Fairness and Diversity in Recommender Systems: A Survey

YUYING ZHAO, Vanderbilt University, USA
YU WANG, Vanderbilt University, USA
YUNCHAO LIU, Vanderbilt University, USA
XUEQI CHENG, Vanderbilt University, USA
CHARU C. AGGARWAL, IBM T.J. Watson Research Center, USA
TYLER DERR, Vanderbilt University, USA

Recommender systems (RS) are effective tools for mitigating information overload and have seen extensive applications across various domains. However, the single focus on utility goals proves to be inadequate in addressing real-world concerns, leading to increasing attention to fairness-aware and diversity-aware RS. While most existing studies explore fairness and diversity independently, we identify strong connections between these two domains. In this survey, we first discuss each of them individually and then dive into their connections. Additionally, motivated by the concepts of user-level and item-level fairness, we broaden the understanding of diversity to encompass not only the item level but also the user level. With this expanded perspective on user and item-level diversity, we re-interpret fairness studies from the viewpoint of diversity. This fresh perspective enhances our understanding of fairness-related work and paves the way for potential future research directions. Papers discussed in this survey along with public code links are available at: https://github.com/YuyingZhao/Awesome-Fairness-and-Diversity-Papers-in-Recommender-Systems.

CCS Concepts: • **Information systems** → *Recommender systems*; *Information retrieval*; *Data mining*.

1 INTRODUCTION

To tackle the challenges of information overload [11], recommender systems (RS) are playing a crucial role in providing personalized services to fit users' interests. Their effectiveness has been demonstrated across various applications, including news recommendations [79], product recommendations [57, 89], friend recommendations [42, 134], and crystal recommendations [90, 91]. These systems not only improve users' experience but also increase entity exposure, which thereby boosts the profits of content providers. The primary goal of these systems is to improve utility performance (e.g., recall, click-through rate) [10]. However, solely pursuing this goal may lead to practical issues (e.g., Mattew Effect [86], Filter Bubble [88], etc). Consequently, researchers have considered other aspects, such as fairness [77], diversity [69], explainability [29], privacy [62], robustness [30, 74], long-term benefits [28], etc. Acknowledging the significance of beyond-utility perspectives, this survey provides an in-depth discussion of fairness and diversity in RS.

Fairness and diversity are of great importance [63, 69, 77, 117]. Studies have revealed that RS might exhibit unfairness, adversely affecting multiple stakeholders [1, 47]. Given the increasing societal influence, any biases

Authors' addresses: Yuying Zhao, yuying.zhao@vanderbilt.edu, Vanderbilt University, USA; Yu Wang, yu.wang.1@vanderbilt.edu, Vanderbilt University, USA; Yunchao Liu, yunchao.liu@vanderbilt.edu, Vanderbilt University, USA; Xueqi Cheng, xueqi.cheng@vanderbilt.edu, Vanderbilt University, USA; Charu C. Aggarwal, charu@us.ibm.com, IBM T.J. Watson Research Center, USA; Tyler Derr, tyler.derr@vanderbilt.edu, Vanderbilt University, USA.

ACM 2157-6912/2024/5-ART
https://doi.org/10.1145/3664928

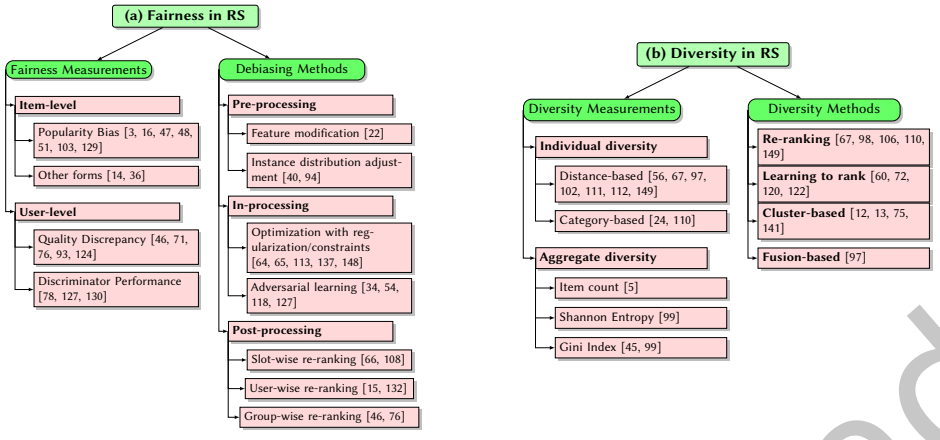

Fig. 1. (a) Fairness in Recommender Systems: fairness measurements and debiasing methods. (b) Diversity in Recommender Systems: diversity measurements and methods to enhance diversity.

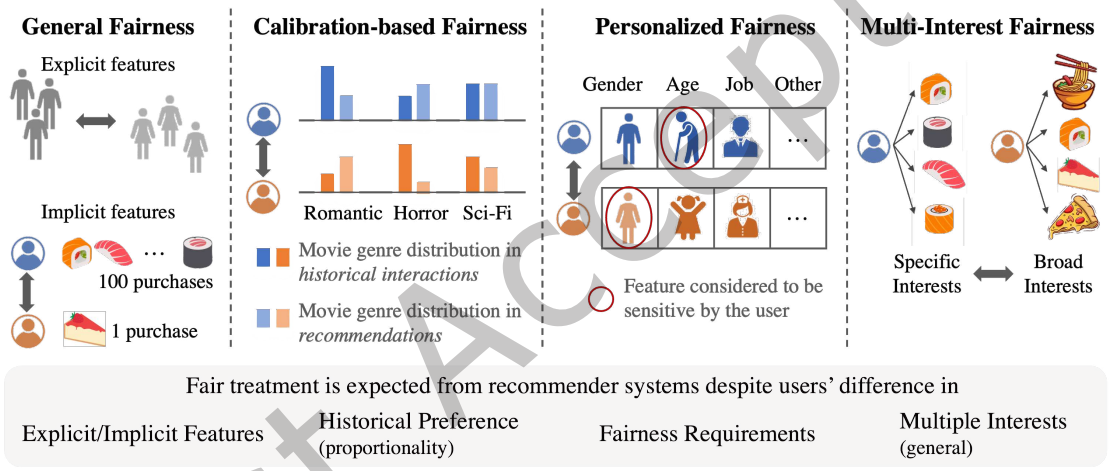

Fig. 2. Fairness and Diversity in RS: users are expected to be treated fairly despite their differences.

within RS have significant impacts. For example, if popular items from big companies dominate the recommendations, the development of small businesses will be hindered, magnifying economic disparities. To address the growing concerns, fairness-aware RS have gained increasing attention and been thoroughly investigated from user and item level. User-level fairness [17, 76, 127] seeks to ensure equitable treatment across different user groups (e.g., groups based on gender, race, etc), while item-level fairness [14, 47, 87] requires that different item groups (e.g., popular and unpopular) have equal opportunities of being recommended. In addition to fairness, diversity is also conducive. Without diversity consideration, RS tend to recommend homogeneous items [126], which may harm customers and providers. For customers, a proliferation of similar items can lead to user fatigue and decreased long-term satisfaction [84]. For providers, small businesses might suffer from low exposure [92] due to the dominance of large companies.

Although fairness and diversity have been exhaustively investigated independently with measurements and methods summarized in Fig. 1(a)(b), their intrinsic connection remains insufficiently explored. The examination of

fairness-diversity relationship presents several benefits, including an enhanced comprehension of the intersection and the revelation of potential research directions which are often overlooked when these domains are studied in isolation. To address this oversight, our analysis encompasses both item and user levels, with an emphasis on the latter, an aspect often underrepresented in diversity research. While both fairness and diversity bear significant implications for users and items alike, discussions on fairness are commonly conducted from both perspectives, whereas diversity tends to be primarily examined in the context of items. Thus, it becomes imperative to expand the scope of diversity to incorporate user aspects as well. We categorize user diversity into *explicit/implicit features, historical preferences (proportionality), fairness requirements, and multiple interests (general)* as shown in Fig. 2. With these expanded diversity definitions, fairness works can be re-interpreted from diversity perspective at both user and item levels. In terms of fairness-diversity connection at the user level, strategies that promote fairness can be construed as mechanisms to alleviate disparate treatment of users, grouped based on different diversity metrics. We further elucidate this by providing a comprehensive table encapsulating these works from the perspective of user diversity. From an item-level standpoint, augmenting item diversity serves as an efficacious strategy for promoting item fairness [81, 107]. We also conduct the experiment to empirically explore the relationship between fairness and diversity.

In conclusion, this survey delivers several key contributions: First, we propose a novel categorization of user diversity, thereby expanding the conventional conceptualization of diversity focusing on the item side. Second, we delve into an exhaustive discussion of fairness-diversity connection at both user and item levels. Our exploration reveals that fairness works can be re-interpreted through the lens of diversity, and strategies enhancing diversity have proven efficacious in improving fairness. Additionally, we delineate pertinent concepts within fairness and diversity individually, advancing existing surveys with more recent works and ensuring the audience is adequately equipped with the requisite contextual understanding prior to delving into their connections.

Relations to other surveys: Various surveys have been published in recent years focusing either on fairness or diversity. Although some of them mention briefly the other aspect (i.e., discuss diversity in fairness surveys [37] or discuss fairness in diversity surveys [126]), none of them have comprehensively discussed the connection between these two domains. In this survey, we aim to fill this crucial gap by focusing on the connections in addition to covering them individually to provide the context. Our aim is not to offer exhaustive discussions on single aspects, as these have been covered in existing surveys for fairness [23, 37, 77, 117] and diversity [19, 63, 69, 126, 128]. Rather, our goal is twofold: firstly, for both fairness and diversity individually, we will augment existing knowledge with more recent developments, recognizing the rapidly expanding body of work in these areas; secondly, we will provide a thorough discussion and categorization of the connection between fairness and diversity. For fairness, we focus on the user and item side. While other more complex categorizations exist (e.g., single-side versus multi-side, dynamic versus static), they are not the major focus of this paper. Regarding diversity, previous works generally discussed it from the item side, while in this survey, we also discuss diversity from the novel user side. The discussions provide new perspectives on understanding fairness from the view of diversity.

Paper organization: The rest of this survey is organized as follows. In Section 2, we introduce RS preliminaries. Next, we discuss fairness and diversity individually in Section 3 and Section 4. Then, in Section 5, we discuss the fairness-diversity connection and extend diversity concepts to user side. Based on the extended diversity concepts, we review the fairness works. We also conduct experiments to empircally investigate the trade-off between fairness and diversity. In the end, we discuss challenges and opportunities in Section 6 and conclude the survey in Section 7.

2 RECOMMENDER SYSTEM PRELIMINARIES

RS are designed to mitigate information overload by recommending items that match users' interests. A typical RS consists of a user set $\mathcal{U} = \{u_1, u_2, ..., u_n\}$ with n users and an item set $\mathcal{I} = \{i_1, i_2, ..., i_m\}$ with m items. The

Table 1. Common notations used throughout the survey and their associated descriptions.

Notation	Description	Notation	Description
K	The number of recommended items	$\mathbf{A}/\mathbf{R}$	Interaction/Predicted preference matrix
λ	Coefficient	$\mathbf{W}$	Binary matrix of recommended items
i^*	Selected item to be added into recommendation	$\mathbf{E}_U/\mathbf{e}_u$	User embeddings/user u's embedding
K'	The number of user interests	$\mathbf{E}_I/\mathbf{e}_i$	Item embeddings/item i's embedding
a	Attention	$\mathbf{Z}_u$	u's representation with multiple interests
$\mathcal{U}$	User set with n users	$Q(\cdot)$	Recommendation quality function
$\mathcal{I}$	Item set with m items	$d(\cdot,\cdot)$	Distance function
$\mathcal{R}_u^K/\mathcal{R}_u$	Top K recommendation list for user u	$f_{\text{rec}}(\cdot)$	Function to measure utility performance
$\mathcal{G}_i$	i-th user/item group	$f_{\text{fair}}(\cdot)$	Function to measure fairness performance
$\mathcal{C}$	Item candidate set (re-ranking)	$f_{\text{div}}(\cdot)$	Function to measure diversity performance
$\mathcal{S}$	(In)complete recommendation set (re-ranking)	$\mathcal{L}_{\text{rec}}$	Loss term for utility performance
$\mathcal{D}$	Provider set	$\mathcal{L}_{\text{fairness}}$	Loss term for fairness performance

user-item historical interactions are represented by a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ where $\mathbf{A}_{ui}$ denotes whether user u has interacted with item i. We note that in most cases $\mathbf{A}_{ui} \in \{0, 1\}$ for binary interactions but could also be weighted, e.g., rating score and purchasing number. The primary goal of recommendations is to predict the list of top K items for each user given the historical user-item interactions $\mathbf{A}$. The top K items are selected based on the preference/relevance scores, which are the dot products between user embedding $\mathbf{E}_{\mathcal{U}}$ and item embedding $\mathbf{E}_{\mathcal{I}}$. For instance, the relevance score between user-item pair (u, i) is $\mathbf{e}_u{}^\top \mathbf{e}_i$ where $\mathbf{e}_u$ and $\mathbf{e}_i$ are the embeddings of user u and item i. After obtaining the scores, K items with the highest values are recommended. To learn the representations for preference calculation, RS can be divided into collaborative filtering (CF)-based and content-based. CF-based methods [9, 119] utilize the user-item interactions and recommend items based on users with similar interaction patterns. Content-based methods [6, 7], on the other hand, usually use additional features (e.g., user profiles) to assist the recommendation process. We direct interested readers to existing RS surveys/books [10, 143] for further details.

Formally, the problem definition of standard recommendation is established as follows: given the interaction matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, the goal is to learn the function $f : \mathcal{U} \times \mathcal{I} \rightarrow \mathbf{R}$, such that the predicted preference matrix $\mathbf{R}$ approximates the true (including unobserved) preference of the users for the items as closely as possible. For top K recommendations, the system recommends the top K items to each user, with these items having highest scores, and formalized as $\mathcal{R}_u^K = \{i | i \in \mathcal{I} \ \ \& \ \ \mathbf{R}_{ui} \in \text{top}K(\mathbf{R}_u)\}$. The notations in this survey are summarized in Table 1. Regarding user and item embeddings, for simplicity, we use $\mathbf{e}_u$ and $\mathbf{e}_i$ to denote user and item embeddings if no confusion. Otherwise, we will use $\mathbf{e}_i^U$ and $\mathbf{e}_i^I$ to distinguish user and item embeddings with the superscripts.

3 FAIRNESS IN RECOMMENDER SYSTEMS

In this section, we introduce the fairness measurements to quantify fairness and debiasing methods to enhance fairness. A summary is shown in Figure 1 (a).

3.1 Fairness Measurements

As one of the most representative multi-stakeholder systems, RS raise fairness concerns from both item (i.e., item-level fairness) and user (i.e., user-level fairness) sides. While other categories exist (e.g., group vs individual), we refer readers to other survey [77, 117] for a comprehensive discussion. In this survey, we focus on user-level and item-level fairness, given their innate links to diversity.

3.1.1 Item-Level Fairness. Item-level fairness focuses on the fair treatment of items during recommendations. One of the most predominant concerns in RS, particularly at the item level, is popularity bias [3], where generally RS tend to recommend popular items to users. Popularity bias would lead to exposure unfairness. In this context, exposure refers to the chance of an item being recommended which is measured by its occurrence in the top K

recommendation. Consequently, popular items receive more exposure and thereby gain more popularity, widening the disparity between popular and less-popular items. This will be detrimental to users, the provider selling the items, and the platform [2, 40]: (1) Users' preference towards unpopular items would be under-represented due to the majority training towards popular items; (2) It becomes hard for the growth of small businesses even if they can provide items with similar quality with popular items; (3) If the producers cannot sell products and make benefits, highly likely they will leave the platform, which may inform corporate monopoly and is unhealthy for the platform development in the long run. Typically, exposure fairness measures the difference in exposure for the items belonging to distinct groups, which is defined based on various constraints, including demographic parity constraints [47, 48, 103] and Extract-K fairness constraint [47]. Next, we first give the formal definition of exposure and define fairness metrics based on exposure function. Note that other functions can also be used to calculate exposure. Interested readers could refer to paper [38] where they provide a summary table and detailed exposure metrics and corresponding fairness metrics.

Definition 3.1. Exposure measures item occurrences in users' recommendation. If an item is recommended to more users, this item has a higher exposure.

$$\text{Exposure}(i) = \sum_{u \in \mathcal{U}} \mathbb{1}(i \in \mathcal{R}_u). \tag{1}$$

Definition 3.2. Demographic parity-based exposure fairness requires the average exposure of item groups to be equal. It is defined as:

$$\left| \frac{1}{|\mathcal{G}_1|} \sum_{i \in \mathcal{G}_1} \text{Exposure}(i) = \frac{1}{|\mathcal{G}_2|} \sum_{i \in \mathcal{G}_2} \text{Exposure}(i) \right|, \tag{2}$$

where $\mathcal{G}_1$ and $\mathcal{G}_2$ are groups divided by item's popularity.

To allow a flexible adjustment in practice, Extract-K fairness constraint [47] introduces α.

Definition 3.3. Extract-K-based exposure fairness requires the exposure of various groups are statistically indistinguishable from a given maximum α:

$$\frac{\sum_{u \in \mathcal{G}_1} \text{Exposure}(i)}{\sum_{i \in \mathcal{G}_2} \text{Exposure}(i)} = \alpha \tag{3}$$

When $\alpha = \frac{|\mathcal{G}_1|}{|\mathcal{G}_1|}$, Eq. (3) equals Eq. (2). While Eq. (3) and Eq. (2) are strictly fair, it is challenging to achieve this goal in practice. Therefore, exposure fairness is often defined in the disparity form.

Definition 3.4. Disparity-based exposure fairness measures the difference between exposures:

$$F(\mathcal{G}_1, \mathcal{G}_2) = \left| \frac{1}{|\mathcal{G}_1|} \sum_{i \in \mathcal{G}_1} \text{Exposure}(i) - \frac{1}{|\mathcal{G}_2|} \sum_{i \in \mathcal{G}_2} \text{Exposure}(i) \right|, \quad \text{or } = \left| \sum_{i \in \mathcal{G}_1} \text{Exposure}(i) - \alpha \sum_{i \in \mathcal{G}_2} \text{Exposure}(i) \right|$$

A lower score indicates a higher level of fairness. We note other forms have been studied to measure popularity bias [16, 51, 129] and item fairness besides popularity bias [14, 36]. Additionally, the above exposure fairness definitions are for two groups. If there are more than two groups, the definitions can be extended to consider the disparities across all pairs of groups.

3.1.2 User-Level Fairness. There are two directions of user-level fairness. One is based on recommendation quality discrepancy, and the other is whether the recommendation encodes sensitive features. We denote the first as *Quality Discrepancy* and the second as *Discriminator Performance* as whether encoding sensitive features is commonly measured with discriminator performance.

Quality Discrepancy: Fairness on the user side is related to the recommendation quality, which can be defined as the performance gap among groups and Gini coefficient [50].

Definition 3.5. Performance unfairness (group level) measures the gap in recommendation performance between groups.

$$\text{Unfairness}_{\text{Gap}}(\mathcal{G}_1, \mathcal{G}_2, \mathbf{R}) = \left| \frac{1}{|\mathcal{G}_1|} \sum_{u \in \mathcal{G}_1} Q(\mathcal{R}_u) - \frac{1}{|\mathcal{G}_2|} \sum_{u \in \mathcal{G}_2} Q(\mathcal{R}_u) \right|, \tag{4}$$

where $\mathbf{R}$ is the recommendations obtained from a recommender system with $\mathcal{R}_u$ denoting the top K recommendation list for user u, and Q is a general quality measurement (e.g., F1, NDCG).

Under this framework, various works define different quality measurements and leveraging different group partition strategies. Fu et al. [46] investigate the unfairness between active users and inactive users based on the number of purchases. They use F1 and NDCG as recommendation quality metrics and also propose explanation quality metrics related to the diversity of explainable paths in the knowledge graph as another quality measurement. Li et al. [76] divide users into the advantaged and disadvantaged groups according to multiple criteria including interaction number, total consumption, and max prize. They use F1 and NDCG as the quality metrics. Rahmani et al. [93] conduct comprehensive experiments on various domains and datasets based on [76]. In addition to the groups divided by the level of activity (i.e., interaction number), they also investigate the unfairness between advantaged and disadvantaged groups based on the consumption of popular items. Wu et al. [124] investigate the unfairness between cold and heavy users which are divided by the number of historical news clicks. They use AUC as the quality measurement. They also define unfairness based on performance gap. However, different from Eq. (4), they obtain the optimal checkpoints performance for all users and cold users respectively and measure the gap.

Definition 3.6. Performance unfairness (individual level) measures pairwise performance disparity between instances without group concept. It is defined based on Gini coefficient where the pairwise disparity between two users is normalized by the average performance:

$$\text{Unfairness}_{Gini}(\mathbf{R}) = \frac{\sum_{u_1, u_2 \in \mathcal{U}} |Q(\mathcal{R}_{u_1}) - Q(\mathcal{R}_{u_2})|}{2|\mathcal{U}| \sum_{u \in \mathcal{U}} Q(\mathcal{R}_u)}. \tag{5}$$

Similar to the group performance gap mentioned above, Fu et al. [46] define the quality as regular NDCG and F1 and also the explanation quality as Q in Eq. (5). Leonhardt et al. [71] define Q based on the preference scores of the recommended items and their top K items.

Discriminator Performance: RS aims to learn high-quality user representations which encode users' preferences for downstream recommendations. It is critical to investigate whether the learned embeddings are fair. This involves sensitive features (e.g., gender) and a discriminator trained to predict sensitive features given user embeddings. If the performance (e.g., accuracy) is low for the discriminator, the recommendation model satisfies the fairness requirement in the embedding space, indicating a high level of fairness. Wu et al. [127] study gender bias and use AUC to measure the discriminator performance of binary classification to mitigate the impact of data imbalance. When the sensitive attribute is not binary (i.e., multiple values), they use micro-averaged F1 measure. Similarly, Wu et al. [130] use F1 score as the measurement for the discriminator performance, where users are allowed to choose sensitive features (i.e., personalized fairness). Li et al. [78] also study personalized fairness and use AUC for discriminator performance.

3.2 Debiasing Methods

Numerous efforts have been devoted to designing debiasing methods [77, 117]. According to the phase of the intervention, they can be summarized into (1) *pre-processing methods* which debias the data before training; (2) *in-processing methods* which incorporate fairness consideration into training process; (3) *post-processing methods* which adjust the recommendation after the model is trained. These methods can be used individually or simultaneously in different phases.

3.2.1 Pre-processing Methods. Pre-processing methods aim to adjust the training data so that it contains less bias. Therefore, the model trained on unbiased data would be fairer. To achieve this, one can either modify the features of the training data [22] or change the instance distribution (e.g., delete, add, re-sample) [40, 94] without feature update. One naive way, which falls in the category of suppression, is to remove sensitive features from input features [70]. However, on the one hand, simply removing features might hurt recommendation performance. On the other hand, features are correlated, and thus removing the sensitive features cannot guarantee fairness [118]. Therefore, other ways are developed based on orthogonalization [22] and marginal distribution mapping [22]. Additionally, adversarial approaches are used to learn fair features but they are more commonly twined with the downstream tasks. Therefore, we categorize them into in-processing methods. Additionally, instances can be adjusted in the training set without changing the features. For example, Rastegarpanah et al. [94] add antidote/fake data to the training dataset following data poisoning attacks, while in [40] they use re-sampling to adjust the proportion of users in groups.

The benefit of pre-processing methods is that they only change the input data and thus existing models can be applied to the adjusted dataset, providing significant flexibility. However, while mitigating bias from the dataset, relevant information to the downstream task might also be removed. This will result in uncertainty in the tradeoff between utility and fairness.

3.2.2 In-processing Methods. Various techniques can guide the training process of the model towards higher utility and better fairness, we introduce the two major directions.

Optimization with Regularization and Constraint: Regularization-based methods integrate fairness with utility in the optimization objective by introducing a fairness regularizer [64, 109, 113]. The new overall loss term is formulated with a Lagrange multiplier as $\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{fairness}}$, where $\mathcal{L}_{\text{rec}}$ is the general recommendation objective, $\mathcal{L}_{\text{fairness}}$ is the fair regularization, and λ is the coefficient to balance two goals. Some works add an independence regularizer to encourage the recommendation to be independent of the sensitive features. Such independence terms include mean matching [64], distribution matching [65], and mutual information [65]. Error correlation loss [113] is designed to regularize the correlation between prediction errors and the distribution of market segments. Four unfairness metrics [137] are proposed as the regularizer, measuring the discrepancy between the prediction behavior of the disadvantaged and advantaged groups. [136] advanced a similar unfairness loss based on Distributionally Robust Optimization(DRO) technique. A tensor-based fairness-aware RS (FATR) [148] is proposed that adds an orthogonality term between the representations of users/items and the corresponding vectors of sensitive features. Counterfactual graphs are generated and utilized in the regularization [27]. [25] add regularization based on data augmentation via generating 'fake' interaction data. [138] add regularization by conducting data augmentation for minority group which utilizes interactions of mainstream users.

The main idea of constrained optimization [103, 104] is similar to regularization-based methods where fair constraints are included during optimization. However, it has a different form as $\textit{minimize} \quad \mathcal{L}_{rec} \quad \textit{s.t. Fairness constraints}$. The difference between optimization with regularization and constraints is that the former tolerates unfairness where the solution might fall into unfair regions but the latter disallows unfairness enforced by the fairness constraints.

Adversarial learning: The main idea is to learn fair representations irrelevant to sensitive features [118, 127]. To achieve this, a discriminator to predict the sensitive label and a generator to generate fair representations are trained. During adversarial learning, the discriminator gains the ability to predict the sensitive label while the generator is optimized to fool the discriminator by generating fair representations so that the discriminator cannot determine whether the representation contains sensitive features. By playing the min-max game between the discriminator and generator, sensitive information will be removed from the final learned representations and only relevant information for the downstream task is maintained. There have been many works in this direction [26, 34, 54, 58, 59, 118, 127], primarily following the aforementioned setup.

The in-processing methods allow more control in utility-fairness trade-off during training. However, they might be designed for specific models and cannot be generalized to other models.

3.2.3 Post-processing Methods. Re-ranking methods are widely used, as post-processing approaches, to adjust the recommendations generated by the model to promote fairness [76, 108, 117, 132]. According to the granularity of adjustment at each time, there are three different re-ranking types [117]: Slot-wise re-ranking [66, 108], User-wise re-ranking [15, 132], and Group-wise re-ranking [46, 76].

Slot-wise re-ranking: These methods add items from candidate list sequentially to recommendation list (i.e., item by item) by following rules considering relevance and fairness simultaneously. The greedy algorithm [66, 108] to select the item with the maximum marginal gain is as follows:

$$i^* = argmax_{i \in C \setminus \mathcal{S}} \lambda f_{\text{rec}}(u, \mathcal{S} \cup \{i\}) + (1-\lambda) f_{\text{fair}}(u, \mathcal{S} \cup \{i\}),$$

where C is the candidate item set with high relevance scores calculated by the trained RS, and $\mathcal{S}$ is the current generated recommendation list. Function f_{rec} measures utility, and f_{fair} measures fairness. The item which will bring the largest contribution to the existing recommendation list will be added to the list from C, leading to an updated list $\mathcal{S} = \mathcal{S} \cup \{i^*\}$. The items will be iteratively added until the recommendation list reaches the required length.

User-wise re-ranking: These methods generate the whole recommendation list for a user at once. A popular strategy is integer programming [31, 123]. The main idea is to treat decisions as variables (e.g., whether an item is in the recommendation list [132] or is in a particular position [15]) and transform the re-ranking problem into an integer programming problem with fairness constraints.

Group-wise re-ranking: Re-ranking at a group level considers several users together rather than adjusting the list for a single user. Similar to user-wise re-ranking, integer programming can also be leveraged. Li et al. [76] use a binary matrix $\mathbf{W} \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{I}|}$ to represent whether an item will be recommended in the top K list for each user. They solve the following optimization problem:

$$\max_{\mathbf{W}} \sum_{i=1}^{|\mathcal{U}|} \sum_{j=1}^{|C|} \mathbf{W}_{ij} \mathbf{R}_{ij}, \quad \text{s.t. GUF}(Z_1, Z_2, \mathbf{W}) < \epsilon, \sum_{j=1}^{|C|} \mathbf{W}_{ij} = K, \mathbf{W}_{ij} \in \{0, 1\},$$

where GUF is the user fairness constraint between two groups and $\mathbf{R}_{ij}$ is the relevance score which predicts the user i's preference towards item j. The optimization problem aims to maximize the relevance score while subjecting to fairness constraint. Another work [46] uses a similar strategy in explainable recommendation with knowledge graphs and adds another fairness constraint for explanation fairness, which is an area with growing attention [53, 145].

The post-processing methods provide model-agnostic flexibility. However, the improvement is restricted by the results from base models, leading to a suboptimal solution. In other words, if the base model generates extremely biased results, the adjustment based on it might be limited.

4 DIVERSITY IN RECOMMENDER SYSTEMS

As another essential beyond-utility perspective, diversity is comprehensively discussed here from three dimensions: (1) the source of diversity, (2) the measurements to quantify diversity, and (3) the methods to promote diversity. Figure 1(b) summarizes the diversity measurements and methods. Various other concepts such as *Serendipity* and *Novelty* are closely related to diversity but slightly different [8]. We refer readers to [63] if interested in a comprehensive discussion on these concepts.

4.1 Source of Item Diversity

Diversity is generally discussed from the item side [67, 97, 99, 111, 149]. In most papers, it is defined based on the redundancy or similarity among the recommended items, where the detailed difference might come from the categories [24, 110] or the distance in item embedding space [102, 112]. Items are naturally different from each other, and the level of diversity is different across items. We will discuss item diversity from three aspects: (1) When category information is available, the items belonging to the same category share higher similarity than items from different categories. For example, two movies both from romance will generally be more similar compared to two motives from romance and horror genre respectively. (2) Within the same category, items might still have different features. Take movies as an example, two movies are both categorized as romance, but they have different stories, topics, budgets, language, etc. (3) Besides the intrinsic features, item diversity is implicitly revealed from users' history interactions. According to the collaborative filtering effect, similar users prefer similar items. Therefore, user interactions can be utilized as diversity indicator. In summary, item diversity might come from categories, at a high level, or from item features, at a detailed level. Item features can be explicit (i.e., intrinsic item features) or implicit (i.e., from learned embeddings considering both item features and other users' history interactions).

4.2 Measurements to Quantify Item Diversity

Diversity can be categorized into individual diversity (i.e., individual-level) which focuses on a single recommendation list and is relevant to each user's satisfaction individually [67, 97, 102], and aggregate diversity (i.e., system-level) which aims to capture the diversity across all recommendations and is relevant to the fairness of providers [5, 45, 99, 135].

4.2.1 Individual Diversity. Individual diversity can be further split into distance-based and category-based. They rely on the distance between different item representations or genre information.

Distance-based diversity: The most widely used definition is called *Intra-List Diversity (ILD)* which measures the pairwise item diversity within one recommendation list.

Definition 4.1. Intra-List Diversity (ILD) is formally defined as follows where the distance function d measures the dissimilarity/distance between items:

$$ILD(\mathcal{R}_u) = \frac{1}{|\mathcal{R}_u|(|\mathcal{R}_u| - 1)} \sum_{i \in \mathcal{R}_u} \sum_{j \in \mathcal{R}_u \setminus i} d(i, j). \tag{6}$$

The variants differs in the way of obtaining item embeddings and the specific distance function. Items can be represented by content descriptors [149], rating scores [67, 111], latent item representations from matrix factorization [102, 112], etc. For distance, various functions have been applied, e.g., Hamming distance [67], Gower diversity [56], complement of cosine similarity [97], Jaccard similarity [111], or Pearson correlation [111]. A larger *ILD* indicates a higher level of diversity.

Category-based diversity: Distance-based diversity is criticized in [110] due to the failure of ensuring the consistency between the diversity value and users' experience. They [110] leverage category/genre information to capture item diversity, which better corresponds to users' perceptions. They propose a novel binomial framework to capture genre-based diversity [110] which considers three perspectives simultaneously: coverage, redundancy, and size-awareness. When genre information is available, they compute the diversity score as $\text{BinomDiv}(\mathcal{R}_u) = \text{Coverage}(\mathcal{R}_u) * \text{NonRed}(\mathcal{R}_u)$ where coverage mainly captures how many different genres are presented in the recommendation and non-redundancy (i.e., NonRed) encourages the genre uniqueness in the recommendation. When explicit category information is unavailable, Chen et. al [24] propose to group items into categories based on their attributes. Thereafter, diversity is defined as the category dissimilarity.

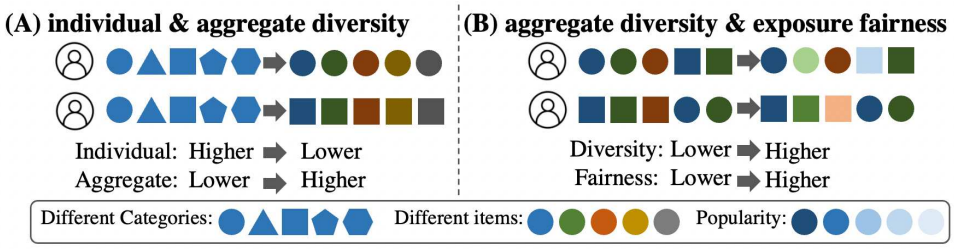

Fig. 3. (A) A toy example to show the difference between individual and aggregate diversity; (B) A toy example to show the connection between aggregate diversity and exposure fairness. Different shapes correspond to different categories, colors correspond to different items, and levels of the darkness of the same color correspond to different levels of popularity (the darker color indicates the more popular item).

4.2.2 Aggregate Diversity. Unlike individual diversity which corresponds to a single recommendation list, aggregate diversity considers all recommendations in an aggregated way. If a RS always recommends popular items rather than a diversified list, from a global view, the aggregate diversity would be poor. Therefore, it reflects the system's ability to recommend less popular or hard-to-find items and thus is related to exposure fairness [16, 103, 129]. The most intuitive definition is to count the number of total diverse items being recommended [5].

Definition 4.2. Aggregate Diversity (Count) is defined as the length of the recommendation set of all users: $Aggdiv = |\bigcup_{u \in \mathcal{U}} \mathcal{R}_u|$.

This measurement focuses on whether the item is recommended but ignores how many users this item is recommended to, which motivates other work to investigate whether such recommendation is evenly distributed when taking the detailed user number into account. For instance, aggregate diversity has been formulated with Shannon entropy (H) [99] and Gini index (Gini) [45, 99].

Definition 4.3. Aggregate Diversity (Shannon Entropy) is defined as:

$$H = -\sum_{i \in \mathcal{I}} p(i) log_2 p(i), \quad p(i) = \frac{|\{u \in \mathcal{U} | i \in \mathcal{R}_u\}|}{\sum_{j \in \mathcal{I}} |\{u \in \mathcal{U} | j \in \mathcal{R}_u\}|}, \tag{7}$$

where $p(i)$ measures the probability of item i being in the recommendation list for all users.

Definition 4.4. Aggregate Diversity (Gini Index) is defined as:

$$Gini = \frac{1}{|\mathcal{I}| - 1} \sum_{k=1}^{|\mathcal{I}|} (2k - |\mathcal{I}| - 1) p(i_k), \tag{8}$$

where p shares the same meaning as in the Shannon entropy formula.

4.2.3 Discussion. Individual and aggregate diversity measure diversity from two distinct levels: individual and system, respectively. High individual diversity does not imply high aggregate diversity and vice versa. For example, in Figure 3(A), the recommendation could be of high individual diversity for each user but of low aggregate diversity from a system level, or the recommendation could be of low individual diversity for each user but provide high aggregate diversity.

4.3 Methods to Promote Diversity

Various methods have been proposed to promote diversity. They can be categorized into four types [19]: re-ranking-based, learning to rank, cluster-based, and fusion-based.

Algorithm 1: Greedy Algorithm for Diversity Enhancement

Input: Recommendation number top K, coefficient λ
return *Recommendation list* $\mathcal{R}_u$
Init recommendation list $\mathcal{R}_u = \{\}$
for $|\mathcal{R}_u| < K$ **do**
  $i^* = argmax_{i \in I \setminus \mathcal{R}_u} s(\mathcal{R}_u \cup \{i\}, \lambda)$
  $\mathcal{R}_u = \mathcal{R}_u \cup \{i^*\}$
return $\mathcal{R}_u$;

(1) Re-ranking methods aim to adjust the ranking from existing RS by combining the diversity constraints to improve individual or aggregate diversity. The traditional rank score is solely based on relevance, while the new one [67, 149] is composed of relevance and diversity in the form of:

$$s(\mathcal{R}_u, \lambda) = \frac{1-\lambda}{|\mathcal{R}_u|} \sum_{i \in \mathcal{R}_u} f_{\text{rec}}(i) + \lambda f_{\text{div}}(\mathcal{R}_u), \tag{9}$$

where $f_{\text{rec}}(i)$ denotes the relevance score of item i, $f_{\text{div}}(R)$ is the diversity score of the recommendation list $\mathcal{R}_u$, and λ is the coefficient to trade-off the utility and diversity goals. As discussed in Section 4.2, various diversity metrics can be adopted for computing $f_{\text{div}}(\mathcal{R}_u)$. After having the new score, a greedy algorithm [18] called Maximum Marginal Relevance (MMR) is performed iteratively to select the item with the maximum score until reaching the expected recommendation length as shown in Algorithm 1. Many works follow this greedy framework [67, 98, 106, 110, 149].

Beyond the greedy framework, other re-ranking algorithms mainly rely on solving constraint optimization problems to find the optimal list. [140] transforms the problem into a series of objective functions with different utility-diversity constraints (e.g., maximizing the relevance under the constraint that diversity is larger than a diversity tolerance, maximizing the diversity when the relevance is larger than a matching tolerance). The benefits of these post-processing methods are that they are model-agnostic and time-efficient (i.e., avoid the computation of re-training). However, the improvement might be restricted by the initial recommendation list.

(2) Learning to rank methods, are in-processing methods that enforce adjustment during training process. Typical RS are trained with utility loss functions such as Bayesian Personalised Ranking (BPR) loss [95] or rank-based loss which aims to improve utility performance, several works add diversity objectives to the existing one so that during the training, the model is trained towards higher utility and to avoid monotony simultaneously. [122] continues and explores the work in [60], which incorporates diversity criteria in regularization terms, by proposing several regularizations on learned item embeddings to incorporate diversity. The regularizations are as follows:

$$reg(\mathbf{E}_U, \mathbf{E}_I) = \sum_{i,j} d(i,j) \|{\mathbf{e}_I}^i - {\mathbf{e}_I}^j\|^2, \quad = \sum_{u,i,j} d(i,j) ({\mathbf{e}_U}^{u\top} (\mathbf{e}_I^i - {\mathbf{e}_I}^j))^2, \quad \text{or } = \sum_{u,i,j} d(i,j) {\mathbf{e}_I}^{i\top} {\mathbf{e}_I}^j, \tag{10}$$

where $\mathbf{E}_U$ is the learned user representation, $\mathbf{E}_I$ is the learned item representation, and $d(i,j)$ is a pre-defined distance between item i and item j. Generally, when the distance $d(i,j)$ is large, the representations between these two items will be trained to be closer due to the regularization. Therefore, more diverse items will be recommended. [72] takes genre diversity into account in the reward model for a multi-armed bandit recommendation. More recently, diversity-aware deep ranking network [120] is proposed to generate accurate and diversified recommendation list during ranking phase. Compared with the post-processing methods, this line of research involves re-training, which might be less efficient. The advantage is that it can ensure diversity in the recommended lists, owing to incorporating diversity considerations during training.

(3) Cluster-based methods leverage the principle that similar items will be grouped into the same cluster, and therefore to promote diversity, items from different clusters rather than a single cluster should be recommended.

Following this idea, different methods are proposed to select items from various clusters [12, 13, 75, 141]. For example, [141] clusters items based on user profiles and recommend items that match these individual clusters rather than the whole user profile. [12] proposes ClusDiv which assigns how many items (i.e., weight) each cluster should be recommended for each user. It initially assigns the weights based on the recommendation list from traditional RS and then adjusts the weights by iteratively decreasing the number of items within the category that is larger than a threshold and increasing the number of items in the least recommended category. After having the adjusted weights, items based on this assignment will be selected for recommendation. For comparison, [141] performs clustering based on local information according to user's tastes while [12] is based on global information.

(4) Fusion-based methods aggregate results from different RS. While a single RS might provide recommendations with high utility performance but low diversity, different RS will provide high-quality recommendations while obtained recommendations are dissimilar from each other. Therefore, model fusion can be leveraged to obtain a result with both high utility and diversity. [97] fuses the rating scores to generate new aggregated scores. It proposes a multi-objective framework that considers utility, diversity, and novelty simultaneously. It first adopts multiple existing RS to generate the predicted ratings for user-item pairs and fuses the predictions by a weighted summation $\hat{r_{ui}} = \sum_{t=1}^{T} w_t r_{ui,t}$, where T denotes different RS, $r_{ui,t}$ is the rating score estimated from the t-th RS, and the weights w_t is the strength for considering the corresponding model during aggregation, which are learnable with a strength Pareto evolutionary algorithm [150, 151]. The aggregated ratings $\hat{r_{ui}}$ are then used for generating the recommendation list.

5 FAIRNESS AND DIVERSITY

As different beyond-utility perspectives, fairness and diversity are generally investigated separately. However, there are various connections that will be discussed in Section.5.1. Based on the connections and differences, we will first summarize diversity from user perspective and then discuss the fairness works for user diversity and item diversity.

5.1 Connections and Differences

With the rapid development of RS, utility performance (e.g., accuracy) is no longer the sole golden standard for determining the quality of the system. There have been various extra considerations which are closely related to users' satisfaction. Fairness and diversity are two such beyond-utility perspectives with each highlighting different aspects. From an ethical perspective, fairness and diversity commonly appear in the same sentence as they are similar in their meaning. Specifically, bias exists when diverse groups are treated dissimilarly and fairness requires these diverse groups to receive similar treatments. From a restricted standpoint, such diversity refers to sensitive features (e.g., race, gender, etc.) which is typically discussed in the fairness field. However, there are also fairness works beyond discussing such sensitive features. They can be summarized in a more comprehensive view. From a more broad standpoint, there are various diversities based on which fairness is proposed. As fairness is generally discussed from both the user and item sides, we naturally discuss connections from the item level and user level as illustrated in Figure 4.

Item-level: as mentioned in Section 4, aggregate diversity is related to exposure fairness. When a RS tends to always recommend popular items which shows a lack of exposure fairness, the aggregate diversity will be low since there will be a large overlapping of popular items. When improving exposure fairness by recommending more unpopular items, the chances that the same unpopular item being recommended to different users will be low, therefore, the aggregate diversity will probably increase. A toy example is shown in Figure 3(B).

User-level: user-level connection is less intuitive than item-level. One of the most important reasons is that typical diversity is discussed from the item perspective. As fairness is discussed at both item and user levels,

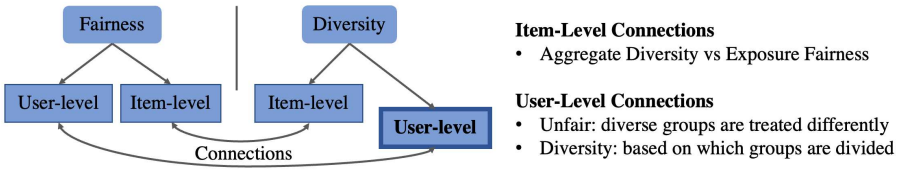

Fig. 4. Fairness and Diversity: in the context of recommender systems, they are commonly investigated separately. However, the connections at item and user level highlight the significance of intersections.

we first summarize diversity from a user perspective, namely, user diversity, which is rarely discussed in the previous literature. User diversity includes explicit/implicit features, historical preferences (proportionality), fairness requirements, and multiple interests (general) and will be summarized in Section 5.2. From the first three perspectives, there are corresponding fairness studies related to that type of user diversity, indicating their strong connections. The last user diversity opens up future direction for fairness in terms of multiple interests.

5.2 User Diversity

While diversity is typically discussed within the context of items, diversity also exists from the aspect of the user side. Users have diverse preferences and interests, which are reflected in their history interactions. In this section, we summarize the user-level diversity from *explicit/implicit features, historical preferences (proportionality), fairness requirements, and multiple interests (general)*. While historical preferences and multiple interests are all related to user preference, the former focus on genre proportionality and the latter is more general.

Explicit/Implicit Features: Users have diverse properties in terms of their inherent features (i.e., explicit features) and behavior features extracted from their interactions (i.e., implicit features). Inherent features include age, gender, race, etc. Behavior features include the interaction number, the average price from interacted items, and so on. These features are sometimes treated as sensitive features from a fairness perspective [44, 61, 76, 124, 127].

Historical Preferences (Proportionality): Users have diverse preferences which are reflected in their history interactions. Specifically, it is reflected in the proportion of interacted genres/categories in the history logs. For example, two users who watched both romance movies and action movies will share similar interests. However, their focus might be different as one user watched 70 romance movies and 30 action movies while another user watched 50 romance and 50 action movies. It is reasonable to expect the recommender system to provide a personalized recommendation list that reflects such differences. This important property is known as calibration [108], which is proposed to avoid the issue of utility-oriented systems where the user's lesser interest gets crowded out by the main interest. The idea of proportionality was first proposed in [35].

Fairness Requirements: Although it is obvious to perceive the users' diversified needs that results in the personalization requirements on recommender system, the other key aspect of this survey, fairness, receives different attitudes from users. First, users would have different tolerance levels of fairness. For example, in the microlending platform [81], lenders' tolerance of fair consideration related to regions varies greatly. Some lenders prefer offering loans to certain regions like their home countries, while others may be open to diverse regions. Secondly, users might treat different features as sensitive features. For instance, some users treat gender as a sensitive feature since they do not want the recommendations to be influenced by this feature, while others may care more about the age feature than the gender feature [78].

Multiple Interests: Unlike focusing on the proportionality of interacted genres, interests capture more general user preference towards certain items. It provides a high-level depiction of the user. Traditional RS learn a single embedding to represent the user. Recently, researchers have proposed that this will lose information during aggregation and cannot fully depict users' diverse interests. Therefore, a line of research focuses on using multiple embeddings to represent users' diverse interests [20, 73]. Additionally, a similar topic is called

Table 2. A summary table of fairness for user diversity. Papers and available codes can be accessed fromlink.

Fairness (Diversity)	Work		Intervention	Approach	Keywords	Code
General Fairness (Explict/ Implict Features)	Gender	[36]	-	-	Fairness Evaluation	✗
		[125]	In-processing	Adversarial learning	News Recommendation	✓
		[131]	In-processing	Constraint-based	Reciprocal	✗
		[17]	In-processing	Regularization-based	Multi-side fairness	✗
	Race	ALG [52]	Post-processing	Re-ranking	Theoretical	✓
		[49]	-	-	Uncertain Inference	✓
		MSRec [146]	Post-processing	Re-ranking	Dating	✗
		FATR [148]	In-processing	Constraint-based	Tensor-based	✓
	Age	[36]	-	-	Fairness Evaluation	✗
		PSL [43]	In-processing	Logical rules	Hybrid	✗
	Behavior	[76]	Post-processing	Constraint-based	User activeness	✓
		[55]	In-processing	Constraint-based	Pareto Optimality	✗
		[46]	Post-processing	Constraint-based	Explainable	✓
		PFGR [133]	In-processing	Constraint-based	Group recommendation	✗
Calibration-based Fairness (Historical Preferences - proportionality)	[108]		Post-processing	Re-ranking	Calibrated	✗
	[3]		-	-	Popularity bias	✗
	[4]		Post-processing	Re-ranking	Popularity bias	✗
	[33]		Post-processing	Re-ranking	Multiple fairness metrics	✓
	[144]		Post-processing	Re-ranking	Taste distortion	✗
Personalized Fairness (Fairness Requirements)	FAR/PFAR [81]		Post-processing	Re-ranking	Diversity tolerance	✗
	OFAiR [107]		Post-processing	Re-ranking	Multi-aspect fairness	✗
	[78]		In-processing	Adversarial learning	Causal notion	✓
	PFRec [130]		Post-processing	Adversarial training	Prompt-based	✓
	UCRS [114]		In-processing	Counterfactual inference	Filter Bubbles	✓
	CUFRL [32]		In-processing	Regularization-based	Fair representation	✗
Multi-Interest Fairness[1] (Multiple Interests - general)	MIND [73]		-	-	Dynamic routing	✓
	ComiRec [20]		-	-	Self-attention	✓
	Re4 [142]		-	-	Backward flow	✓
	PinnerSage [89]		-	-	Cluster-based	✗
	MIP [101]		-	-	Time-aware	✓
	MacridVAE [83]		-	-	VAE	✓
	DisenGCN [82]		-	-	Neighborhood routing	✓
	DGCF [115]		-	-	Collaborative filtering	✓

disentangled learning [82, 115], where the single embedding is disentangled to multiple sub-embeddings such that these latent sub-embeddings would reflect different intentions towards various items. While users have multiple interests/intentions during decision-making, different users have diversified interests.

5.3 Fairness for User Diversity

We discuss fairness works for proposed diversity, which are summarized in Table 2 and Fig. 2.

5.3.1 General Fairness. Most works related to user fairness fall into this category where groups are divided by explicit sensitive features (e.g., gender, race, religion) [44, 61, 127] or implicit sensitive features (e.g., degree) [76, 124]. The goal is to ensure that diverse groups would receive similar treatments. To achieve this, methods discussed in Section 3.2, such as regularization-based optimization and adversarial learning, can be applied. Specifically, these works focus on the explicit sensitive features which annot be easily changed by users, including gender [17, 36, 125, 131], race [49, 52, 146, 148], age [36, 43]. Other works focus on implicit features extracted from user behaviors (e.g., interaction number, price of purchased items) [46, 55, 76, 133].

5.3.2 Calibration-based Fairness. Calibration in classification requires that the predicted class distribution aligns with the input data's actual distribution. Similarly, calibration in recommendations aims to ensure that the genre distribution in the recommended list aligns with that in users' historical interactions, thereby more accurately reflecting users' interests. For instance, in movie recommendations encompassing various movie categories like action and romance, if a user watches 90% action movies and 10% romantic movies, the recommendation list should reflect this property rather than exclusively recommending all action movies. Recommending items without following users' history preference would lead to unfairness in terms of preserving preference proportionality. To

[1]To the best of our knowledge, there are no related multi-interest fairness works and we review works related to multi-interest. Due to the same reason, while summarizing these works, we will not categorize them in the intervention and approach columns designed for fairness works.

better preserve preference proportionality, Steck [108] propose calibrated recommendation with post-processing methods to adjust the genre distribution with maximum marginal relevance [18]. The main idea is to first obtain recommendations with baseline recommender systems and then update the list to fit the distribution in historical interactions. As demonstrated in Eq. 11, the adjusted recommendation list $\mathcal{R}_u^*$ considers two aspects (1) utility, indicated by the relevance score $f_{\text{rec}}(\mathcal{R}_u)$, (2) fairness, indicated by the distribution divergence where p denotes the genre distribution in users' historical interactions and q represents the distribution in the recommendation and C_{KL} measures the Kullback–Leibler(KL)-divergence between these two distributions.

$$\mathcal{R}_u^* = argmax_{\mathcal{R}_u, |\mathcal{R}_u|=K}(1-\lambda) f_{\text{rec}}(\mathcal{R}_u) - \lambda C_{KL}(p, q(\mathcal{R}_u)) \tag{11}$$

Minimizing the divergence allows the recommendation to better align with user interests reflected in interactions, with λ controlling the trade-off between utility and fairness. Inspired by Steck's work, Abdollahpouri et al. [3] investigate unfairness of popularity bias from the user perspective. They discover that the property of consistent genre distribution in recommendation and interactions is maintained at varying levels for users with diverse interests in popular items. Users are categorized into three groups based on the ratio of popular items in their historical interactions: Niche Users (those with the least interest in popular items), Blockbuster-focused Users (those with the highest interest in popular items), and Diverse Users (all others). They find that Niche Users suffer more from popularity bias. Because although these users are interested in unpopular/long-tail items, the recommendation still concentrates on popular items. This finding suggests that the preferences of Niche Users are not as well-served as those of Blockbuster-focused Users. After observing this phenomenon, in a follow-up work [4], Abdollahpouri et al. design a metric called User Popularity Deviation (UPD) to quantify the popularity bias from a user-centered view and propose a re-ranking method called Calibrated Popularity which is similar to Eq. 11 with a different fairness term based on Jensen–Shannon divergence. Silva et al. [33] conduct a comprehensive experiment to investigate calibration in fair recommendations. They evaluate six recommender algorithms applied in the movie domain and analyze variations of three fairness measures, including three distribution metrics: Kullback–Leibler, Hellinger, and Pearson chi-square [21]. Previous works generally assume static user interest. However, Zhao et al. [144] emphasize that interest would evolve over time. They extend the static setting by predicting the future genre distribution that matches the user interest and conducting the calibration based on the predicted distribution rather than the one extracted from historical interactions.

5.3.3 Personalized Fairness. Users have diverse fairness demands. They have different levels of demands/tolerance on fairness consideration. For example, some users put more emphasis on fairness than others. Additionally, individuals also differ in which features should be considered sensitive. For instance, some people think age is a regular feature that helps recommend popular songs among their peers and some consider it as sensitive as they would like to follow the trend and not limited by age. Therefore, this suggests fairness may need to be personalized to individuals. Liu et al. [81] investigate personalized fairness in microlending, where recommender systems are designed to recommend loans to lenders. In the design, besides the fairness need that borrowers from diverse demographic groups should have a fair chance of being recommended, they also consider personalized fairness on the lender side. Lenders' receptivity to the diversification of recommended loans varies greatly. Some lenders prefer certain regions while others are open to diverse areas. They use the information entropy [100] to identify the lender diversity tolerance and assign it as the weight of fairness term in the re-ranking process so that different lenders show personalized focus on the fairness aspects. Sonboli et al. [107] follow their work and extend it into the scenario with multiple sensitive features. They propose a more fine-grained personalization according to each feature. For different protected features, users have different tolerance which is also measured by information entropy. Li et al. [78] study personalized fairness based on the causal notion. Users are allowed to specify sensitive features. Feature-independent user embeddings are generated so that the recommendation outcomes maintain the same in the counterfactual world where the other features are unchanged except for specified sensitive features. They design an adversarial learning approach to remove sensitive information while

keeping relevant information for the recommendation. Wu et al. [130] define selective fairness task in sequential recommendation where users can flexibly choose which features will be considered sensitive. To satisfy users' diverse fairness demands, they adopt a pre-training and prompt-tuning framework. A traditional recommender system without fair consideration is obtained via pre-training, and diverse fairness needs are satisfied with both task-specific and user-specific prompts using adversarial training. Wang et al. [114] present the initial step towards personalized filter bubbles mitigation. While the filter bubble issue [88] would lead to recommending homogeneous items, it is unreasonable for users to passively accept the recommendation strategy to mitigate such an issue. Their work proposes a new framework called user controllable recommender system which allows users to actively control the mitigation of filter bubbles. Cui et al. [32] tackle the limitation of existing works that only focus on debiasing pre-defined sensitive features. However, users might be interested in several sensitive groups which are unknown in advance. To solve this, they propose controllable universal fair representation learning to make the representations fair to all possible sensitive attributes.

5.3.4 Multi-Interest Fairness. To the best of our knowledge, no studies have yet investigated fairness related to multiple interests. In this section, we review research on multi-interest and suggest potential future directions for fairness in the context of multiple interests. Traditional RS utilize a single embedding to learn user preferences. However, users might have diverse interests that cannot be adequately represented by one embedding. For instance, a user could be interested in sports (e.g., basketball) and art (e.g., painting) simultaneously. One single embedding representing the overall interest could be insufficient to identify these interests and to make corresponding recommendations. Therefore, a single representation would lead to a sub-optimal solution. To mitigate this, researchers use multiple embeddings to represent diverse interests. User u has embedding $\mathbf{Z}_u \in \mathbb{R}^{d\times K'} = \{\mathbf{z}_u^k\}_{k=1,...,K'}$ where $\mathbf{z}_u^k$ denotes user u's k-th interest among all K interests. After learning the representations, the relevance score of user u to item i is calculated by $max_{k=1}^{K'} \mathbf{e}_i{}^T \mathbf{z}_u^k$ where $\mathbf{e}_i$ is the item representation. The items with the highest scores are recommended to the user.

There have been various ways to learn multi-interest user representations ($\mathbf{Z}_u$). MIND [73] represents the initial effort where they extract the interests based on dynamic capsule routing. ComiRec [20] utilizes self-attention to learn multiple interests based on item interactions. Denote user's interaction as $x_1, ..., x_{nx}$ where nx is the length of interacted items and the representation of i's item in the interaction as $\mathbf{e}_i$. The attention of how much an item is correlated with k-th interest is computed with the softmax operations. The user's k-th interest is the weighted average of the interacted items based on the calculated attention. These are shown in Eq. 12.

$$a_{k,i} = \frac{\exp(\mathbf{w}_k^T \tanh(\mathbf{W_1}\mathbf{e}_i))}{\sum_j \exp(\mathbf{w}_k^T \tanh(\mathbf{W_1}\mathbf{e}_j))}, \text{ where } \mathbf{z_k} = \sum_j a_{k,j}\mathbf{W_2}\mathbf{e}_j. \tag{12}$$

Following ComiRec, which considers the item-to-interest relationship, Re4 [142] proposes backward flow to model interest-to-item relationship by adding three regularizations including re-contrast which leverages contrastive learning to learn distinct interest representations, re-attend to ensure that the learned attention correlates to the relevance score for recommendations, and re-construct to highlight that interest representations should reflect the content of representative items. PinnerSage [89] clusters items from the users' interactions with the Ward hierarchical clustering method [121] and uses one representative in each item cluster to represent one of the user's interests. The representative embedding is selected by minimizing the sum of distance with the items in the same cluster. In their work, they find that while the proposed multi-interest strategy improves utility performance, it also increases recommendation diversity. MIP [101] also utilizes cluster-based methods to achieve multi-interest. They assign each interest as the latest item representation in each cluster. In addition, rather than treating each interest with uniform importance, they learn the weight to represent the preference over each interest embedding.

There is another line of similar research called disentangled learning. Unlike the multi-interest work that assigns multiple interest embeddings to each user, disentangled learning seperate user embeddings into several sub-embeddings that each represent one interest/intention. MacridVAE [83] performs disentanglement at both a macro (i.e., to buy a shirt or a cellphone) and a micro (i.e., the size or the color of the shirt) level based on VAE [68, 96]. Macro disentanglement is achieved by learning several prototypes based on users' intentions. Micro disentanglement is realized by magnifying the KL divergence. DisenGCN [82] updates the traditional aggregation strategy where nodes gather information from all neighbors uniformly or based on degrees. It aggregates information from related neighbors according to their closeness with the factors obtained from Softmax. For example, sports-related factors will be mainly updated by items like baseball rather than paintings. DGCN [115] enhances DisenGCN by applying the distance correlation for factor independence and a new aggregation mechanism. It models the distribution over intents for each user-item interaction and iteratively refines the intent-aware interaction graphs and representations.

While multi-interest and disentanglement strategies have demonstrated effectiveness in improving utility performance, most of these studies assume that all users have the same number of interests/intentions. However, users' interests are diverse, and the level of diversity might vary between individuals. Some users might be interested in many aspects, while some users have more specific preferences towards certain categories. Naturally, users with diverse interests should be assigned more interest numbers to capture this diversity. Thus, it could be unfair to assign the same number to every user. The impacts of multi-interest or disentanglement for users with varying levels of interest diversity (to the best of our knowledge) have not been investigated, and potential unfairness problems might arise. In addition to the user-side unfairness, investigating item-side fairness is also valuable, as one potential reason for improved performance might be recommending more popular items, thereby increasing popularity bias.

5.4 Fairness for Item Diversity

On the one hand, diversifying recommendations helps discover potential user interests to improve user experience, as discussed in Section. 4. On the other hand, it helps improve the item visibility [128], especially those unpopular items from small providers that initially have a low opportunity of being recommended. Researchers have highlighted one of the benefits of diversifying is to satisfy the equal market exposure of providers [126], which naturally connects diversity with a fairness point of view. More specifically, aggregate diversity measures system-level diversity which reflects the systems' ability to recommend less popular or hard-to-find items and thus is relevant to the exposure fairness among providers. Liu et al. [80, 81] propose two fairness-aware re-ranking extensions called Fairness-Aware Re-ranking (FAR) and Personalized Fairness-Aware Re-ranking (PFAR) based on xQuAD [98] which is designed for result diversification. FAR enhances diversity by boosting the scores for items that belong to new providers. Following the general framework in Alg. 1, they substitute the objective in line 4 into the following:

$$v^* = \text{argmax}_{v \in \mathcal{R}_u \setminus \mathcal{S}} P(v|u) + \lambda \sum_{d \in \mathcal{D}} P(d|u) \mathbb{1}_{v \in d} \prod_{i \in S} \mathbb{1}_{i \notin d}, \tag{13}$$

where $\mathbb{1}$ is the indicator function and $\mathcal{D}$ is the provider set. The first term corresponds to utility performance based on relevance scores, and the second term is to assign an incentive score for the provider that never appears in the existing recommendation list. If none of the recommended items belongs to provider d, the second term is effective, Otherwise, the second term equals zero. In this way, it increases the chance of small providers being recommended. PFAR incorporates personalized consideration based on the assumption that users have different tolerance to the level of diversification. They obtain a tolerance score τ_u based on information entropy and update Eq. 13 by multiplying this weight in the second term with λ.

Following Liu et al [80, 81], Sonboli et al [107] also consider fairness-promoting diversity to improve provider fairness. The main idea is to increase the diversity of recommended item list, which benefits the protected class

Table 3. Empirical relationship between fairness and diversity in RS. All metrics are the higher the better. The best performance of each metric is marked in bold.

Model	Utility↑	User fairness↑	Item fairness↑	Individual diversity↑	Aggregate diversity↑
Vanilla (MF)	**0.298**	0.856	0.029	0.682	0.730
+User fairness	0.255	**0.953**	0.029	0.708	0.711
+Item fairness	0.257	0.810	**0.094**	0.530	0.781
+Individual diversity	0.272	0.857	0.058	**0.759**	0.793
+Aggregate diversity	0.296	0.832	0.032	0.685	**0.955**

the most. They update the second term in Eq. 13 so that the objective of re-ranking is to find the adjusted recommendation list that maximizes the relevance while maximizing the diversity (i.e., minimizing the similarity). The similarity metric takes into account the item's difference from the existing list and users' tolerance of diversification towards one feature based on information entropy. Specifically, they define a weighted cosine similarity as:

$$\text{wcos}(\mathbf{e}_i, \mathbf{e}_j, \mathbf{t}_u) = \sum_{f}^{|F|} \mathbf{t}_{ut}\mathbf{e}_{if} \times \mathbf{e}_{jf} \frac{1}{\sqrt{\sum_f \mathbf{t}_{uj}\mathbf{e}_{if}^2} \times \sqrt{\sum_j \mathbf{t}_{uj}\mathbf{e}_{jf}^2}}, \tag{14}$$

where $\mathbf{e}_i$ and $\mathbf{e}_j$ are the representations of two items, $\mathbf{e}_{if}$ is the f-th feature in the representation, and $\mathbf{t}_u$ is user's tolerance of diversification. Based on the pair-wise similarity function, the similarity between one item to an existing recommendation list is defined as $\text{sim}(\mathcal{S}, i) = \sum_{i' \in \mathcal{S}} \text{wcos}(\mathbf{e}_i, \mathbf{e}_{i'}, \mathbf{t}_u)$. Compared with FAR and PFAR where the second term loses the boosting effect once the provider has appeared in the recommendation list, Eq. 14 continues work. Additionally, this work provides a more fine-grained personalization considering per feature for users. FA*IR [139] enhances diversity to improve group fairness in another way. They create two queues for protected and unprotected items and integrate them to satisfy a probabilistic ranked fairness test. In this way, they ensure that the proportion of protected candidates would remain statistically above a given minimum.

5.5 Empirical Investigation between Fairness and Diversity

We empirically investigate the relationship between fairness and diversity through experiments on MovieLens dataset[2] where the gender attribute and popularity property[3] are regarded as sensitive features in user and item sides, respectively. Due to space limitation, it is infeasible to conduct a comprehensive study on various base models, fairness/diversity measurements and algorithms which requires dedicated efforts in a new survey. We train the base model based on the representative RS matrix factorization (MF) [95], optimize and evaluate according to specific measurements and methods for improving fairness and diversity. Regularization-based [137] and reranking-based [67] methods are adopted to enhance fairness and diversity. Fairness is defined as the ratio discrepancy between advantage and disadvantage groups. Aggregate diversity is the total number of different recommended items and we adopt ILD in Eq. 6 as individual diversity where the distance function is the cosine distance between pre-trained embeddings. We repeat the experiments three times for each method and report the average evaluation. Note that the learned embeddings in models with regularizations are different from those of vanilla model, this makes individual diversity based on the embedding space incomparable. Therefore, we leave out comparing the individual diversity for fair models. Reranking does not change embeddings and we will include the discussion for the diversity-enhanced methods. From the results in Table 3, we draw the following observations:

[2]https://grouplens.org/datasets/movielens/1m/
[3]Top 20% items with highest interaction numbers are regarded as popular and the remaining as less-popular.

- User fairness regularization achieves a significant improvement on the user fairness while the other metrics remain at a similar level with vanilla model.
- When item fairness improves, the aggregate diversity improves. It indicates that unpopular items are probably recommended due to fairness requirement and thus aggregate diversity is improved.
- When individual diversity increases, aggregate diversity also increases since new items are recommended. The item fairness improves since unpopular items have more chance to be recommended based on diversity consideration. Improving individual diversity cannot ensure increasing aggregate diversity or item fairness, but empirically we observe the positive effect.
- With reranking to enhance aggregate diversity, the aggregate diversity metric improves significantly while other metrics maintain similarly. There is a slight improvement in individual diversity but it is not as strong as the gain brought by increasing individual diversity to aggregate diversity. Similar case happens for item fairness. The reason for marginal change in other aspects is that few instances need to adjusted for a high aggregate diversity. Therefore, the performance are close to vanilla model. By recommending new items that have never been recommended globally, there is no guarantee on improvement of fairness aspect or individual diversity.

6 CHALLENGES AND OPPORTUNITIES

Researchers have raised the awareness of beyond-utility perspectives to evaluate recommendation including fairness and diversity and have also started the exploration of their intersections. While there exist many challenges, there are also valuable open opportunities for future research directions.

- **Understanding the Relationship between Fairness and Diversity:** This survey delves into the high-level conceptual connections between fairness and diversity and present an initial empirical discussion. However, their practical relationship is intricate and multifaceted. Various metrics exist for evaluating fairness and diversity. It is recommended to conduct extensive empirical studies to determine the feasibility of simultaneously optimizing these metrics or if they inherently conflict. A quantitative analysis of their mathematical relationships is essential. Additionally, how to incorporate diversity into fairness or vice versa rather than treating them as two goals during the method design is also interesting. Discussions on other types of fairness-diversity intersections beyond user-level and item-level (e.g., single-side vs multi-side) are encouraged.
- **Multi/Many-objective Selection, Optimization and Evaluation:** There are trade-offs between utility and beyond-utility objectives, and also within beyond-utility objectives with diverse metrics. Several research questions need to be addressed *(1) Metric selection: Within various metrics, how to choose the specific ones in practice.* Guidelines for the applications of different fairness and diversity measurements and a thorough theoretical and empirical investigation of the relationship between metrics are encouraged [105]. *(2) Model optimization: How to balance different objectives, especially when the objectives conflict with each other.* The research field needs to go beyond assuming a single "best" model can be obtained due to these inherent trade-offs. Instead, efforts should switch to multi-objective approaches [116, 147] that are then evaluated according to their discovered Pareto frontiers. This not only helps better benchmarking and comparison across published works but can provide a suite of non-dominated options/solutions for industry practitioners allowing an increasingly fair decision-making process. *(3) Model evaluations: How to compare and evaluate the model performances when multiple metrics are provided.* It is worth further investigating how to aggregate diverse metrics into one single metric for comparison purposes where the scales and variations of metrics might be different. Rank-based evaluations via the average of the ranks in multiple metrics could avoid the scale issue but cannot be applied in model selection. Addressing these challenges requires dedicated and extensive research efforts.

- **Intersection among Beyond-Utility Perspectives:** While each beyond-utility perspective is often investigated independently, their intersections are of significance in practice. Other intersections beyond fairness-diversity intersection are worth investigating. For example, fairness and explainability are closely connected. Giving the causality for a prediction can expose why certain recommendations, providing insights into the source of bias and potential better ways to mitigate bias [46, 48, 85]. However, researchers should be cautious when utilizing explanations for fairness where the generated explanations heavily depend on the detailed technique. It is also unclear whether improving fairness related to explainability could ensure fair outcomes.
- **Fairness and Diversity in a Dynamic Setting:** Most works in these two fields focus on static settings where the interactions and user/item profiles remain the same. However, the interactions, user/item profiles are dynamic in practice. It is unclear how the bias and diversity evolve when the distribution of users and items shift. The solutions for static setting can be applied in the latest snapshot which is effective but time-consuming. How to efficiently provide fair or diverse recommendations after the change also needs to be answered. Fair or diverse solutions will make an impact on recommendations and thus influence user's behaviors, what would be the impact in the long run. Furthermore, in these dynamic settings, it would be of interest to investigate how adversarial attacks on RS [39, 41] interact with these more responsibly/ethically developed RS.

7 CONCLUSION

In this survey, we aim to explore the connections between fairness and diversity in recommender systems. We begin the survey with introductions to preliminaries of recommender systems and relevant concepts on fairness and diversity in recommender systems. After reviewing existing works in fairness and diversity independently, we extend the diversity concept from the item level to include the user level where categorization is provided, including explicit/implicit features, historical preferences (proportionality), fairness needs, and multiple interests (general). With the expanded diversity perspective, we discuss the connections between fairness and diversity from both levels by interpreting fairness works from a diversity point of view. This novel perspective enables a better understanding of existing fairness works and reveals potential future directions. Finally, we discuss the challenges and opportunities with the hope of inspiring future innovations and highlighting the focus on beyond-utility aspects along with their intersections. We hope this survey serves as a valuable resource for future research in recommender systems, particularly in exploring the intersections of fairness and diversity.

ACKNOWLEDGMENTS

This research is supported by the National Science Foundation (NSF) under grant number IIS2239881, The Home Depot, and Snap Inc. The authors appreciate reviewers and editors for their dedication and effort throughout the review and publication process, as well as for their constructive feedback.

REFERENCES

[1] Himan Abdollahpouri and Robin Burke. 2019. Multi-stakeholder recommendation and its connection to multi-sided fairness. *arXiv* (2019).
[2] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling popularity bias in learning-to-rank recommendation. In *RecSys*. 42–46.
[3] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The unfairness of popularity bias in recommendation. *arXiv* (2019).
[4] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. 2021. User-centered evaluation of popularity bias in recommender systems. In *UMAP*. 119–129.
[5] Gediminas Adomavicius and YoungOk Kwon. 2011. Improving aggregate recommendation diversity using ranking-based techniques. *TKDE* 24, 5 (2011), 896–911.
[6] Gediminas Adomavicius and Alexander Tuzhilin. 2010. Context-aware recommender systems. In *Recommender systems handbook*. Springer, 217–253.

[7] Charu C Aggarwal. 2016. Content-based recommender systems. *Recommender systems: The textbook* (2016).
[8] Charu C Aggarwal. 2016. Evaluating recommender systems. *Recommender Systems: The Textbook* (2016), 225–254.
[9] Charu C Aggarwal. 2016. Neighborhood-based collaborative filtering. *Recommender Systems: The Textbook* (2016).
[10] Charu C Aggarwal et al. 2016. *Recommender systems*. Vol. 1. Springer.
[11] Muhammad Aljukhadar, Sylvain Senecal, and Charles-Etienne Daoust. 2012. Using recommendation agents to cope with information overload. *International Journal of Electronic Commerce* 17, 2 (2012), 41–70.
[12] Tevfik Aytekin and Mahmut Özge Karakaya. 2014. Clustering-based diversity improvement in top-N recommendation. *Journal of Intelligent Information Systems* 42, 1 (2014), 1–18.
[13] Chems Eddine Berbague, Nour El-islem Karabadji, Hassina Seridi, Panagiotis Symeonidis, Yannis Manolopoulos, and Wajdi Dhifli. 2021. An overlapping clustering approach for precision, diversity and novelty-aware recommendations. *Expert Systems with Applications* 177 (2021), 114917.
[14] 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 *KDD*. 2212–2220.
[15] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *SIGIR*. 405–414.
[16] Rodrigo Borges and Kostas Stefanidis. 2021. On mitigating popularity bias in recommendations via variational autoencoders. In *Proceedings of the 36th annual ACM symposium on applied computing*. 1383–1389.
[17] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In *Conference on fairness, accountability and transparency*. PMLR, 202–214.
[18] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*. 335–336.
[19] Pablo Castells, Neil Hurley, and Saul Vargas. 2021. Novelty and diversity in recommender systems. In *Recommender systems handbook*. Springer, 603–646.
[20] Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Controllable multi-interest framework for recommendation. In *KDD*. 2942–2951.
[21] Sung-Hyuk Cha. 2007. Comprehensive survey on distance/similarity measures between probability density functions. *City* 1, 2 (2007), 1.
[22] Haoyu Chen, Wenbin Lu, Rui Song, and Pulak Ghosh. 2021. Counterfactual Fairness through Data Preprocessing.
[23] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.
[24] Li Chen and Pearl Pu. 2007. Preference-based organization interfaces: aiding user critiques in recommender systems. In *International Conference on User Modeling*. Springer, 77–86.
[25] Lei Chen, Le Wu, Kun Zhang, Richang Hong, Defu Lian, Zhiqiang Zhang, Jun Zhou, and Meng Wang. 2023. Improving Recommendation Fairness via Data Augmentation. *arXiv* (2023).
[26] Weixin Chen, Li Chen, Yongxin Ni, Yuhan Zhao, Fajie Yuan, and Yongfeng Zhang. 2023. FMMRec: Fairness-aware Multimodal Recommendation. *arXiv* (2023).
[27] Wei Chen, Yiqing Wu, Zhao Zhang, Fuzhen Zhuang, Zhongshi He, Ruobing Xie, and Feng Xia. 2024. FairGap: Fairness-aware Recommendation via Generating Counterfactual Graph. *ACM TOIS* 42, 4 (2024), 1–25.
[28] Xiong-Hui Chen, Bowei He, Yang Yu, Qingyang Li, Zhiwei Qin, Wenjie Shang, Jieping Ye, and Chen Ma. 2023. Sim2Rec: A Simulator-based Decision-making Approach to Optimize Real-World Long-term User Engagement in Sequential Recommender Systems. *arXiv* (2023).
[29] Ziheng Chen, Fabrizio Silvestri, Jia Wang, Yongfeng Zhang, Zhenhua Huang, Hongshik Ahn, and Gabriele Tolomei. 2022. GREASE: Generate Factual and Counterfactual Explanations for GNN-based Recommendations. *arXiv* (2022).
[30] Ziheng Chen, Fabrizio Silvestri, Jia Wang, Yongfeng Zhang, and Gabriele Tolomei. 2023. The Dark Side of Explanations: Poisoning Recommender Systems with Counterfactual Examples. *arXiv* (2023).
[31] Michele Conforti, Gérard Cornuéjols, Giacomo Zambelli, et al. 2014. *Integer programming*. Vol. 271. Springer.
[32] Yue Cui, Ma Chen, Kai Zheng, Lei Chen, and Xiaofang Zhou. 2023. Controllable Universal Fair Representation Learning. In *WebConf*. 949–959.
[33] Diego Corrêa da Silva, Marcelo Garcia Manzato, and Frederico Araújo Durão. 2021. Exploiting personalized calibration and metrics for fairness recommendation. *Expert Systems with Applications* 181 (2021), 115112.
[34] Enyan Dai and Suhang Wang. 2021. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of WSDM*. 680–688.
[35] Van Dang and W Bruce Croft. 2012. Diversity by proportionality: an election-based approach to search result diversification. In *SIGIR*. 65–74.

[36] Yashar Deldjoo, Vito Walter Anelli, Hamed Zamani, Alejandro Bellogin, and Tommaso Di Noia. 2021. A flexible framework for evaluating user and item fairness in recommender systems. *UMUAI* (2021), 1–55.
[37] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2022. A survey of research on fair recommender systems. *arXiv* (2022).
[38] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2023. Fairness in recommender systems: research landscape and future directions. *UMUAI* (2023), 1–50.
[39] Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2021. A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks. *ACM Computing Surveys (CSUR)* 54, 2 (2021).
[40] Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Conference on fairness, accountability and transparency*. PMLR, 172–186.
[41] Wenqi Fan, Tyler Derr, Xiangyu Zhao, Yao Ma, Hui Liu, Jianping Wang, Jiliang Tang, and Qing Li. 2021. Attacking black-box recommendations via copying cross-domain user profiles. In *ICDE*. IEEE, 1583–1594.
[42] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph neural networks for social recommendation. In *WebConf*. 417–426.
[43] Golnoosh Farnadi, Pigi Kouki, Spencer K Thompson, Sriram Srinivasan, and Lise Getoor. 2018. A fairness-aware hybrid recommender system. *arXiv* (2018).
[44] Andres Ferraro, Xavier Serra, and Christine Bauer. 2021. Break the loop: Gender imbalance in music recommenders. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 249–254.
[45] Daniel M Fleder and Kartik Hosanagar. 2007. Recommender systems and their impact on sales diversity. In *EC*.
[46] 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. In *SIGIR*.
[47] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, et al. 2021. Towards long-term fairness in recommendation. In *Proceedings of the WSDM*. 445–453.
[48] Yingqiang Ge, Juntao Tan, Yan Zhu, Yinglong Xia, Jiebo Luo, Shuchang Liu, Zuohui Fu, Shijie Geng, Zelong Li, and Yongfeng Zhang. 2022. Explainable fairness in recommendation. In *SIGIR*. 681–691.
[49] Avijit Ghosh, Ritam Dutt, and Christo Wilson. 2021. When fair ranking meets uncertain inference. In *SIGIR*. 1033–1043.
[50] Corrado Gini. 1921. Measurement of inequality of incomes. *The economic journal* 31, 121 (1921), 124–125.
[51] Elizabeth Gómez, Carlos Shui Zhang, Ludovico Boratto, Maria Salamó, and Mirko Marras. 2021. The winner takes it all: geographic imbalance and provider (un) fairness in educational recommender systems. In *SIGIR*. 1808–1812.
[52] Sruthi Gorantla, Amit Deshpande, and Anand Louis. 2021. On the problem of underranking in group-fair ranking. In *International Conference on Machine Learning*. PMLR, 3777–3787.
[53] Przemyslaw A Grabowicz, Nicholas Perello, and Aarshee Mishra. 2022. Marrying fairness and explainability in supervised learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1905–1916.
[54] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. 2023. Adversarial learning for counterfactual fairness. *Machine Learning* 112, 3 (2023), 741–763.
[55] Qianxiu Hao, Qianqian Xu, Zhiyong Yang, and Qingming Huang. 2021. Pareto optimality for fairness-constrained collaborative filtering. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5619–5627.
[56] Jayant R Haritsa. 2009. The KNDN Problem: A Quest for Unity in Diversity. *IEEE Data Eng. Bull.* 32, 4 (2009), 15–22.
[57] Bowei He, Xu He, Yingxue Zhang, Ruiming Tang, and Chen Ma. 2023. Dynamically Expandable Graph Convolution for Streaming Recommendation. In *WebConf*. 1457–1467.
[58] Hengchang Hu, Yiming Cao, Zhankui He, Samson Tan, and Min-Yen Kan. 2023. Automatic Feature Fairness in Recommendation via Adversaries. In *SIGIR*. 245–252.
[59] Wenyue Hua, Yingqiang Ge, Shuyuan Xu, Jianchao Ji, and Yongfeng Zhang. 2023. UP5: Unbiased Foundation Model for Fairness-aware Recommendation. *arXiv* (2023).
[60] Neil J Hurley. 2013. Personalised ranking with diversity. In *Recommender Systems*. 379–382.
[61] Rashidul Islam, Kamrun Naher Keya, Ziqian Zeng, Shimei Pan, and James Foulds. 2021. Debiasing career recommendations with neural fair collaborative filtering. In *WebConf*. 3779–3790.
[62] Arjan JP Jeckmans, Michael Beye, Zekeriya Erkin, Pieter Hartel, Reginald L Lagendijk, and Qiang Tang. 2013. Privacy in recommender systems. *Social media retrieval* (2013), 263–281.
[63] Marius Kaminskas and Derek Bridge. 2016. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM TiiS* 7, 1 (2016), 1–42.
[64] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2013. Efficiency Improvement of Neutrality-Enhanced Recommendation.. In *Decisions@ RecSys*. Citeseer, 1–8.
[65] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2018. Recommendation independence. In *FAccT*.

[66] Chen Karako and Putra Manggala. 2018. Using image fairness representations in diversity-based re-ranking for recommendations. In *Adjunct Publication of UMAP*. 23–28.
[67] John Paul Kelly and Derek Bridge. 2006. Enhancing the diversity of conversational collaborative recommendations: a comparison. *Artificial Intelligence Review* 25, 1 (2006), 79–95.
[68] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv* (2013).
[69] Matevž Kunaver and Tomaž Požrl. 2017. Diversity in recommender systems–A survey. *Knowledge-based systems*.
[70] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *NeurIPS* 30 (2017).
[71] Jurek Leonhardt, Avishek Anand, and Megha Khosla. 2018. User fairness in recommender systems. In *Companion Proceedings of the The Web Conference 2018*. 101–102.
[72] Chang Li, Haoyun Feng, and Maarten de Rijke. 2020. Cascading hybrid bandits: Online learning to rank for relevance and diversity. In *RecSys*. 33–42.
[73] Chao Li, Zhiyuan Liu, Mengmeng Wu, Yuchi Xu, Huan Zhao, Pipei Huang, Guoliang Kang, Qiwei Chen, Wei Li, and Dik Lun Lee. 2019. Multi-interest network with dynamic routing for recommendation at Tmall. In *CIKM*. 2615–2623.
[74] Dong Li, Ruoming Jin, Zhenming Liu, Bin Ren, Jing Gao, and Zhi Liu. 2022. Towards Reliable Item Sampling for Recommendation Evaluation. *arXiv* (2022).
[75] Xiaohui Li and Tomohiro Murata. 2012. Multidimensional clustering based collaborative filtering approach for diversified recommendation. In *ICCSE*. IEEE, 905–910.
[76] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented fairness in recommendation. In *WebConf*. 624–632.
[77] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. 2022. Fairness in recommendation: A survey. *arXiv* (2022).
[78] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2021. Towards personalized fairness based on causal notion. In *SIGIR*. 1054–1063.
[79] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized news recommendation based on click behavior. In *IUI*. 31–40.
[80] Weiwen Liu and Robin Burke. 2018. Personalizing fairness-aware re-ranking. *arXiv* (2018).
[81] Weiwen Liu, Jun Guo, Nasim Sonboli, Robin Burke, and Shengyu Zhang. 2019. Personalized fairness-aware re-ranking for microlending. In *RecSys*. 467–471.
[82] Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, and Wenwu Zhu. 2019. Disentangled graph convolutional networks. In *International conference on machine learning*. PMLR, 4212–4221.
[83] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning disentangled representations for recommendation. *NeurIPS* 32 (2019).
[84] Sean M McNee, John Riedl, and Joseph A Konstan. 2006. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*. 1097–1101.
[85] Giacomo Medda, Francesco Fabbri, Mirko Marras, Ludovico Boratto, Mihnea Tufis, and Gianni Fenu. 2023. GNNUERS: Fairness Explanation in GNNs for Recommendation via Counterfactual Reasoning. *arXiv* (2023).
[86] Robert K Merton. 1968. The Matthew Effect in Science: The reward and communication systems of science are considered. *Science* 159, 3810 (1968), 56–63.
[87] Preetam Nandy, Cyrus Diciccio, Divya Venugopalan, Heloise Logan, Kinjal Basu, and Noureddine El Karoui. 2022. Achieving Fairness via Post-Processing in Web-Scale Recommender Systems. In *FAccT*. 715–725.
[88] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *WWW*. 677–686.
[89] Aditya Pal, Chantat Eksombatchai, Yitong Zhou, Bo Zhao, Charles Rosenberg, and Jure Leskovec. 2020. Pinnersage: Multi-modal user embedding framework for recommendations at pinterest. In *KDD*. 2311–2320.
[90] Shubham Pandey, Jiaxing Qu, Vladan Stevanović, Peter St John, and Prashun Gorai. 2021. Predicting energy and stability of known and hypothetical crystals using graph neural network. *Patterns* 2, 11 (2021), 100361.
[91] Jiaxing Qu, Yuxuan Richard Xie, and Elif Ertekin. 2023. Leveraging Language Representation for Material Recommendation, Ranking, and Exploration. *arXiv* (2023).
[92] Hossein A Rahmani, Yashar Deldjoo, Ali Tourani, and Mohammadmehdi Naghiaei. 2022. The unfairness of active users and popularity bias in point-of-interest recommendation. In *BIAS*. Springer, 56–68.
[93] Hossein A Rahmani, Mohammadmehdi Naghiaei, Mahdi Dehghan, and Mohammad Aliannejadi. 2022. Experiments on generalizability of user-oriented fairness in recommender systems. In *SIGIR*. 2755–2764.
[94] Bashir Rastegarpanah, Krishna P Gummadi, and Mark Crovella. 2019. Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *WSDM*. 231–239.
[95] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv* (2012).

[96] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*. PMLR, 1278–1286.
[97] Marco Tulio Ribeiro, Anisio Lacerda, Adriano Veloso, and Nivio Ziviani. 2012. Pareto-efficient hybridization for multi-objective recommender systems. In *RecSys*. 19–26.
[98] Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting query reformulations for web search result diversification. In *WWW*. 881–890.
[99] Guy Shani and Asela Gunawardana. 2011. Evaluating recommendation systems. In *Recommender systems handbook*. Springer, 257–297.
[100] Claude Elwood Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE* 5, 1 (2001), 3–55.
[101] Hui Shi, Yupeng Gu, Yitong Zhou, Bo Zhao, Sicun Gao, and Jishen Zhao. 2022. Every Preference Changes Differently: Neural Multi-Interest Preference Model with Temporal Dynamics for Recommendation. *arXiv* (2022).
[102] Yue Shi, Xiaoxue Zhao, Jun Wang, Martha Larson, and Alan Hanjalic. 2012. Adaptive diversification of recommendation results via latent factor portfolio. In *SIGIR*. 175–184.
[103] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *KDD*. 2219–2228.
[104] Ashudeep Singh and Thorsten Joachims. 2019. Policy learning for fairness in ranking. *NeurIPS* 32 (2019).
[105] Jessie J Smith, Lex Beattie, and Henriette Cramer. 2023. Scoping Fairness Objectives and Identifying Fairness Metrics for Recommender Systems: The Practitioners' Perspective. In *WebConf*. 3648–3659.
[106] Barry Smyth and Paul McClave. 2001. Similarity vs. diversity. In *ICCBR*. Springer, 347–361.
[107] Nasim Sonboli, Farzad Eskandanian, Robin Burke, Weiwen Liu, and Bamshad Mobasher. 2020. Opportunistic multi-aspect fairness through personalized re-ranking. In *Proceedings of the 28th ACM UMAP*. 239–247.
[108] Harald Steck. 2018. Calibrated recommendations. In *RecSys*.
[109] Jiakai Tang, Shiqi Shen, Zhipeng Wang, Zhi Gong, Jingsen Zhang, and Xu Chen. 2023. When Fairness meets Bias: a Debiased Framework for Fairness aware Top-N Recommendation. In *RecSys*. 200–210.
[110] Saúl Vargas, Linas Baltrunas, Alexandros Karatzoglou, and Pablo Castells. 2014. Coverage, redundancy and size-awareness in genre diversity for recommender systems. In *Proceedings of Recommender systems*. 209–216.
[111] Saúl Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *RecSys*. 109–116.
[112] Saul Vargas, Pablo Castells, and David Vallet. 2011. Intent-oriented diversity in recommender systems. In *SIGIR*.
[113] Mengting Wan, Jianmo Ni, Rishabh Misra, and Julian McAuley. 2020. Addressing marketing bias in product recommendations. In *Proceedings of the 13th international conference on web search and data mining*. 618–626.
[114] Wenjie Wang, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. 2022. User-controllable Recommendation Against Filter Bubbles. In *SIGIR*. 1251–1261.
[115] Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. 2020. Disentangled graph collaborative filtering. In *SIGIR*. 1001–1010.
[116] Yuhao Wang, Ha Tsz Lam, Yi Wong, Ziru Liu, Xiangyu Zhao, Yichao Wang, Bo Chen, Huifeng Guo, and Ruiming Tang. 2023. Multi-Task Deep Recommender Systems: A Survey. *arXiv* (2023).
[117] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems* 41, 3 (2023), 1–43.
[118] Yu Wang, Yuying Zhao, Yushun Dong, Huiyuan Chen, Jundong Li, and Tyler Derr. 2022. Improving fairness in graph neural networks via mitigating sensitive attribute leakage. In *KDD*. 1938–1948.
[119] Yu Wang, Yuying Zhao, Yi Zhang, and Tyler Derr. 2023. Collaboration-Aware Graph Convolutional Network for Recommender Systems. In *WebConf*. 91–101.
[120] Zihong Wang, Yingxia Shao, Jiyuan He, Jinbao Liu, Shitao Xiao, Tao Feng, and Ming Liu. 2023. Diversity-aware Deep Ranking Network for Recommendation. In *CIKM*. 2564–2573.
[121] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *JASA* 58, 301 (1963), 236–244.
[122] Jacek Wasilewski and Neil Hurley. 2016. Incorporating diversity in a learning to rank recommender system. In *The twenty-ninth international flairs conference*.
[123] Laurence A Wolsey. 2020. *Integer programming*. John Wiley & Sons.
[124] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2022. Are Big Recommendation Models Fair to Cold Users? *arXiv* (2022).
[125] Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. 2021. Fairness-aware news recommendation with decomposed adversarial learning. In *AAAI*, Vol. 35. 4462–4469.
[126] Haolun Wu, Yansen Zhang, Chen Ma, Fuyuan Lyu, Fernando Diaz, and Xue Liu. 2022. A Survey of Diversification Techniques in Search and Recommendation. *arXiv* (2022).
[127] Le Wu, Lei Chen, Pengyang Shao, Richang Hong, Xiting Wang, and Meng Wang. 2021. Learning fair representations for recommendation: A graph-based perspective. In *WebConf*. 2198–2208.

[128] Qiong Wu, Yong Liu, Chunyan Miao, Yin Zhao, Lu Guan, and Haihong Tang. 2019. Recent advances in diversified recommendation. *arXiv* (2019).
[129] Yao Wu, Jian Cao, Guandong Xu, and Yudong Tan. 2021. TFROM: A two-sided fairness-aware recommendation model for both customers and providers. In *SIGIR*. 1013–1022.
[130] Yiqing Wu, Ruobing Xie, Yongchun Zhu, Fuzhen Zhuang, Ao Xiang, Xu Zhang, Leyu Lin, and Qing He. 2022. Selective fairness in recommendation via prompts. In *SIGIR*. 2657–2662.
[131] Bin Xia, Junjie Yin, Jian Xu, and Yun Li. 2019. WE-Rec: A fairness-aware reciprocal recommendation based on Walrasian equilibrium. *Knowledge-Based Systems* 182 (2019), 104857.
[132] Lin Xiao, Zhang Min, Zhang Yongfeng, Gu Zhaoquan, Liu Yiqun, and Ma Shaoping. 2017. Fairness-aware group recommendation with pareto-efficiency. In *Proceedings of recommender systems*. 107–115.
[133] Yang Xiao, Qingqi Pei, Lina Yao, Shui Yu, Lei Bai, and Xianzhi Wang. 2020. An enhanced probabilistic fairness-aware group recommendation by incorporating social activeness. *J. Netw. Comput. Appl.* 156 (2020), 102579.
[134] Xing Xie. 2010. Potential friend recommendation in online social network. In *2010 IEEE/ACM CPSCom*. IEEE, 831–835.
[135] Chen Xu, Sirui Chen, Jun Xu, Weiran Shen, Xiao Zhang, Gang Wang, and Zhenhua Dong. 2023. P-MMF: Provider Max-min Fairness Re-ranking in Recommender System. In *WebConf*. 3701–3711.
[136] Hao Yang, Zhining Liu, Zeyu Zhang, Chenyi Zhuang, and Xu Chen. 2023. Towards Robust Fairness-aware Recommendation. In *RecSys*. 211–222.
[137] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. *NeurIPS* 30 (2017).
[138] Yuxin Ying, Fuzhen Zhuang, Yongchun Zhu, Deqing Wang, and Hongwei Zheng. 2023. CAMUS: Attribute-Aware Counterfactual Augmentation for Minority Users in Recommendation. In *WebConf*. 1396–1404.
[139] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of CIKM*. 1569–1578.
[140] Mi Zhang and Neil Hurley. 2008. Avoiding monotony: improving the diversity of recommendation lists. In *RecSys*.
[141] Mi Zhang and Neil Hurley. 2009. Novel item recommendation by user profile partitioning. In *WI-IAT*, Vol. 1. 508–515.
[142] Shengyu Zhang, Lingxiao Yang, Dong Yao, Yujie Lu, Fuli Feng, Zhou Zhao, Tat-Seng Chua, and Fei Wu. 2022. Re4: Learning to Re-contrast, Re-attend, Re-construct for Multi-interest Recommendation. In *WebConf*. 2216–2226.
[143] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)* 52, 1 (2019), 1–38.
[144] Xing Zhao, Ziwei Zhu, and James Caverlee. 2021. Rabbit holes and taste distortion: Distribution-aware recommendation with evolving interests. In *WebConf*. 888–899.
[145] Yuying Zhao, Yu Wang, and Tyler Derr. 2023. Fairness and Explainability: Bridging the Gap Towards Fair Model Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence* 9 (2023), 11363–11371.
[146] Yong Zheng, Tanaya Dave, Neha Mishra, and Harshit Kumar. 2018. Fairness in reciprocal recommendations: A speed-dating study. In *Adjunct publication of UMAP*. 29–34.
[147] Yong Zheng and David Xuejun Wang. 2022. A survey of recommender systems with multi-objective optimization. *Neurocomputing* 474 (2022), 141–153.
[148] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-aware tensor-based recommendation. In *CIKM*. 1153–1162.
[149] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *WWW*. 22–32.
[150] Eckart Zitzler, Marco Laumanns, and Lothar Thiele. 2001. SPEA2: Improving the strength Pareto evolutionary algorithm. *TIK-report* 103 (2001).
[151] Eckart Zitzler and Lothar Thiele. 1999. Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *IEEE transactions on Evolutionary Computation* 3, 4 (1999), 257–271.