A Reproducibility Study of Product-side Fairness in Bundle Recommendation

Nguyen, Huy Son; Mansoury, M.; Hanjalic, A.

DOI
10.1145/3705328.3748169

Publication date
2025

Document Version
Final published version

Published in
Proceedings of the 19th ACM Conference on Recommender Systems

A Reproducibility Study of Product-side Fairness in Bundle Recommendation

Huy-Son Nguyen*
Delft University of Technology
Delft, Netherlands
H.S.Nguyen@tudelft.nl

Yuanna Liu*
University of Amsterdam
Amsterdam, Netherlands
y.liu8@uva.nl

Masoud Mansoury
Delft University of Technology
Delft, Netherlands
m.mansoury@tudelft.nl

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

Alan Hanjalic
Delft University of Technology
Delft, Netherlands
a.hanjalic@tudelft.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, Netherlands
M.deRijke@uva.nl

Abstract

Recommender systems are known to exhibit fairness issues, particularly on the product side, where products and their associated suppliers receive unequal exposure in recommended results. While this problem has been widely studied in traditional recommendation settings, its implications for bundle recommendation (BR) remain largely unexplored. This emerging task introduces additional complexity: recommendations are generated at the bundle level, yet user satisfaction and product (or supplier) exposure depend on both the bundle and the individual items it contains. Existing fairness frameworks and metrics designed for traditional recommender systems may not directly translate to this multi-layered setting. In this paper, we conduct a comprehensive reproducibility study of product-side fairness in BR across three real-world datasets using four state-of-the-art BR methods. We analyze exposure disparities at both the bundle and item levels using multiple fairness metrics, uncovering important patterns. Our results show that exposure patterns differ notably between bundles and items, revealing the need for fairness interventions that go beyond bundle-level assumptions. We also find that fairness assessments vary considerably depending on the metric used, reinforcing the need for multi-faceted evaluation. Furthermore, user behavior plays a critical role: when users interact more frequently with bundles than with individual items, BR systems tend to yield fairer exposure distributions across both levels. Overall, our findings offer actionable insights for building fairer bundle recommender systems and establish a vital foundation for future research in this emerging domain.

CCS Concepts

• **Information systems** → **Evaluation of retrieval results**.

Keywords

Bundle recommendation, Fairness, Popularity bias, Exposure bias

*Both authors contributed equally to this research.

RecSys '25, Prague, Czech Republic

ACM ISBN 979-8-4007-1364-4/25/09
https://doi.org/10.1145/3705328.3748169

ACM Reference Format:
Huy-Son Nguyen, Yuanna Liu, Masoud Mansoury, Mohammad Aliannejadi, Alan Hanjalic, and Maarten de Rijke. 2025. A Reproducibility Study of Product-side Fairness in Bundle Recommendation. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25), September 22–26, 2025, Prague, Czech Republic.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3705328.3748169

1 Introduction

Fairness in recommender systems has become an increasingly important area of research [12, 24, 31, 40, 57]. These systems often serve multiple stakeholders – such as users, products, providers (or suppliers), and platforms – each with different expectations and needs [1]. Although recommender systems are typically optimized to deliver accurate and personalized results to users, fairness requires that products and providers are equitably represented in these outputs.

On the product side, exposure is a key concern for fairness. The visibility of a product on a recommendation list can shape users' opinions (e.g., in *news* [46, 59]), influence opportunities (e.g., in *job matching* [23, 32]), affect preferences (e.g., in *music* [42]), and drive economic outcomes (e.g., in *e-commerce* [11]). Product-side fairness aims to ensure that products receive equal exposure or exposure proportional to their utility, so that suppliers or products have a fair chance of being seen and selected by users [39, 51].

Prior research has made significant progress in measuring and addressing product-side fairness across a variety of recommendation tasks. For example, Raj and Ekstrand [48] studied fairness in top-K recommendation by modeling item exposure based on user browsing behavior. Similarly, Li et al. [30] explored fairness in next-basket recommendation, focusing on how repeated and exploratory recommendation patterns influence product exposure.

Despite these advances, the issue of fairness remains largely unexplored in the context of bundle recommendation (BR), which recommends sets of items grouped together as meaningful packages. BR is increasingly common in diverse domains such as e-commerce (e.g., *fashion* [15], *electronic kits* [53], *online games* [13]), digital media (e.g., *curated playlists* [52]), services (e.g., *meals* [29]), and even the medical field (e.g., *drug packages* [65]). Unlike single-item recommendation, bundle recommendation must account not only for the relevance of individual items but also for their compatibility as a collection, introducing more complex relationships between users, bundles, and items [45, 52, 53]. This added complexity raises critical

new challenges to fairness. Specifically, ensuring fair exposure in BR requires attention not just to which bundles are recommended, but also to how items within those bundles are presented, both of which may be influenced by biases in historical interaction data.

In this paper, we conduct the first comprehensive study of product-side fairness in bundle recommendation. Our goal is to assess how fairly current BR methods allocate exposure to both bundles and the individual items they contain, and to understand how exposure disparities arise and propagate through these systems.

We frame our investigation around the following questions:

(RQ1) To what extent does popularity bias in historical user interactions lead to unfair exposure (i.e., unequal representation) at both the bundle and item levels in BR methods?

(RQ2) How fairly do BR methods allocate exposure to bundles of varying popularity, and how does this affect the exposure of items within those bundles?

(RQ3) How does variation in users' interaction preferences, such as a tendency to engage more with bundles versus individual items, impact fairness outcomes in BR scenario?

To answer these questions, we implement a thorough empirical study using four state-of-the-art BR methods across three benchmark datasets from diverse domains: books, music playlists, and fashion outfits. We employ fairness evaluation protocols based on six widely adopted exposure fairness metrics [30, 34, 48], measuring both bundle- and item-level disparities. In addition, we devise a novel approach to investigate the impact of user preferences on bundle- and item-level fairness through a user grouping strategy. Thereby, this study offers nuanced views of how fairness manifests in BR systems.

2 Related Work

This section describes relevant studies on bundle recommendation and fairness issues, which serve as the background of our work.

2.1 Bundle Recommendation

In bundle recommendation, a variety of methods have been developed to improve the accuracy of recommending pre-defined item sets to users. An early approach of Chen et al. [9], named DAM, emphasizes the necessity of affiliated items within bundles by jointly optimizing user-item and user-bundle interactions through attention mechanisms and multi-task learning. Using advances in graph neural networks [18, 22, 27], BGCN [7] surpasses DAM [9] by unfolding user preferences into an item view and bundle view, adopting graph convolutional networks [27] to encode each type of preference separately. Deng et al. [13] employ BundleNet, performing message passing on the user-bundle-item tripartite graph.

With the growth of multi-view architectures, the adoption of contrastive learning has addressed challenges related to the inconsistency and synergy of information propagation between different views in bundle recommendation [36, 37, 45, 64]. CrossCBR [37], a multi-view learning approach with LightGCNs [22], achieves notable advancements by using cross-view contrastive losses to capture cooperative signals from two distinct views. Building on CrossCBR [37], MultiCBR [36] restructures the learning process by reversing the sequence of fusion and contrast operations. Instead of relying on a quadratic number of cross-view contrastive losses, MultiCBR [36] simplifies the framework by employing only two self-supervised contrastive losses, enhancing both its effectiveness and efficiency. Recent approaches based on distillation [49] or synergy optimization [25] have achieved state-of-the-art performance.

Investigation of intricate item connections that affect bundle tactics and user preferences has led to several in-depth studies on bundle recommendation [45, 52, 53, 58, 64]. First, Zhao et al. [64] introduces the intention disentanglement technique by modeling multiple latent features from both local and global views, using contrastive loss to align these views. This architecture faces performance limitations due to its hyperparameter sensitivity and computational complexity. BundleGT [58] leverages hierarchical graph transformer networks to capture strategy-aware representations by modeling latent associations at the bundle and user levels. To enhance bundle representation learning at the item-level, EBRec [16] augments user-item interactions by exploiting user-bundle-item correlations generated through pre-trained models. BunCa [45] harnesses item-level causation within user preferences and bundle construction, employing a refined attention operation on item-item graphs to better capture asymmetric relationships among items in real-world scenarios. Beside modeling item relations in set recommendation via graphs [44, 45, 58], some generative approaches, such as BRIDGE [6] and DisCo [5], have been further developed to produce new appropriate bundles aligning with user preferences.

These advancements listed above [45, 58, 64] uncover user preferences underlying inferred bundling strategies, but have not inspected the impact of product-side fairness on the final prediction. Our study reproduces prominent BR models [36, 37, 45, 58] within a unified experimental framework and concentrates on investigating item-level and bundle-level fairness based on their outcomes. Our scrutiny can also be extended to enhance current approaches in bundle construction [38, 53], or bundle generation [8, 54] tasks.

2.2 Fairness in Recommender Systems

Information access systems, such as search engines and recommendation systems, exhibit characteristics of multi-stakeholder, ranking-based displays, centrality of personalization, and user responses in real applications [17]. These factors pose challenges in evaluating and optimizing fairness. In recent years, the research community has shown a growing interest in fairness within recommender systems, focusing on fairness definitions [3, 62], evaluation metrics [50, 63], and optimization algorithms [4, 41] to foster a fair and healthy recommendation ecosystem.

Fairness of recommender systems can be divided into user fairness and product fairness [57]. Product fairness investigates whether products are allocated a fair distribution of exposure by being recommended based on various fairness principles, such as equal opportunity and statistical parity [48]. On the user side, user fairness measures the deviation of recommendation performance among different user groups [56]. Specifically, users can be grouped by the level of activity or the consumption of popular items [47]. To optimize product fairness and user fairness in a recommender system, fairness algorithms, including in-processing methods [61] and post-processing methods [43], have been proposed to enhance fairness at different stages of the recommendation algorithm pipeline. A recent open-source toolkit, FairDiverse [60], collects and develops multiple fairness-aware algorithms in search and recommendation.

Apart from the general recommendation scenarios, fairness is also explored in some special domains, such as next basket recommendation (NBR), where users exhibit both repetitive and exploratory purchase behaviors [28]. Liu et al. [34] re-consider item fairness metrics to assess representative NBR methods and test the robustness of these metrics. Li et al. [30] discover a shortcut to achieving better fairness and accuracy performance in NBR and proposes a fine-grained evaluation paradigm. Liu et al. [33] jointly optimize item fairness and repeat bias for NBR methods through mixed-integer linear programming. These works show that fairness behaves differently in different recommendation scenarios, which may impact the evaluation design and optimization strategies.

Fairness in bundle recommendation remains unexplored. Bundle recommendation algorithms directly recommend bundle lists to users, and indirectly influence the exposure of items within the predicted bundles. Therefore, our work aims to measure the bundle-level and item-level fairness of BR methods and investigate their potential correlation. To the best of our knowledge, we are the first to study product-side fairness in bundle recommendation.

3 Methodology

This section outlines our experimental setup used to evaluate exposure at both the bundle and item levels.

3.1 Problem Formulation

Given a set of users $\mathcal{U} = \{u_1, u_2, \ldots, u_{|\mathcal{U}|}\}$, a set of bundles $\mathcal{B} = \{b_1, b_2, \ldots, b_{|\mathcal{B}|}\}$, and a set of items $\mathcal{I} = \{i_1, i_2, \ldots, i_{|\mathcal{I}|}\}$, the BR task represents user-bundle interactions, user-item interactions, and bundle-item affiliations through three binary matrices: $X \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{B}|}$, $Y \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$, and $Z \in \{0,1\}^{|\mathcal{B}|\times|\mathcal{I}|}$ as inputs to train BR models. In these matrices, an entry of 1 indicates an observed connection between the corresponding user-bundle, user-item, or bundle-item pair, while 0 denotes the absence of such an interaction. The objective of the BR task is to predict accurately unobserved user-bundle interactions by leveraging the implicit feedback contained within X, Y, and Z matrices. Specifically, for each user $u \in \mathcal{U}$, a list of top-K bundles can be inferred and ranked through user-specific relevance probability $\hat{y}(b|u)$, where $b \in \mathcal{B}$.

Fairness approach. Our study evaluates fairness at both the item-level and bundle-level across recommended bundles for all users using fairness metrics described in Section 3.3.2. Each user receives a list of bundles L, as a recommendation, where each bundle $b \in L$ contains a set of items. We aim to evaluate whether the bundles and items obtained by these BR algorithms are fairly exposed, and to investigate the correlation between bundle-level fairness and item-level fairness in the BR scenario. Table 1 presents the notation used in this paper.

3.2 Datasets

Following [7, 16, 36, 37, 45], we use three benchmark datasets from various sectors with the same training ratio to evaluate BR models: Youshu[1], NetEase[2], and iFashion [10].

The Youshu dataset involves predicting which book lists (bundles) users are likely to purchase. In the NetEase dataset, individual tracks represent items, and user-generated playlists serve as the target bundles for music recommendation. The iFashion dataset constructs outfits as bundles, which are combinations of clothing and accessories, to attract user engagement in a fashion recommendation scenario. The key statistics of the three datasets are summarized in Table 2. These datasets cover diverse application domains and show varying interaction densities and bundle sizes, ensuring a general evaluation.

Three types of historical interaction are used as model inputs: *user-item*, *user-bundle*, and *bundle-item*. Following [37], datasets are randomly split into training, validation, and test sets on user-bundle interactions for each user by a 7 : 1 : 2 ratio.

Notably, the *c-score* metric, as reported in [49], measures the consistency between collaborative user-user pairs via interaction matrices X and Y. On the iFashion dataset, it demonstrates high consistency, indicating that users with similar item-level preferences also tend to have similar bundle-level preferences. In contrast, on the Youshu and NetEase datasets, the low *c-score* values suggest that in large-bundle datasets, there is weaker alignment between user-bundle and user-item collaborative relationships.

Table 1: Notation used in the paper.

X	binary user-bundle interaction matrix
Y	binary user-item interaction matrix
Z	binary bundle-item affiliation matrix
L	ranked list of K bundles
$L(b)$	rank of bundle b in L
$y(b\|u)$	relevance of b to u
$y(i\|i \in b)$	relevance of item i in bundle b
G^+	popular group
G^-	unpopular group
$\mathbf{a}_L$	exposure vector for bundles in L
$\mathbf{a}_L(b)$	exposure of b in L
$\mathbf{a}_L(i\|i \in b)$	exposure of item i in bundle b
$G(L)$	group alignment matrix for bundles in L
ϵ_L	the exposure of groups in L $(G(L)^T\mathbf{a}_L)$

3.3 Evaluation Protocols

3.3.1 Utility ranking metrics. Following common evaluation practices in bundle recommendation [16, 37, 45, 52], we use recall at K ($R@K$) and normalized discounted cumulative gain at K ($N@K$) to assess the utility of BR methods. $R@K$ computes the proportion of test bundles that appear within the top-K ranked prediction, while $N@K$ evaluates the ranking quality by emphasizing the placement of relevant bundles at higher positions in the predicted list.

3.3.2 Fairness ranking metrics. Following [34], we measure fairness using a popular bundle group and an unpopular bundle group over distributions of ranked bundle lists among all users. Since fair exposure is unlikely to be satisfied in a single list in the recommendation task [48], we exclude exposure fairness metrics designed for single ranking, such as attention-weighted rank fairness (AWRF) [50]. Assume $\pi(L|u)$ is a user-dependent distribution and $\rho(u)$ is a distribution over all users, then $\rho(u)\pi(L|u)$ is the distribution of overall rankings among all the users. $\epsilon_L = G(L)^T\mathbf{a}_L$ is the group exposure

[1] https://www.yousuu.com/
[2] https://music.163.com/

Table 2: Statistics of the Youshu, NetEase, and iFashion datasets.

Dataset	$\|\mathcal{U}\|$	$\|\mathcal{I}\|$	$\|\mathcal{B}\|$	$\mathcal{U}$–$\mathcal{I}$	$\mathcal{U}$–$\mathcal{B}$	#Avg.I/B	$\mathcal{U}$–$\mathcal{B}$ Dens.	c–$score$	User groups		
									$\|g_1\|$	$\|g_2\|$	$\|g_3\|$
Youshu	8,039	32,770	4,771	138,515	51,377	37.03	0.13%	0.0812	677	135	7,194
NetEase	18,528	123,628	22,864	1,128,065	302,303	77.80	0.07%	0.0599	498	240	17,790
iFashion	53,897	42,563	27,694	2,290,645	1,679,708	3.86	0.07%	0.2917	5,565	3,397	44,935

of a single ranking; its expectation $\epsilon_\pi = E_{\pi\rho}[\epsilon_L]$ is the group exposure among all the rankings. We select six representative fair ranking metrics covering two categories of fairness principles:

Equal opportunity takes into account the utility of the ranked lists, and requires that exposure should be proportional to relevance. The Exposed Utility Ratio (EUR) [51] uses a ratio-based metric to measure how much the exposure of each group deviates from the ideal of being proportional to its utility $Y(G)$:

$$\text{EUR} = \frac{\epsilon_\pi\left(G^+\right)/Y\left(G^+\right)}{\epsilon_\pi\left(G^-\right)/Y\left(G^-\right)}. \tag{1}$$

The Realized Utility Ratio (RUR) [51] further ensures that click-through rates for each group $\mathbf{\Gamma}(G)$ should be proportional to utility:

$$\text{RUR} = \frac{\mathbf{\Gamma}\left(G^+\right)/Y\left(G^+\right)}{\mathbf{\Gamma}\left(G^-\right)/Y\left(G^-\right)}. \tag{2}$$

The Expected Exposure Loss (EEL) [14] measures the Euclidean distance between expected exposure ϵ_π and target exposure ϵ^* among groups; similarly, the Expected Exposure Relevance (EER) [14] measures the dot product of expected exposure and target exposure:

$$\text{EEL} = \|\epsilon_\pi - \epsilon^*\|_2^2,\ \text{EER} = 2{\epsilon_\pi}^T\epsilon^* \tag{3}$$

Statistical parity ensures equal exposure across groups without considering utility. In particular, the Expected Exposure Disparity (EED) [14] quantifies the disparities in how overall exposure is distributed among groups:

$$\text{EED} = \|\epsilon_\pi\|_2^2. \tag{4}$$

Demographic Parity (DP) [51] calculates the ratio of average exposure obtained by the two groups:

$$\text{DP} = \epsilon_\pi\left(G^+\right)/\epsilon_\pi\left(G^-\right). \tag{5}$$

Following [48], we take logDP, logEUR, and logRUR in the experiments to deal with the empty-group case. We use the Geometric browsing model [4] to compute exposure for all fair ranking metrics, i.e., $\gamma(1-\gamma)^{L(d)-1}$, where γ is the patience parameter.

Item-level fairness evaluation. We measure item-level fairness between a popular item group and an unpopular item group. First, we obtain the exposure of bundles within a list via the browsing model. Then, each item within a bundle shares equal average exposure of the bundle exposure, i.e., item exposure $\mathbf{a}_L(i|i \in b) = \mathbf{a}_L(b)/|b|$, where $|b|$ is the number of items within bundle b. For the test data, the relevance of item i within a bundle is: $y(i|i \in b) = y(b|u)/|b|$. In this way, we can use the above fair ranking metrics to measure the item-level fairness.

3.4 Bundle Recommendation Models

To reproduce BR baselines, we investigate the following representative BR methods with open-source materials, including multi-view learning, graph neural network, and data augmentation approaches. Four well-known BR algorithms are used to facilitate our evaluations of utility and fairness.

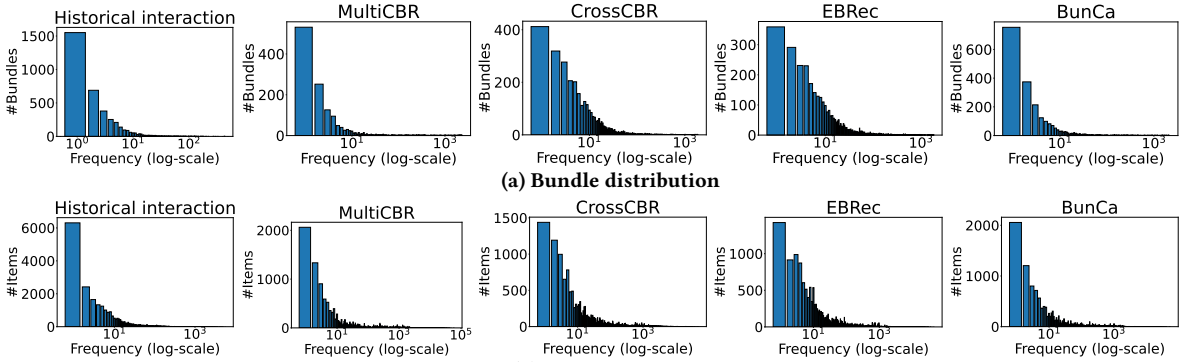
- **CrossCBR** [37] uses contrastive learning between two distinct views (*user-item view* and *user-bundle view*) to exploit their cooperation. In each view, a corresponding bipartite graph (*user-item graph* or *user-bundle graph*) is constructed and then propagated using LightGCN operation [22]. Aligning separately learned views enables each view to distill augmented data from the other, resulting in mutual enhancement. CrossCBR improves the self-discrimination of representations by increasing the dispersion among different users and bundles.
- **MultiCBR** [36] is a multi-view learning framework that jointly captures user-bundle, user-item, and bundle-item interactions. Built on CrossCBR [37], it uses bundle-item affiliations to enhance sparse bundle representations. By adopting an "*early fusion - late contrast*" method, MultiCBR models both cross-view and ego-view preferences for improved user modeling, and reduces computational cost by requiring only two contrastive losses instead of a quadratic number like CrossCBR.
- **EBRec** [16] learns item-level bundle representations through two key modules, which incorporate high-order B-U-I (bundle-user-item) correlations to capture richer collaborative signals. Moreover, it enhances B-U-I correlations by augmenting observed user-item interactions with synthetic data generated by pre-trained models, further improving representation quality.
- **BunCa** [45] is a causation-aware multi-view learning framework for BR that captures asymmetric item interconnection via bundle construction and user preferences to enhance the ultimate representation of bundles/users. Contrastive learning is inherited from prior work [36, 37] to enhance representation alignment and discrimination across views.

3.5 Implementation Details

To examine group fairness using the fair ranking metrics from Section 3.3.2, we create a popularity-based bundle and item partition, which is determined by the frequency of purchases across all historical user interactions. Following [19, 34], highly interacted bundles making up roughly 20% of interactions are categorized as the popular group (G^+), while the remaining 80% constitute the unpopular group (G^-). Particularly, the interaction frequency for each item is determined to compute item popularity by first reconstructing the user-item matrix, Y', as follows:

$$Y' = (X \times Z) + Y \tag{6}$$

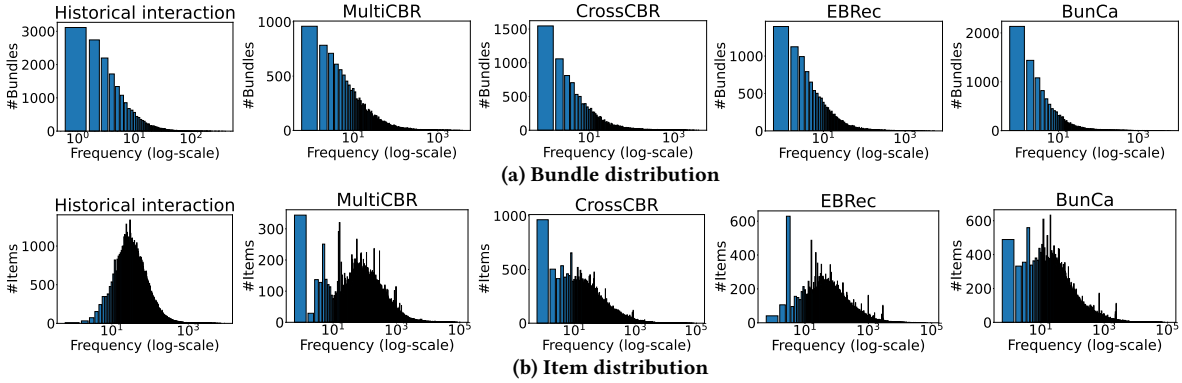
then we normalize the matrix by dividing its entries by the total number of interactions. Given Y', we follow the same procedure as [19, 34] (also described above for bundles) to compute item popularity. This way of computing item popularity takes into account the impact of users' interactions with bundles.



(a) Bundle distribution

(b) Item distribution

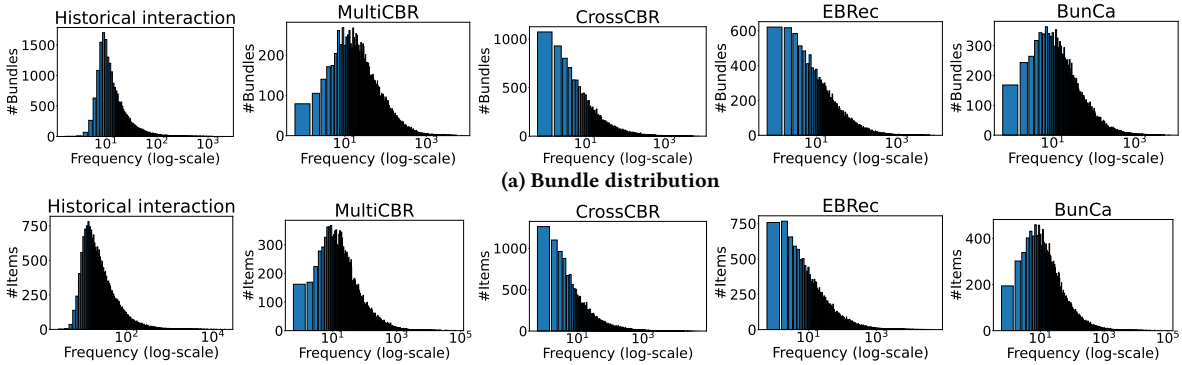
Figure 1: Distribution of bundle and item frequency in historical interaction and BR outcomes on Youshu dataset.



(a) Bundle distribution

(b) Item distribution

Figure 2: Distribution of bundle and item frequency in historical interaction and BR outcomes on NetEase dataset.



(a) Bundle distribution

(b) Item distribution

Figure 3: Distribution of bundle and item frequency in historical interaction and BR outcomes on iFashion dataset.

To ensure convincing comparisons, our implementation uses settings from state-of-the-art publications on bundle recommendation [16, 36, 37, 45]. The initial embedding dimension is set to 64, Xavier [21] is used to initialize the trainable parameters, and the Adam optimizer [26] is applied to ensure stable and efficient training. Regarding validating hyperparameters, we follow the optimal configurations provided in the original papers. To facilitate performance validation, the utility metrics, including R@K and N@K, are employed with $K = 20$. In the Geometric browsing model, γ is set to 0.5. For the training and test phases, our computational environment is built on NVIDIA P100 and T4 GPUs.Our repository is available on Github via https://github.com/Rec4Fun/Fairness_Bundle_RecSys25.

4 Experimental Results

In this section, we present the results from our experiments to answer our three research questions from the introduction.

4.1 Popularity and Exposure Analysis (RQ1)

RQ1 investigates the relationship between the distribution of user interactions in the input data and the distribution of exposure in recommendation results, analyzed at both the bundle and item levels. Fig. 1–3 visualize the distributions of bundles and items in the user interaction data and the outputs of four BR methods across three datasets. These results reveal key differences in distributional patterns (i.e., degrees of popularity bias) in user interaction with

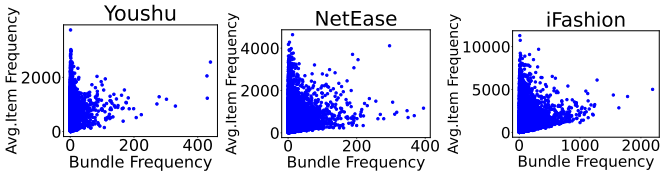
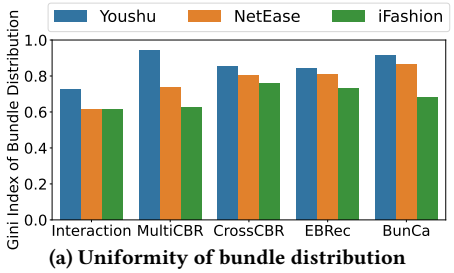
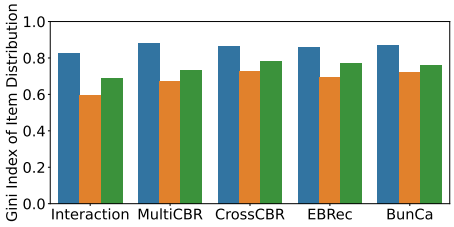


Figure 4: The relationship between each bundle popularity and average popularity of its items.



(a) Uniformity of bundle distribution



(b) Uniformity of item distribution

Figure 5: Uniformity of (a) bundle and (b) item distribution measured by Gini Index in user interaction data and the recommendation results obtained from four BR methods.

bundles and items, as well as how such biases propagate through recommendation models into their output exposure.

In the leftmost plots of each figure, corresponding to the input interaction data, it becomes evident that user interactions with bundles and with individual items do not always follow the same distributional trend. These histograms show the number of bundles/items on the y-axis and the (log-transformed) interaction frequency on the x-axis. This transformation helps to visualize distributional differences more clearly across varying frequency scales.

On the Youshu and iFashion datasets (Figs. 1 and 3, respectively), bundle and item interactions follow similar overall shapes: a long-tail distribution for the Youshu dataset and an approximately normal distribution for the iFashion dataset. However, on the NetEase dataset (Fig. 2), this alignment does not hold. Bundle interactions show a long-tail distribution, whereas item interactions are closer to a normal distribution. This discrepancy aligns with NetEase's low *c-score* (shown in Table 2), suggesting that user preferences over bundles and individual items are less consistent in this dataset.

To better understand this discrepancy, Fig. 4 presents the relationship between each bundle's interaction frequency and the average interaction frequency of the items it contains. In these plots, the x-axis represents bundle frequency (computed from matrix X in Eq. 6), and the y-axis represents the average frequency of items within each bundle (computed from matrix Y' in Eq. 6). This analysis confirms that user behavior toward bundles and items can differ significantly. For instance, a frequently interacted bundle may include items that are themselves infrequently interacted with, and vice versa–a highly popular item may appear in bundles that are rarely engaged with. This suggests that bundle-level interaction patterns do not necessarily follow item-level interaction patterns.

Beyond the input data, we also observe how BR methods influence exposure at both levels. Although BR systems recommend only bundles, we infer item-level exposure by distributing each bundle's exposure score equally among its constituent items as described in Section 3.3.2. For example, if a bundle receives an exposure score of 0.9 and contains 10 items, each item is assigned 0.09 exposure.

The results show method- and dataset-specific patterns. On the Youshu dataset, item and bundle exposure distributions across BR methods closely resemble those in the interaction data. On the NetEase dataset, bundle distributions remain consistent with interaction data, but item-level exposure varies across methods. MultiCBR and EBRec produce item distributions resembling the normal distribution seen in the input data, while CrossCBR and BunCa exhibit long-tail patterns. On the iFashion dataset, MultiCBR and BunCa match the normal distribution observed in the input data, whereas CrossCBR and EBRec tend toward long-tail exposure.

Fig. 5 quantifies distributional uniformity using Gini Index [55], computed for both bundles and items. A lower Gini Index indicates a more uniform distribution (fairer representation). While NetEase and iFashion datasets exhibit similar levels of uniformity in the input interaction data, the uniformity in recommendation results varies across BR methods. Two trends emerge: (1) BR methods typically decrease the uniformity of exposure compared to the input data—indicating an amplification of popularity bias, and (2) this reduction in uniformity is consistent with the original bias level in the input data: datasets with less uniform interaction distributions, such as the Youshu dataset, tend to result in less uniform exposure distributions in both bundles and items. These findings align with prior work [2, 35] and highlight the critical role of interaction bias in shaping fairness outcomes in BR systems.

4.2 Fairness Evaluation (RQ2)

To address RQ2, the fairness of BR methods is evaluated by fairness metrics described in Section 3.3.2. Table 3 presents the results.

In terms of accuracy, MultiCBR and BunCa consistently perform well across datasets. However, when assessing fairness at the bundle level, CrossCBR (on the Youshu and NetEase dataset) and EBRec tend to outperform the other methods, reflecting a trade-off between fairness and accuracy similar to findings in previous studies [20, 42]. At the item level, no single method consistently performs best, and the fairness metrics often disagree on which BR method is superior.

Overall, these results suggest that fairness in bundle recommendation is multi-faceted. Achieving fairness at the bundle level does not necessarily translate to fairness at the item level. This highlights the need for fairness-aware approaches that explicitly consider exposure disparities at both the bundle and item levels.

4.3 Fairness Analysis on User Tendency (RQ3)

To understand how different user behaviors affect fairness outcomes, we group users based on their interaction tendencies: group g_1 tends to purchase bundles rather than individual items; group g_2 combines neutral users who treat bundles and items equally; group g_3 prefers to buy single items. We quantify this behavior

Table 3: Overall performance of bundle recommendation methods in terms of utility and fairness metrics. Bold values indicate the best performance for each metric, while underlined values represent the second-best. Arrows denote metric direction: ↑ means higher is better, ↓ means lower is better, and ○ indicates optimal fairness when values are closer to 0.

Dataset	Method	Accuracy		Bundle-level Fairness						Item-level Fairness					
		R@20↑	*N*@20↑	logEUR○	logRUR○	EEL↓	EER↑	EED↓	logDP○	logEUR○	logRUR○	EEL↓	EER↑	EED↓	logDP○
Youshu	MultiCBR	0.286	0.168	2.667	2.249	0.476	0.486	0.660	6.024	0.565	0.083	0.049	0.826	**0.564**	4.152
	CrossCBR	0.276	0.166	**2.094**	**2.101**	**0.289**	**0.569**	**0.556**	**5.444**	**0.437**	**0.079**	0.049	**0.848**	0.585	**4.025**
	EBRec	0.281	0.168	2.153	2.187	0.308	0.560	0.565	5.504	0.486	0.093	**0.048**	0.840	0.577	4.074
	BunCa	**0.295**	**0.172**	2.390	2.131	0.385	0.524	0.606	5.743	0.523	0.084	**0.048**	0.834	0.571	4.111
Netease	MultiCBR	**0.090**	**0.050**	2.173	0.690	0.387	0.560	0.589	4.744	0.458	**0.001**	0.035	0.958	0.614	2.835
	CrossCBR	0.081	0.043	**1.908**	**0.644**	**0.288**	**0.613**	**0.544**	**4.464**	0.505	0.002	0.035	0.949	**0.605**	2.872
	EBRec	0.088	0.048	2.241	0.693	0.413	0.547	0.603	4.817	**0.443**	**0.001**	0.035	**0.961**	0.617	**2.811**
	BunCa	0.089	0.048	2.515	0.755	0.519	0.500	0.661	5.110	0.452	**0.001**	0.035	0.959	0.615	2.820
iFashion	MultiCBR	**0.150**	**0.120**	1.464	0.604	0.335	0.825	0.529	4.525	0.547	0.076	0.039	1.087	0.529	3.868
	CrossCBR	0.115	0.088	1.383	**0.458**	0.293	0.853	0.518	4.418	**0.350**	**0.041**	**0.014**	**1.146**	0.563	**3.624**
	EBRec	0.133	0.105	**1.314**	0.472	**0.259**	**0.882**	**0.510**	**4.328**	0.356	0.043	**0.014**	1.144	0.561	**3.632**
	BunCa	0.143	0.112	1.578	0.594	0.399	0.781	0.550	4.680	0.601	0.076	0.048	1.070	**0.522**	3.935

using a *tendency score* that measures the disparity ratio between bundle-based and item-based interaction for each user:

$$r_u = \frac{\sum_{b \in \mathcal{B}} X_{ub}}{\sum_{i \in \mathcal{I}} Y_{ui}} \tag{7}$$

where r_u represents the proportion of bundle-based interactions to item-based interactions for each user u. A higher value of r_u depicts a stronger tendency of user u toward bundle-oriented purchases.[3]

Fig. 6–8 present the fairness outcomes of four BR methods across these user groups and datasets. Fairness varies meaningfully across user types, confirming that fairness is not solely a property of the recommendation algorithm or dataset, but also a function of user behavior or tendency in interacting with bundles or items.

For the iFashion dataset, consistency between fairness performances across groups is achieved at both item-level and bundle-level, while it is more confusing on other datasets. This is reasonably expected by the disparity issues between learning item-level and bundle-level user preference [49], the meaning of *c-score* [49], and the influence of high-frequency items within bundles on BR models [45] demonstrated in related work. The NetEase dataset exhibits the most inconsistent patterns. The limited accuracy of models and data properties (low *c-score*, large bundle-sizes, and high item-frequency variance) likely introduce noise in fairness evaluations. Moreover, bundles in the iFashion dataset are outfits, i.e., constructed more elaborately, based on the providers' strategy. Meanwhile, bundles in the Youshu and NetEase datasets are simply defined by grouping items based user sessions.

BR methods generally achieve fairer exposure distributions when serving users who primarily interact with bundles, especially according to the EEL and EER metrics. This may be due to the relatively low overlap among bundles, which promotes greater exposure diversity. EED and logDP have opposite item-level fairness rankings of the three user groups on all datasets. This means that recommendations for g_3 have the closest overall exposure between popular and unpopular item groups, while recommendations for g_1 have the closest average exposure between two item groups.

[3]In our experiment, users with $r_u > 1.1$ are bundle-oriented (g_1), those with $r_u < 0.9$ are item-oriented (g_3), and users with r_u in between are considered neutral (g_2). These thresholds were selected based on the observed distribution of r_u, though alternative values may also be appropriate depending on dataset characteristics.

5 Conclusion

We conducted the first in-depth study of product-side fairness in the bundle recommendation task. Unlike conventional recommendation settings where individual items are recommended, BR systems deliver bundles—each consisting of multiple items. While exposure is explicitly given to bundles, items within them also receive indirect exposure. This dual-layer exposure requires fairness assessments at both the bundle-level and item-level aspects.

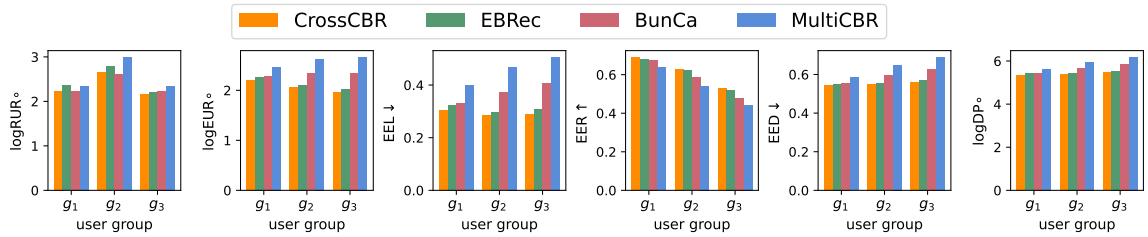
We found that the distribution of exposure in historical interaction data and in the outputs of BR methods differs notably between bundles and items. Hence, fairness interventions cannot rely solely on bundle-level assumptions and must consider item-specific dynamics. Moreover, consistent with prior research, we observed that different fairness metrics often disagree in their assessments, reinforcing the need for multi-perspective evaluation. Finally, we showed that user behavior plays a key role: when users interact more with bundles than individual items, BR methods tend to yield fairer exposure distributions at both the bundle and item levels.

These findings highlight the complex nature of fairness in bundle recommendation and the need for more targeted strategies that address its unique challenges. To the best of our knowledge, this is the first comprehensive exploration of fairness in the BR scenario.

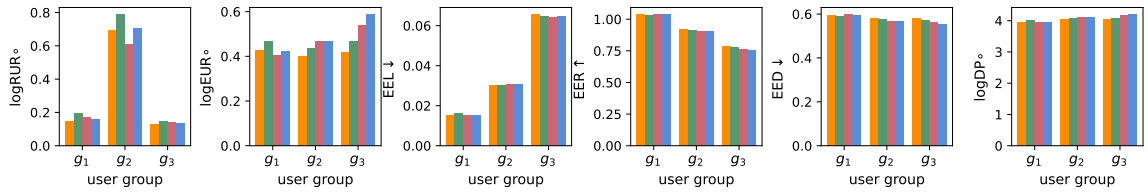
In future work, we plan to expand our analysis beyond the popularity of bundles in recommender systems. In this study, we defined fairness subgroups based on popularity. However, other grouping strategies—such as item categories, supplier identities, or sensitive user and item attributes—could reveal additional fairness issues. Investigating these dimensions should lead to the development of more holistic and inclusive fairness-aware BR methods.

Acknowledgments

This research was (partially) supported by Ahold Delhaize, through AIRLab Amsterdam, the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.1389.20.183, and KICH3.LTP.-20.006, and the European Union's Horizon Europe program under

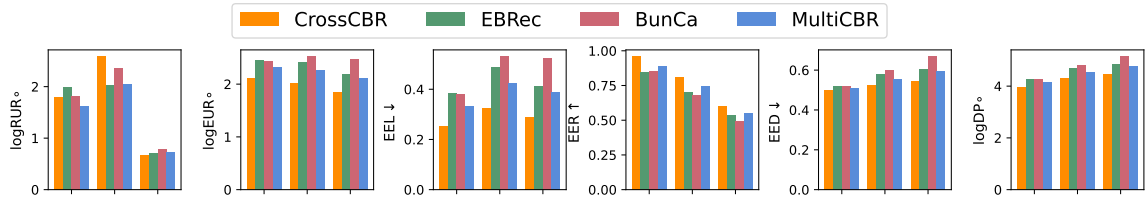


(a) Bundle-level fairness

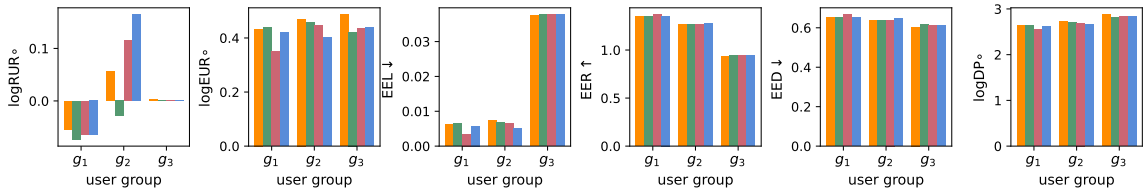


(b) Item-level fairness

Figure 6: Product fairness evaluation of BR methods for different user groups on Youshu dataset.

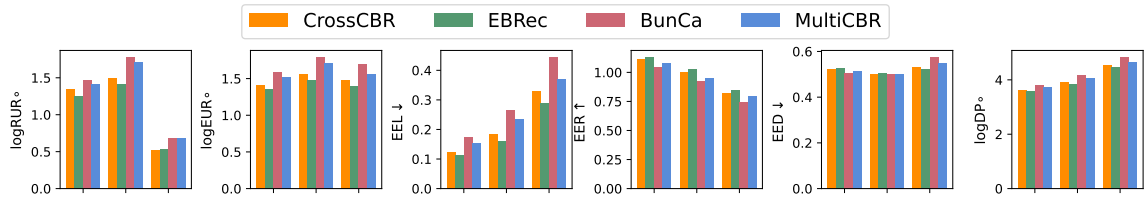


(a) Bundle-level fairness

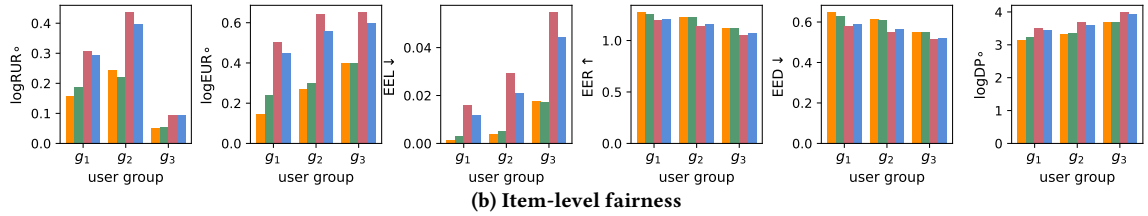


(b) Item-level fairness

Figure 7: Product fairness evaluation of BR methods for different user groups on NetEase dataset.



(a) Bundle-level fairness



(b) Item-level fairness

Figure 8: Product fairness evaluation of BR methods for different user groups on iFashion dataset.

grant agreement No 101070212, and the China Scholarship Council under grant nrs. 202206290080. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. 2020. Multi-stakeholder Recommendation: Survey and Research Directions. *User Modeling and User-Adapted Interaction* 30 (2020), 127–158.

[2] Himan Abdollahpouri and Masoud Mansoury. 2020. Multi-sided Exposure Bias in Recommendation. *ACM KDD Workshop on Industrial Recommendation Systems* (2020).

[3] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2212–2220.

[4] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *The 41st international ACM SIGIR Conference on Research & Development in Information Retrieval*. 405–414.

[5] Tuan-Nghia Bui, Huy-Son Nguyen, Cam-Van Thi Nguyen, Hoang-Quynh Le, and Duc-Trong Le. 2025. Personalized Diffusion Model Reshapes Cold-Start Bundle Recommendation. In *Companion Proceedings of the ACM on Web Conference 2025*. 3088–3091.

[6] Tuan-Nghia Bui, Huy-Son Nguyen, Cam-Van Nguyen Thi, Hoang-Quynh Le, and Duc-Trong Le. 2024. BRIDGE: Bundle Recommendation via Instruction-Driven Generation. *arXiv preprint arXiv:2412.18092* (2024).

[7] Jianxin Chang, Chen Gao, Xiangnan He, Depeng Jin, and Yong Li. 2020. Bundle Recommendation with Graph Convolutional Networks. In *Proceedings of the 43rd international ACM SIGIR Conference on Research and Development in Information Retrieval*. 1673–1676.

[8] Jianxin Chang, Chen Gao, Xiangnan He, Depeng Jin, and Yong Li. 2021. Bundle recommendation and generation with graph neural networks. *IEEE Transactions on Knowledge and Data Engineering* 35, 3 (2021), 2326–2340.

[9] Liang Chen, Yang Liu, Xiangnan He, Lianli Gao, and Zibin Zheng. 2019. Matching User with Item Set: Collaborative Bundle Recommendation with Deep Attention Network. In *IJCAI*. 2095–2101.

[10] Wen Chen, Pipei Huang, Jiaming Xu, Xin Guo, Cheng Guo, Fei Sun, Chao Li, Andreas Pfadler, Huan Zhao, and Binqiang Zhao. 2019. POG: Personalized Outfit Generation for Fashion Recommendation at Alibaba iFashion. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2662–2670.

[11] Abhisek Dash, Abhijnan Chakraborty, Saptarshi Ghosh, Animesh Mukherjee, and Krishna P Gummadi. 2022. FaiRIR: Mitigating Exposure Bias from Related Item Recommendations in Two-sided Platforms. *IEEE Transactions on Computational Social Systems* 10, 3 (2022), 1301–1313.

[12] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2024. Fairness in Recommender Systems: Research Landscape and Future Directions. *User Modeling and User-Adapted Interaction* 34, 1 (2024), 59–108.

[13] Qilin Deng, Kai Wang, Minghao Zhao, Zhene Zou, Runze Wu, Jianrong Tao, Changjie Fan, and Liang Chen. 2020. Personalized Bundle Recommendation in Online Games. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2381–2388.

[14] Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 275–284.

[15] Yujuan Ding, Zhihui Lai, PY Mok, and Tat-Seng Chua. 2023. Computational Technologies for Fashion Recommendation: A Survey. *Comput. Surveys* 56, 5 (2023), 1–45.

[16] Xiaoyu Du, Kun Qian, Yunshan Ma, and Xinguang Xiang. 2023. Enhancing Item-level Bundle Representation for Bundle Recommendation. *ACM Transactions on Recommender Systems* (2023).

[17] Michael D Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz, et al. 2022. Fairness in Information Access Systems. *Foundations and Trends in Information Retrieval* 16, 1-2 (2022), 1–177.

[18] Chen Gao, Yu Zheng, Nian Li, Yinfeng Li, Yingrong Qin, Jinghua Piao, Yuhan Quan, Jianxin Chang, Depeng Jin, Xiangnan He, et al. 2023. A Survey of Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions. *ACM Transactions on Recommender Systems* 1, 1 (2023), 1–51.

[19] 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 14th ACM international conference on web search and data mining*. 445–453.

[20] Yingqiang Ge, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. 2022. Toward Pareto Efficient Fairness-utility Trade-off in Recommendation through Reinforcement Learning. In *Proceedings of the fifteenth ACM international Conference on Web Search and Data Mining*. 316–324.

[21] Xavier Glorot and Yoshua Bengio. 2010. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 249–256.

[22] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 639–648.

[23] Rashidul Islam, Kamrun Naher Keya, Ziqian Zeng, Shimei Pan, and James Foulds. 2021. Debiasing Career Recommendations with Neural Fair Collaborative Filtering. In *Proceedings of the Web Conference 2021*. 3779–3790.

[24] Di Jin, Luzhi Wang, He Zhang, Yizhen Zheng, Weiping Ding, Feng Xia, and Shirui Pan. 2023. A Survey on Fairness-aware Recommender Systems. *Information Fusion* 100 (2023), 101906.

[25] Kyungho Kim, Sunwoo Kim, Geon Lee, and Kijung Shin. 2024. Towards Better Utilization of Multiple Views for Bundle Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 3827–3831.

[26] Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*. San Diego, CA, USA.

[27] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, 2017, Conference Track Proceedings*.

[28] Ming Li, Sami Jullien, Mozhdeh Ariannezhad, and Maarten de Rijke. 2023. A Next Basket Recommendation Reality Check. *ACM Transactions on Information Systems* 41, 4 (2023), 1–29.

[29] Ming Li, Lin Li, Xiaohui Tao, and Jimmy Xiangji Huang. 2024. MealRec+: A Meal Recommendation Dataset with Meal-Course Affiliation for Personalization and Healthiness. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 564–574.

[30] Ming Li, Yuanna Liu, Sami Jullien, Mozhdeh Ariannezhad, Andrew Yates, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Are We Really Achieving Better Beyond-Accuracy Performance in Next Basket Recommendation?. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 924–934.

[31] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. 2023. Fairness in Recommendation: Foundations, Methods, and Applications. *ACM Transactions on Intelligent Systems and Technology* 14, 5 (2023), 1–48.

[32] Yunqi Li, Michiharu Yamashita, Hanxiong Chen, Dongwon Lee, and Yongfeng Zhang. 2023. Fairness in Job Recommendation under Quantity Constraints. In *AAAI-23 Workshop on AI for Web Advertising*.

[33] Yuanna Liu, Ming Li, Mohammad Aliannejadi, and Maarten de Rijke. 2025. Repeat-Bias-Aware Optimization of Beyond-Accuracy Metrics for Next Basket Recommendation. In *European Conference on Information Retrieval*. Springer, 214–229.

[34] Yuanna Liu, Ming Li, Mozhdeh Ariannezhad, Masoud Mansoury, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Measuring item fairness in next basket recommendation: A reproducibility study. In *European Conference on Information Retrieval*. Springer, 210–225.

[35] Zhongzhou Liu, Yuan Fang, and Min Wu. 2023. Mitigating Popularity Bias for Users and Items with Fairness-centric Adaptive Recommendation. *ACM Transactions on Information Systems* 41, 3 (2023), 1–27.

[36] Yunshan Ma, Yingzhi He, Xiang Wang, Yinwei Wei, Xiaoyu Du, Yuyangzi Fu, and Tat-Seng Chua. 2024. MultiCBR: Multi-view Contrastive Learning for Bundle Recommendation. *ACM Transactions on Information Systems* 42, 4 (2024), 1–23.

[37] Yunshan Ma, Yingzhi He, An Zhang, Xiang Wang, and Tat-Seng Chua. 2022. CrossCBR: Cross-view Contrastive Learning for Bundle Recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1233–1241.

[38] Yunshan Ma, Xiaohao Liu, Yinwei Wei, Zhulin Tao, Xiang Wang, and Tat-Seng Chua. 2024. Leveraging multimodal features and item-level user feedback for bundle construction. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 510–519.

[39] Masoud Mansoury. 2022. Understanding and Mitigating Multi-sided Exposure Bias in Recommender Systems. *ACM SIGWEB Newsletter* 2022, Autumn (2022), 1–4.

[40] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 2145–2148.

[41] Masoud Mansoury, Bamshad Mobasher, and Herke van Hoof. 2024. Mitigating exposure bias in online learning to rank recommendation: A novel reward model for cascading bandits. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 1638–1648.

[42] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the Trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2243–2251.

[43] Mohammadmehdi Naghiaei, Hossein A Rahmani, and Yashar Deldjoo. 2022. CPFair: Personalized Consumer and Producer Fairness Re-ranking for Recommender Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 770–779.

[44] Huy-Son Nguyen, Tuan-Nghia Bui, Long-Hai Nguyen, Duy-Cat Can, Cam-Van Thi Nguyen, Duc-Trong Le, and Hoang-Quynh Le. 2023. HHMC: a heterogeneous x homogeneous graph-based network for multimodal cross-selling recommendation. In *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 1–6.

[45] Huy-Son Nguyen, Tuan-Nghia Bui, Long-Hai Nguyen, Hung Hoang, Cam-Van Thi Nguyen, Hoang-Quynh Le, and Duc-Trong Le. 2024. Bundle Recommendation with Item-Level Causation-Enhanced Multi-view Learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 324–341.

[46] Tao Qi, Fangzhao Wu, Chuhan Wu, Peijie Sun, Le Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. 2022. ProFairRec: Provider Fairness-aware News Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1164–1173.

[47] Hossein A Rahmani, Mohammadmehdi Naghiaei, Mahdi Dehghan, and Mohammad Aliannejadi. 2022. Experiments on Generalizability of User-oriented Fairness in Recommender Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2755–2764.

[48] Amifa Raj and Michael D. Ekstrand. 2022. Measuring Fairness in Ranked Results: An Analytical and Empirical Comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 726–736.

[49] Yuyang Ren, Zhang Haonan, Luoyi Fu, Xinbing Wang, and Chenghu Zhou. 2023. Distillation-enhanced Graph Masked Autoencoders for Bundle Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1660–1669.

[50] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attentionon fair Group Representation in Ranked Lists. In *Companion proceedings of the 2019 world wide web conference*. 553–562.

[51] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2219–2228.

[52] Meng Sun, Lin Li, Ming Li, Xiaohui Tao, Dong Zhang, Peipei Wang, and Jimmy Xiangji Huang. 2024. A Survey on Bundle Recommendation: Methods, Applications, and Challenges. *arXiv preprint arXiv:2411.00341* (2024).

[53] Zhu Sun, Kaidong Feng, Jie Yang, Hui Fang, Xinghua Qu, Yew-Soon Ong, and Wenyuan Liu. 2024. Revisiting Bundle Recommendation for Intent-aware Product Bundling. *ACM Transactions on Recommender Systems* 2, 3 (2024), 1–34.

[54] Zhu Sun, Kaidong Feng, Jie Yang, Xinghua Qu, Hui Fang, Yew-Soon Ong, and Wenyuan Liu. 2024. Adaptive In-Context Learning with Large Language Models for Bundle Generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 966–976.

[55] Saúl Vargas and Pablo Castells. 2014. Improving Sales Diversity by Recommending Users to Items. In *Proceedings of the 8th ACM Conference on Recommender systems*. 145–152.

[56] Lequn Wang and Thorsten Joachims. 2021. User Fairness, Item Fairness, and Diversity for Rankings in Two-sided Markets. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 23–41.

[57] 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.

[58] Yinwei Wei, Xiaohao Liu, Yunshan Ma, Xiang Wang, Liqiang Nie, and Tat-Seng Chua. 2023. Strategy-aware Bundle Recommender System. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1198–1207.

[59] Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. 2021. Fairness-aware News Recommendation with Decomposed Adversarial Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4462–4469.

[60] Chen Xu, Zhirui Deng, Clara Rus, Xiaopeng Ye, Yuanna Liu, Jun Xu, Zhicheng Dou, Ji-Rong Wen, and Maarten de Rijke. 2025. FairDiverse: A Comprehensive Toolkit for Fair and Diverse Information Retrieval Algorithms. *arXiv preprint arXiv:2502.11883* (2025).

[61] Chen Xu, Yuxin Li, Wenjie Wang, Liang Pang, Jun Xu, and Tat-Seng Chua. 2025. Bridging Jensen Gap for Max-Min Group Fairness Optimization in Recommendation. *arXiv preprint arXiv:2502.09319* (2025).

[62] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. 1–6.

[63] 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 the 2017 ACM on Conference on Information and Knowledge Management*. 1569–1578.

[64] Sen Zhao, Wei Wei, Ding Zou, and Xianling Mao. 2022. Multi-view Intent Disentangle Graph Networks for Bundle Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 4379–4387.

[65] Zhi Zheng, Chao Wang, Tong Xu, Dazhong Shen, Penggang Qin, Xiangyu Zhao, Baoxing Huai, Xian Wu, and Enhong Chen. 2023. Interaction-aware Drug Package Recommendation via Policy Gradient. *ACM Transactions on Information Systems* 41, 1 (2023), 1–32.