the positive samples in the validation or test set to form the user's candidates list. Then we compute the metric scores over this candidates list to evaluate the models top-K ranking performance. The result of all metrics in our experiments are averaged over all users.

5.2 Main Results

In this section, we show the performance of our re-ranking method on both of the recommendation quality and fairness effectiveness compared with traditional fairness-unaware recommendation algorithms.

Table 5, Table 6 and Table 7 show the main results on the three Amazon datasets about dividing user groups based on their number of interactions, total consumption and maximum price respectively. The overall performance, advantaged performance and disadvantaged performance are calculated on the whole test set, the group of advantaged users in the test set, and the group of disadvantaged users in the test set respectively. UGF is computed as Equation 2 to evaluate the difference of recommendation quality (NDCG@K or F_1) between the advantaged and disadvantaged groups. We set the upper bound of the constraint ϵ to the half of the metric differences between two groups of the fairness-unaware baseline. The original results in the table are from baselines, and the fair results in the table are from our model.

Comparing advantaged and disadvantaged groups under four baselines, we can find that there is a big difference in recommendation performance between the two groups. Take the results of NeuMF on Grocery as an example, in Table 5, the difference of NDCG@10 between two groups is 32.21%, and the difference of F_1@10 between two groups is 23.51%; in Table 6, the difference of NDCG@10 is 27.43%, and the difference of F_1@10 is 20.34%; in Table 7, the difference of NDCG@10 is 13.41%, and the difference of F_1@10 is 8.86%. Such disparity could be caused by the nature of collaborative filtering. In other words, the advantaged users may dominate the learning algorithm, and thus the disadvantaged users are more likely to receive biased recommendations due to their insufficient training data, which results in extremely unfair treatments by the system.

We can see from the three tables that our re-ranking method has the ability to significantly reduce the fairness disparity as well as improve the overall recommendation performance of all baselines. For example, also from results of NeuMF on Grocery in Table.5, fair-NeuMF improves the overall NDCG@10 from 38.86% to 40.09%, and improves the overall F_1@10 from 14.51% to 15.48%, as well as reduces the difference between NDCG@10 from 32.21% to 10.85%, and reduces the difference between F_1@10 from 23.51% to 5.82%. What's

Table 6: The recommendation performance of overall, advantaged, and disadvantaged users of our re-ranking method and corresponding baselines on three Amazon datasets, with the method of grouping users by their total consumption. The results are reported in percentage (%). All re-ranking results here are obtained under the fairness constraint on F_1. The evaluation metrics here are calculated based on the top-10 predictions in the test set. Our best results are highlighted in bold.

			Beauty				Grocery				Health			
			Overall	Adv.	Disadv.	UGF	Overall	Adv.	Disadv.	UGF	Overall	Adv.	Disadv.	UGF
BiasedMF	F1	Orig.	14.27	29.76	13.01	16.75	15.78	33.85	14.15	19.70	13.92	33.04	12.31	20.73
		Fair	**15.05**	16.91	**14.90**	**2.01**	**16.14**	17.99	**15.97**	**2.02**	**14.50**	19.11	**14.11**	**5.00**
	NDCG	Orig.	43.25	65.94	41.40	24.54	45.08	66.29	43.16	23.13	41.37	67.72	39.15	28.57
		Fair	**43.82**	47.07	**43.55**	**3.52**	**45.49**	47.97	**45.27**	**2.70**	**42.15**	52.20	**41.31**	**10.89**
PMF	F1	Orig.	13.72	29.74	12.42	17.32	15.62	33.88	13.97	19.91	11.34	22.79	10.38	12.41
		Fair	**14.56**	18,54	**14.23**	**4.31**	**16.00**	20.59	**15.58**	**5.01**	**11.42**	14.28	**11.17**	**3.11**
	NDCG	Orig.	41.06	64.57	39.15	25.42	44.85	67.67	42.78	24.89	36.23	54.80	34.67	20.13
		Fair	**41.98**	50.35	**41.29**	**9.06**	**45.52**	54.40	**44.71**	**9.69**	**36.65**	47.93	**35.70**	**12.23**
NeuMF	F1	Orig.	12.44	28.28	11.15	17.13	14.51	33.16	12.82	20.34	12.20	32.10	10.53	21.57
		Fair	**14.04**	15.91	**13.89**	**2.02**	**15.49**	17.34	**15.32**	**2.02**	**13.06**	14.91	**12.91**	**2.00**
	NDCG	Orig.	35.13	60.00	33.10	26.90	38.86	64.02	36.59	27.43	33.55	61.74	31.17	30.57
		Fair	**36.71**	39.11	**36.51**	**2.60**	**39.84**	41.55	**39.69**	**1.86**	**34.39**	38.60	**34.04**	**4.56**
STAMP	F1	Orig.	12.75	26.71	11.61	15.10	14.35	30.94	12.85	18.09	13.15	30.85	11.66	19.19
		Fair	**12.81**	16.32	**12.53**	**3.79**	**14.45**	18.58	**14.08**	**4.50**	13.15	17.58	**12.78**	**4.80**
	NDCG	Orig.	35.54	57.38	33.75	23.63	38.59	61.41	36.52	24.89	36.53	61.34	34.44	26.90
		Fair	**35.70**	46.08	**34.85**	**11.23**	**38.86**	47.77	**38.06**	**9.71**	**36.66**	47.31	**35.76**	**11.55**

Table 7: The recommendation performance of overall, advantaged, and disadvantaged users of our re-ranking method and corresponding baselines on three Amazon datasets, with the method of grouping users by the maximum price of items they bought. The results are reported in percentage (%). All re-ranking results here are obtained under the fairness constraint on F_1. The evaluation metrics are calculated based on the top-10 predictions in the test set. Our best results are highlighted in bold.

			Beauty				Grocery				Health			
			Overall	Adv.	Disadv.	UGF	Overall	Adv.	Disadv.	UGF	Overall	Adv.	Disadv.	UGF
BiasedMF	F1	Orig.	14.27	21.31	13.86	7.45	15.78	24.29	15.25	9.04	13.92	22.21	13.41	8.80
		Fair	**14.57**	15.51	**14.51**	**1.00**	**15.84**	20.55	**15.55**	**5.00**	**14.15**	15.09	**14.09**	**1.00**
	NDCG	Orig.	43.25	53.59	42.63	10.96	45.08	57.29	44.30	12.99	41.37	53.16	40.64	12.52
		Fair	**43.44**	44.56	**43.37**	**1.19**	**45.09**	52.34	**44.63**	**7.71**	**41.44**	41.81	**41.41**	**0.40**
PMF	F1	Orig.	13.72	20.83	13.30	7.53	15.62	24.09	15.08	9.01	11.34	16.43	11.02	5.41
		Fair	**14.01**	15.80	**13.90**	**1.90**	**15.84**	18.01	**15.71**	**2.30**	**11.35**	12.68	**11.27**	**1.41**
	NDCG	Orig.	41.06	50.35	40.52	9.83	44.85	58.39	43.99	14.40	36.23	46.37	35.60	10.77
		Fair	**41.28**	42.28	**41.22**	**1.06**	**44.98**	48.74	**44.74**	**4.00**	**36.27**	42.06	**35.91**	**6.15**
NeuMF	F1	Orig.	12.44	20.36	11.97	8.39	14.51	22.84	13.98	8.86	12.20	21.12	11.65	9.47
		Fair	**12.81**	14.70	**12.70**	**2.00**	**14.82**	15.76	**14.76**	**1.00**	**12.36**	16.13	**12.13**	**4.00**
	NDCG	Orig.	35.13	46.92	34.43	12.49	38.86	51.48	38.07	13.41	33.55	46.51	32.74	13.77
		Fair	**35.31**	34.87	**35.34**	**0.47**	**38.98**	38.79	**39.00**	**0.21**	**33.66**	39.17	**33.32**	**5.85**
STAMP	F1	Orig.	12.75	19.04	12.38	6.66	14.41	21.75	13.96	7.79	12.86	20.02	12.42	7.60
		Fair	**12.84**	13.98	**12.77**	**1.21**	**14.53**	16.31	**14.41**	**1.90**	**12.88**	16.46	**12.66**	**3.80**
	NDCG	Orig.	35.54	44.27	35.02	9.25	38.16	48.39	37.51	10.88	35.83	46.57	35.17	11.40
		Fair	**35.60**	37.31	**35.50**	**1.81**	**38.22**	41.76	**38.00**	**3.76**	**35.84**	42.27	**35.44**	**6.83**

more, the performance of disadvantaged groups has also been improved due to the fairness constraint. For example, NDCG@10 is improved from 36.13% to 39.17%, and F_1@10 from 12.51% to 14.99%. However, the performance of advantaged users is reduced to satisfy our fairness constraint. Although we sacrifice some of the average recommendation performance of the advantaged users, the constraint that decreases the disparity between two groups substantially improves the performance of the disadvantaged group, which accounts for much more users than the advantaged users. The total performance compromise of the advantaged users is much smaller than the total improvement of the disadvantaged users, which is the reason why the overall performance gets boosted.

Among the three grouping methods, we see that the fairness disparity of dividing users based on their maximum purchase price is not as significant as dividing groups according to their number of records or total consumption. This could be a possible reason that

limits the ability of performance improvement of our re-ranking method. The less significant fairness disparity of the max price grouping method may lie in the following two aspects. On the one hand, although users who interact more actively with platforms tend to have a greater consumption range, using the maximum price to capture users' activity may be less informative than the other two methods, since it cannot properly capture the accumulative influence in the long term. On the other hand, Figure 4 in Section 3 shows the user distribution under three grouping methods. We can see that limited by the maximum price of items in each dataset, the difference of user distributions between advantaged and disadvantaged groups under the maximum price method is not as obvious as the other two methods, thus resulting in less significant unfair treatment by the systems under this grouping method. However, the unfair treatment of recommendation systems is still obvious under max price grouping method, which reflects its capability to capture user activity level.

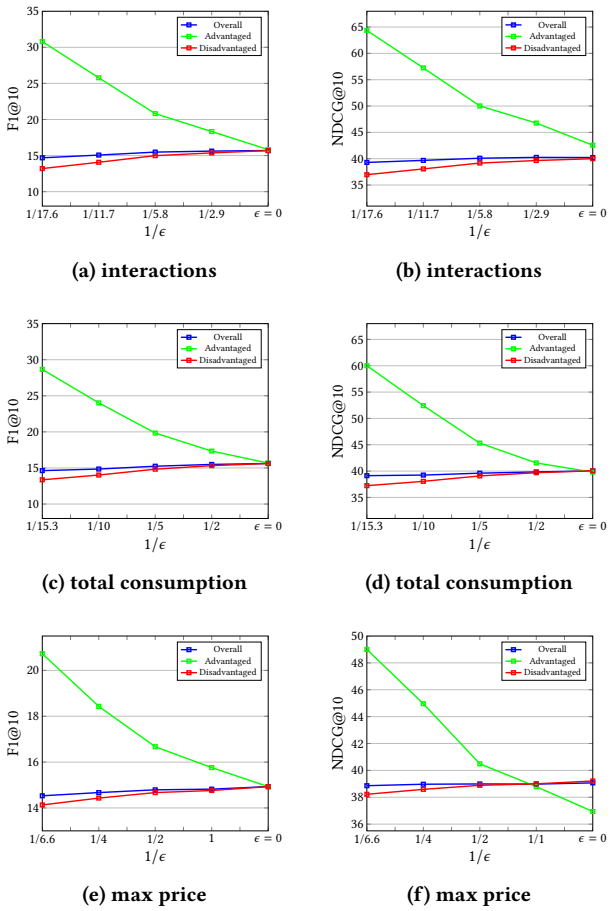
Overall, the experiments show that our re-ranking algorithm can not only shrink the fairness disparity between the two groups of users, but also provide better overall recommendation results than the baseline methods. With these more fairer recommendation lists, those disadvantaged users who account for a large proportion of the user community can get more benefits.

5.3 Ablation Study

In different scenarios, the definition of fairness and the strictness of fairness requirement may be different. From Definition 2, we know that the smaller ϵ is, the fairer our model will be. However, the excessive pursuit of fairness sometimes is unnecessary and will result in a great impact on the recommendation performance. As shown in the main results we presented above, to achieve fairer performance, the average recommendation quality of the advantaged users could be sacrificed, since their original scores were so high that the scores of disadvantaged users could not approach them through directly re-ranking the recommendation lists based on their own preference scores. Therefore, we are interested in how the value of ϵ, which can evaluate fairness between two groups, will affect the recommendation quality of different groups.

In this section, we study how the value of ϵ in Equation 2 can influence the performance of the overall, advantaged group and disadvantaged group. Take the performance of fair-NeuMF on the Grocery dataset as an example, Figure.5 shows how recommendation quality changes with the degree of relaxation of fairness requirements.

All results are under the fairness constraint of F_1@10. From the figures, we can see that the more stringent the requirement of fairness is, the more performance reduction of the advantaged group, and the more performance improvement of overall and disadvantaged group. And the F_1@10 of these three groups will be almost the same when we set $\epsilon = 0$. Although the results are generated under the constraint of F_1@10, we can see that the recommendation quality on NDCG@10 also shows a similar trend in performance change. In Figure.5(f), the performance of the advantaged group is even lower than the disadvantaged group when we set $\epsilon = 0$. Therefore, there may be a trade-off between pursuing fairness and reducing the sacrifice of the advantaged group under some scenarios, although we can still get improvements on the overall performance.



(a) interactions **(b) interactions**

(c) total consumption **(d) total consumption**

(e) max price **(f) max price**

Figure 5: The metrics F_1@10 and NDCG@10 change with respect to the ϵ on Overall, Advantaged and Disadvantaged groups.

6 CONCLUSIONS

In this work, we study the problem of fairness in recommendation algorithms from the user perspective. We show that current recommendation algorithms will capture the data imbalance that lies on the user side, thus produces unfair treatment between different user groups. We first conduct a data-driven observation analysis on three Amazon datasets with several shallow or deep recommendation algorithms, to show that users who interact more actively with platforms only account for a small proportion of users in data. However, the recommendation quality for these advantaged users is significantly higher than those disadvantaged users, which gives rise to unfair issues in recommender systems. The unfair treatment between different groups of users can also reduce the overall performance since the less active users are in the majority. We then quantify unfairness at the group level and provide a fairness constrained re-ranking method to mitigate the unfairness between advantaged and disadvantaged groups while maintaining the recommendation quality. Our extensive experiments show that our method can reduce the unfairness between advantaged and disadvantaged groups significantly, and also improve the overall recommendation quality through providing more satisfying recommendations to the majority of disadvantaged users.

ACKNOWLEDGEMENT

We appreciate the reviews and suggestions of the reviewers. This work was supported in part by NSF IIS-1910154 and IIS-2007907. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

REFERENCES

[1] Himan Abdollahpouri and Robin Burke. 2019. Multi-stakeholder recommendation and its connection to multi-sided fairness. *arXiv preprint arXiv:1907.13158* (2019).
[2] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 42–46.
[3] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2019. Managing popularity bias in recommender systems with personalized re-ranking. *arXiv preprint arXiv:1901.07555* (2019).
[4] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286* (2019).
[5] Gediminas Adomavicius and YoungOk Kwon. 2011. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering* 24, 5 (2011), 896–911.
[6] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453* (2018).
[7] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2212–2220.
[8] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*. 405–414.
[9] Robin Burke. 2017. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093* (2017).
[10] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 13–18.
[11] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2017. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840* (2017).
[12] Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. 2019. Two-player games for efficient non-convex constrained optimization. In *Algorithmic Learning Theory*. 300–332.
[13] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *RecSys*. 101–109.
[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.
[15] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, et al. 2020. Fairness-Aware Explainable Recommendation over Knowledge Graphs. *arXiv preprint arXiv:2006.02046* (2020).
[16] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, and Yongfeng Zhang. 2021. Towards Long-term Fairness in Recommendation. In *Proceedings of the 14th ACM Conference on Web Search and Data Mining (WSDM)*.
[17] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. 2016. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*. 2415–2423.
[18] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.
[19] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *WWW*. 173–182.
[20] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.
[21] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2014. Correcting Popularity Bias by Enhancing Recommendation Neutrality.. In *RecSys Posters*.
[22] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 8 (2009), 30–37.
[23] Jurek Leonhardt, Avishek Anand, and Megha Khosla. 2018. User Fairness in Recommender Systems. In *Companion Proceedings of the The Web Conference 2018*. 101–102.
[24] Shuai Li. 2016. *The art of clustering bandits*. Ph.D. Dissertation. Università degli Studi dell'Insubria.
[25] Shuai Li, Claudio Gentile, and Alexandros Karatzoglou. 2016. Graph clustering bandits for recommendation. *arXiv preprint arXiv:1605.00596* (2016).
[26] Xiao Lin, Hongjie Chen, Changhua Pei, Fei Sun, Xuanji Xiao, Hanxiao Sun, Yongfeng Zhang, Wenwu Ou, and Peng Jiang. 2019. A pareto-efficient algorithm for multiple objective optimization in e-commerce recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 20–28.
[27] Xiao Lin, Min Zhang, Yongfeng Zhang, Zhaoquan Gu, Yiqun Liu, and Shaoping Ma. 2017. Fairness-aware group recommendation with pareto-efficiency. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 107–115.
[28] Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. 2018. STAMP: short-term attention/memory priority model for session-based recommendation. In *SIGKDD*. ACM, 1831–1839.
[29] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. 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*. 2243–2251.
[30] Andriy Mnih and Russ R Salakhutdinov. 2008. Probabilistic matrix factorization. In *Advances in neural information processing systems*. 1257–1264.
[31] Harikrishna Narasimhan. 2018. Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics*. 1646–1654.
[32] Harikrishna Narasimhan, Andrew Cotter, Maya R Gupta, and Serena Wang. 2020. Pairwise Fairness for Ranking and Regression.. In *AAAI*. 5248–5255.
[33] Yoon-Joo Park and Alexander Tuzhilin. 2008. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM conference on Recommender systems*. 11–18.
[34] Gourab K Patro, Arpita Biswas, Niloy Ganguly, Krishna P Gummadi, and Abhijnan Chakraborty. 2020. FairRec: Two-Sided Fairness for Personalized Recommendations in Two-Sided Platforms. In *Proceedings of The Web Conference 2020*. 1194–1204.
[35] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2009. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the 2009 SIAM international conference on data mining*. SIAM, 581–592.
[36] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 560–568.
[37] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).
[38] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2219–2228.
[39] Ashudeep Singh and Thorsten Joachims. 2019. Policy learning for fairness in ranking. In *Advances in Neural Information Processing Systems*. 5426–5436.
[40] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. 2017. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081* (2017).
[41] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems*. 2921–2930.
[42] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*. 1171–1180.
[43] Meike Zehlike and Carlos Castillo. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference 2020*. 2849–2855.
[44] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.
[45] Wayne Xin Zhao, Junhua Chen, Pengfei Wang, Qi Gu, and Ji-Rong Wen. 2020. Revisiting Alternative Experimental Settings for Evaluating Top-N Item Recommendation Algorithms. *CIKM* (2020).