Causal Intervention for Fairness in Multibehavior Recommendation

Xi Wang ©, Wenjie Wang ©, Fuli Feng ©, Wenge Rong ©, Chuantao Yin ©, and Zhang Xiong ©

Abstract—Recommender systems usually learn user interests from various user behaviors, including clicks and postclick behaviors (e.g., like and favorite, which reflects the true interests of users). However, these behaviors inevitably exhibit popularity bias, leading to some unfairness issues: 1) for items with similar quality, more popular ones get more exposure; and 2) even worse the popular items with lower popularity might receive more exposure. Existing work on mitigating popularity bias blindly eliminates the bias and usually ignores the effect of item quality. We argue that the relationships between different user behaviors (e.g., conversion rate) actually reflect the item quality. Therefore, to handle the unfairness issues, we propose to mitigate the popularity bias by considering multiple user behaviors. In this work, we examine causal relationships behind the interaction generation procedure in multibehavior recommendation. Specifically, we find that: 1) item popularity is a confounder between the exposed items and users’ postclick interactions, leading to the exposure unfairness; and 2) some hidden confounders (e.g., the reputation of item producers) affect both item popularity and quality, resulting in the quality unfairness. To alleviate these confounding issues, we propose a causal framework to estimate the causal effect, which leverages backdoor adjustment to block the backdoor paths caused by the confounders. In the inference stage, we remove the negative effect of popularity and utilize the good effect of quality for recommendation. Experiments on two real-world datasets validate the effectiveness of our proposed framework, which enhances fairness without sacrificing recommendation accuracy on postclick behavior.

Index Terms—Causal intervention, fairness, multibehavior, popularity bias, recommender system.

Manuscript received 26 March 2023; revised 16 January 2024; accepted 12 March 2024. Date of publication 30 May 2024; date of current version 2 October 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61977003 and in part by the State Key Laboratory of Complex and Critical Software Environment under Grant CCSE-2024ZX-16. (Corresponding authors: Wenjie Wang; Fuli Feng.)

Xi Wang is with the State Key Laboratory of Complex and Critical Software Environment, Beihang University, Beijing 100191, China (e-mail: samuel.wang.xi@centralepekin.cn).

Wenjie Wang is with the National University of Singapore 119077, Singapore (e-mail: wangwenjie@u.nus.edu).

Fuli Feng is with the University of Science and Technology of China, Hefei 230026, China (e-mail: fulifeng93@gmail.com).

Wenge Rong and Zhang Xiong are with the School of Computer Science, Beihang University, Beijing 100191, China (e-mail: w.rong@buaa.edu.cn; xiongz@buaa.edu.cn).

Chuantao Yin is with Sino-French Engineer School, Beihang University, Beijing 100191, China (e-mail: chuantao.yin@buaa.edu.cn).

Digital Object Identifier 10.1109/TCSS.2024.3379903

I. INTRODUCTION

RECOMMENDER systems are becoming increasingly important for personalized information filtering on various platforms, such as E-commerce and video-streaming. To learn user preferences, a promising choice is jointly mining multiple user behaviors. Specifically, user behaviors can be divided into click and postclick behaviors (e.g., buy and like) [1]. Due to the uneven distribution of user behaviors, recommender models typically suffer from popularity bias, which hinders the model fidelity [3]. Worse still, recommender systems tend to amplify popularity bias, e.g., over-recommending popular items, causing various issues such as unfairness or filter bubbles [4]. Therefore, this work focuses on mitigating the impact of popularity bias from the perspective of multiple user behaviors.

Existing work on alleviating popularity bias in recommendation mainly falls into two categories.

1) Causal inference for debiasing inspects the causal relationships behind interaction generation procedure, and removes the confounding effect of popularity through causal intervention [3], [5] or counterfactual inference [6]. However, they focus on reducing the impact of popularity bias or using popularity to achieve better accuracy, but do not pay attention to the impact of popularity on fairness in recommendation. Besides, they ignore the relationships between different behaviors and only mitigate popularity bias on one of them, which might lead to harmful unfairness issues.

2) Multibehavior recommendation uses multitask learning to predict both click and postclick behaviors, which usually incorporates inverse propensity scoring (IPS) to reduce the impact of popularity on click behaviors [1], [7], [8]. Nevertheless, the propensity is hard to estimate properly, and thus IPS-based methods often suffer from the high variance problem [9].

In this work, we identify two kinds of unfairness caused by popularity bias in multibehavior recommendation. Fig. 1 provides an empirical evidence with a representative model ESMM [1] on the Kwai dataset [3]. 1) As shown in Fig. 1(a), more popular items will get higher ranks and more exposure opportunities, while having similar quality. 2) Items with lower quality but high popularity might receive more exposure. For example, for items that receive similar postclick behaviors (e.g.,

1For brevity, we use clicks to represent various implicit feedback, such as page-view, listen, play, and download.
Variables $X$ that contains an arrow into $X$ are considered confounders and are treated as unobserved variables. The backdoor criterion relative to small like/click ratios and usually have lower quality. How-the same number of likes), the ones with more clicks have smaller like/click ratios and usually have lower quality. However, as illustrated in Fig. 1(b), high popularity leads to more exposures on such items. To address these unfairness issues, it is critical to mitigate the popularity bias of both click and postclick behaviors.

To explore the reasons for the unfairness in multibehavior recommendation, we inspect the causal relationships by the causal graph in Fig. 2: $U$ and $I$ represent user and item, $C$ and $L$ denote the user’s click probability and postclick probability on the item, $Z$ is item popularity, $Q$ is item quality and $T$ represents some unobserved variables. First, we will introduce the definition of backdoor criterion [2] in causal intervention. In order to measure the direct effect that one variable, say $X$, has on another one, say $Y$, we must first make sure to isolate the effect from any other spurious correlations that might be present. The easiest way to do this is to make sure all noncausal paths between $X$ and $Y$ are blocked off. We can easily identify the variables we need to condition on by applying the so-called “backdoor criterion” which is defined as: Given an ordered pair of variables $(X, Y)$ in a directed acyclic graph $G$, a set of variables $Z$ satisfies the backdoor criterion relative to $X$ and $Y$ if no node in $Z$ is a descendant of $X$, and $Z$ blocks every path between $X$ and $Y$ that contains an arrow into $X$. And $Z$ is a confounder between variables $X$ and $Y$.

In the proposed causal graph, $Z \rightarrow L$ means that popularity affects user’s postbehaviors because of the user’s herd mentality (also known as conformity), and $Z \rightarrow I$ denotes the impact of popularity on item exposure: popular items are more likely to be recommended by the systems [3], [10]. As such, $Z$ is a confounder, causing the backdoor path $I \leftarrow Z \rightarrow L$ from $I$ to $L$. Consequently, for items with similar quality, the more popular ones will receive higher postclick probability $L$ due to the confounding effect of $Z$. As to the quality unfairness, we find that some unobserved confounders $T$ (e.g., reputation of item producer) affect both item popularity $Z$ and quality $Q$, resulting in another backdoor path $Q \leftarrow T \rightarrow Z \rightarrow L$ from $Q$ to $L$ (refer to Section II-B for more details about the causal graph). As a result, items with the worse quality might correlate with high popularity because of unobserved $T$, and thus they are more frequently recommended.

The key to solving the two unfairness issues lies in cutting off the two backdoor paths: $I \leftarrow Z \rightarrow L$ and $Q \leftarrow T \rightarrow Z \rightarrow L$. To block the first backdoor path, we can incorporate backdoor adjustment [2] to perform the intervention $do(I = i)$ thanks to the observed $Z$ [2], [3]. However, $T$ and $Q$ are usually unobserved variables in the second backdoor path $Q \leftarrow T \rightarrow Z \rightarrow L$, making backdoor adjustment intractable. Fortunately, the items’ postclick conversion ratio (i.e., the number of postclick behaviors divided by the click number) is able to reflect the item quality (see Fig. 3 for the evidence), and thus we use it as the approximation of $Q$ [11]. Furthermore, to cut off the second backdoor path via backdoor adjustment, we can control $Z$ without accessing the values of $T$ because $Z$ blocks all the backdoor paths from $Q$ to $L$ [2]. As such, we propose a multibehavior debiasing framework (MBD) to block the two backdoor paths by two interventions. Specifically, MBD additionally considers item quality for prediction and estimates the postclick probability $P(L|U, do(I), do(Q))$ during training. Thereafter, MBD removes the bad effect of $Z$ on $L$ and keeps the good effect of $Q$ for inference. We conduct experiments on two real-world datasets, showing that MBD can effectively alleviate the two unfairness issues and achieve superior accuracy. We release the code and data at https://anonymous.4open.science/r/MBD-DCBC/.

The main contributions of our work are as follows.
1) We study a new problem of popularity bias in multibehavior recommendation, and address the two consequent unfairness issues from a causal view.
2) We propose a multibehavior debiasing framework, which blocks the backdoor paths leading to two kinds of unfairness via two interventions.
3) To achieve the causal intervention in the experimental model, during the training process, we model the influence of popularity on the postclick recommendation
score, which contains the backdoor path impact between $Z$ and $L$. To remove the influence of popularity on item exposure and quality of recommendations by cutting off the backdoor path, we set the popularity as a fixed constant in the inference process.

4) Extensive experiments on two real-world datasets verify the effectiveness of MBD in mitigating the unfairness issues and maintaining the recommendation accuracy.

II. METHODOLOGY

In this section, we first formulate two fairness problems by considering multiple user behaviors, and then we discover the reasons for unfairness from a causal view. Lastly, we introduce the MBD framework to mitigate the unfairness problem via two interventions.

A. Fairness in Multibehavior Recommendation

Previous work either removes popularity bias (e.g., MACR [12]) or leverages popularity via causal inference to achieve better accuracy (e.g., PDA [3]). However, they only focus on the accuracy regarding one kind of user behavior, ignoring the relations between multiple user behaviors and the unfairness caused by popularity bias. Different from MACR [12] and PDA [3], we propose utilizing multiple user behaviors to measure item quality, and mitigate the unfairness issues by explicitly modeling the relations between popularity and quality. In this article, we focus on the relations between popularity and item quality in multibehavior recommendation, and then find two kinds of unfairness problems. Next, we will introduce the definition and the quantitative calculation of two kinds of fairness.

1) Approximation of Item Quality: Since not all the datasets have the item quality score, it is difficult to estimate it directly. In this work, we propose to use the item’s postclick conversion ratio to make an approximation. Previous research [11] and our experiments both demonstrate the positive correlation between the item quality and postclick conversion ratio. We conduct experiments on the popular Zhihu² dataset, where the collection number of articles strongly reflects the item quality. Fig. 3 clearly presents the positive correlation between collection number and the postclick conversion ratio. In this figure, the value of Spearman’s correlation coefficient [13], [14] is 0.8361 and the p-value is $4.4e^{-6}$, which also prove the rationality of approximating quality by the postclick conversion ratio.

2) Exposure Fairness: The exposure unfairness is that for items with same quality, more popular ones usually get more exposure. To evaluate this unfairness, we need to quantify the item’s exposure probability. To calculate the exposure for each item in the ranking list, we refer to the position bias defined in the discounted cumulative gain [15]. Formally, the exposure of item $i$ can be obtained by

where $v_j$ denotes the exposure probability of the $j$th item in the ranking list, and $P_{i,j}$ is the probability of item $i$ recommended at the position of $j$. To evaluate the exposure fairness, we argue that the items with similar quality should receive similar exposure. As such, we divide items into groups based on item quality, and then split each group into subgroups according to different popularity. Thus, we can measure the exposure difference between these subgroups to evaluate the exposure fairness. Formally, we estimate the exposure of a subgroup $G$ by averaging the item-level exposure

Thereafter, given two subgroups $G_1$ and $G_2$ with similar quality but different clicks (i.e., different popularity), the exposure fairness aims to minimize the exposure disparity $D_r$ between them

3) Quality Fairness: The quality unfairness is that for items with the lower quality, a higher popularity may lead to more exposure. Following previous work [16], we introduce a fairness criteria that relates exposure to item quality. Specifically, for the evaluation, we first estimate the average quality of groups, and then use the exposure/quality disparity between different groups to evaluate the quality unfairness. Formally

where $r_i$ is the postclick conversion ratio of item $i$, and $D_l(G_1, G_2)$ describes the unfairness between the groups $G_1$ and $G_2$. A small disparity $D_l$ implies that items receive the exposure proportional to their quality.

B. Causal View of Unfairness

The causal graph is a directed acyclic graph $G = \{V, E\}$, where $V$ denotes a set of variables, and $E$ represents the causal relations among variables [2]. To analyze the reasons for the two kinds of unfairness, we abstract a causal graph in Fig. 2(a) that describes the causal relations between item popularity, quality, and multiple user behaviors. We explain the causal graph as follows.

1) $U$ and $I$ denote users and items, respectively. We assume that users can only interact with the items exposed to them, and thus $I$ represents the exposed items.
2) $C$ is the click probability of the user clicking the exposed item.
3) $L$ is the postclick probability, indicating whether the user has a postclick behavior (e.g., favorite and like) on the exposed item.
4) $Z$ represents the item popularity, which is estimated by item’s received interactions.
5) $Q$ represents the item quality, which is a hidden variable and cannot be easily quantified. As discussed in

²https://github.com/THUIR/ZhihuRec-Dataset.
Section II-A, we propose to approximately estimate it by the conversion ratio from clicks to postclick behaviors. If the clicked item has better quality, users should have a higher probability of liking it.

6) $T$ represents some unobserved variables, such as the reputation and scale of item producers.

7) \{$U, I\} \rightarrow C$: the click probability is affected by the user and exposed items.

8) $Z \rightarrow I$: item popularity affects the exposure of items. The recommendation models are usually trained on long-tailed data. Therefore, they tend to give popular items more exposure [17].

9) \{$U, I\} \rightarrow L$: the user and exposed items determine the probability of postclick behaviors.

10) $C \rightarrow L$: only clicked items can have postclick behaviors.

11) \{$Z, Q\} \rightarrow L$: item popularity and quality affect the probability of postclick behaviors. Meanwhile, popularity influences users’ postbehaviors because of the users’ herd mentality (also known as conformity) [3].

12) $Z \leftarrow T \rightarrow Q$: $T$ affects both $Z$ and $Q$, causing the confounding effect between the two variables. For example, item producer reputation can affect the popularity of items because users tend to consume items with famous brands or well-reputed producers, making such items more likely to become popular. Besides, the quality of items is often affected by item producer reputation or the business scale of the producer.

1) **Reason for Two Unfairness Issues**: With the help of the causal graph in Fig. 2(a), we can explore the true reasons for the two kinds of unfairness. First, we find that item popularity $Z$ is a confounder between $I$ and $L$, causing the first backdoor path $I \leftarrow Z \rightarrow L$. The backdoor path results in that for the same quality items, more popular ones are more likely to be exposed to the user via $Z \rightarrow I$ and receive higher postclick probabilities by $Z \rightarrow L$ due to users’ herd mentality, which reflects the exposure unfairness [3]. Moreover, $T$ and $Z$ are confounders between $Q$ and $L$, causing the second backdoor path $Q \leftarrow T \rightarrow Z \rightarrow L$. Thus, item quality is correlated with popularity because of an unobserved confounder $T$ via $Q \leftarrow T \rightarrow Z$. Then some low-quality items might correlate with high popularity, which will bring more exposure to such popular items. This explains the quality unfairness.

To summarize, two backdoor paths cause two kinds of unfairness in multibehavior recommendation. To alleviate the unfairness, we need to cut off these two backdoor paths for removing the bad impact of such confounders.

C. **MBD Framework**

In this work, we propose a MBD framework that leverages backdoor adjustment [2] to cut off the backdoor paths and remove the bad effect caused by the confounders. Specifically, as shown in Fig. 2(b), we perform the intervention $\text{do}(I = i)$ by controlling the observed variable $Z$ to block the backdoor path $I \leftarrow Z \rightarrow L$. However, for the second backdoor path $Q \leftarrow T \rightarrow Z \rightarrow L$, $Q$ and $T$ are unobserved variables that make the backdoor adjustment infeasible. Thanks to the postclick conversion ratio that reflects the item quality (see Section II-A), we propose to use it for the approximation of item quality $Q$. Instead of controlling the unobserved $T$, we can achieve the intervention on $Q$ by controlling $Z$, which also blocks the second backdoor path. Besides, considering the critical impact of item quality in the recommendation, we estimate the postclick probability via $P(L|U, \text{do}(I), \text{do}(Q))$.

We denote the causal graph shown in Fig. 2(a) as $G$ and the intervened causal graph shown in Fig. 2(b) as $G'$, and then we will derive the following expression on $G$:

$$P(L|U, \text{do}(I), \text{do}(Q)) = P_{G'}(L|U, I, Q)$$

$$= \sum_{Z} \sum_{C} P_{G'}(L|U, I, Q, Z, C) \times P_{G'}(C|U, I, Q) P_{G'}(Z|U, I, Q)$$

(5a)

$$= \sum_{Z} \sum_{C} P_{G'}(L|U, I, Q, Z, C) P_{G'}(C|U, I) P_{G'}(Z)$$

(5b)

$$= \sum_{Z} P_{G'}(L|U, I, Q, Z, C = 1) P_{G'}(C = 1|U, I) P_{G'}(Z)$$

(5c)

$$+ \sum_{Z} P_{G'}(L|U, I, Q, Z, C = 0) \times P_{G'}(C = 0|U, I) P_{G'}(Z)$$

(5d)

$$\approx \sum_{Z} P_{G'}(L|U, I, Q, Z, C = 1) P_{G'}(C = 1|U, I) P(Z)$$

(5e)

(5f)

where $P_{G'}(\cdot)$ denotes the probability function estimated by $G'$. The derivation of (5) is as follows.

1) Equations (5a) and (5c) are due to Bayes’ theorem;
2) Equation (5b) is because of the do-calculus on $I$ and $Q$. The first backdoor path $I \leftarrow Z \rightarrow L$ is blocked by $\text{do}(I = i)$, which mitigates the exposure unfairness. The second backdoor path $Q \leftarrow T \rightarrow Z \rightarrow L$ is blocked by $\text{do}(Q = r_i)$, which mitigates the quality unfairness. Thus, $P_{G'}(Z|U, I, Q)$ can be deduced as $P_{G'}(Z)$. $P_{G'}(C|U, I, Q)$ changes into $P_{G'}(C|U, I)$ because $C$ is independent from $Q$. The user cannot know the item quality $Q$ before clicking the item;
3) Equation (5d) is because we do not consider the nonclicked data;
4) Equation (5e) is because of the same prior of $U, I, Z, Q$ on two graphs. $\{U, I, Z, Q\} \rightarrow L$ will not be changed by cutting off two paths $Z \rightarrow I$ and $T \rightarrow Q$.

Next, we need to consider how to estimate $P(L|U, \text{do}(I), \text{do}(Q))$ from the data. From the derived expression (5e), we need to firstly estimate $P(L|U, I, Q, Z, C = 1)$ and $P(C = 1|U, I)$, respectively, and then estimate $\sum_{Z} P(L|U, I, Q, Z, C = 1) P(C = 1|U, I) P(Z)$ to get the prediction score. The estimations are as follows.

1) Estimating $P(C = 1|U, I)$: This is the conditional click probability between users and items (i.e., the click-through ratio (CTR) [18]). To parameterize $P(c = 1|u, i)$, we use $f_c$. 

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
are the same scale. Next, we can estimate the click through rate by the min–max normalization [20] to make sure that they are the user–item matching functions \(f_c\) and \(f_l\), respectively. The purple arrows indicate that the item popularity \(Z\) is used in the training stage but removed from the inference stage.

can be realized by any user–item matching model (e.g., matrix factorization (MF) [19]). Hence, the probability can be calculated as

\[
p_{ctu} = P(c = 1|u, i) = \sigma(f_c(u, i))
\]

(6)

where \(\sigma\) is the sigmoid function to normalize the matching score between 0 and 1. We instantiate the CTR loss by the cross-entropy loss on the historical click data \(\mathcal{D}\), which consists of users’ sampled unclicked items as negative samples

\[
L_{ctu} = \sum_{(u,i) \in \mathcal{D}} -\hat{y}_c \log(p_{ctu}) - (1 - \hat{y}_c) \log(1 - p_{ctu})
\]

(7)

where \(\hat{y}_c \in \{0, 1\}\) is the user click feedback on items. In addition, we apply \(L_2\) regularization but omit it for brevity.

2) **Estimating** \(P(L|U, I, Q, Z, C = 1)\): This conditional probability estimates that given \(U = u, I = i\), item quality \(Q = r_i\) and item popularity \(Z = z_i\), how likely the user will have postclick behavior on the clicked item [i.e., the postclick conversion ratio (CVR)]. Same with CTR, we calculate this conditional probability by

\[
p_{ctu} = P(l = 1|u, i, r_i, z_i, c = 1) = \sigma(f_l(u, i))r_i^rz_i^c
\]

(8)

where \(\sigma\) is the sigmoid function, and \(f_l\) can be implemented by any user–item matching model. The hyper parameter \(\gamma_r\) is used to smooth the item popularity and \(\gamma_r\) is used to control the influence of the like/click ratio of item \(i\), i.e., \(r_i = (n_{i,l}/n_{i,c})\), where \(n_{i,l}\) and \(n_{i,c}\) are the number of postclick behaviors and clicks, respectively. We normalize the ratio \(r\) and the popularity \(z\) by the min–max normalization [20] to make sure that they are the same scale. Next, we can estimate the click through and conversion ratio (CTCVR) as

\[
p_{ctcvr} = P(l = 1|u, i, r_i, z_i, c = 1) * P(c = 1|u, i) = \sigma(f_l(u, i))r_i^rz_i^c \sigma(f_c(u, i)).
\]

(9)

We calculate the CTCVR loss by the cross-entropy loss

\[
L_{ctcvr} = \sum_{(u,i) \in \mathcal{D}_t} -\hat{y}_l \log(p_{ctcvr}) - (1 - \hat{y}_l) \log(1 - p_{ctcvr})
\]

(10)

where \(\hat{y}_l \in \{0, 1\}\) denotes whether the user has postclick behavior on the item, and \(\mathcal{D}_t\) collects the historical postclick data.

The overall loss \(L\) consists of two loss terms from the CTR and CTCVR predictions

\[
L = L_{ctu} + L_{ctcvr}.
\]

(11)

3) **Estimating** \(P(L|U, I, Q, Z, C = 1)\): At last, with the help of above two steps, we can estimate the interventional probability \(P(L|U, do(I), do(Q))\) as follows:

\[
P(L|U, do(I), do(Q)) = \sum_{z} P(L|U, I, Q, Z, C = 1)P(C = 1|U, I)P(Z)
\]

(12a)

\[
= \sum_{z} \sigma(f_l(u, i))r_i^rz_i^c \sigma(f_c(u, i))p(Z)
\]

(12b)

\[
= \sigma(f_c(u, i))\sigma(f_l(u, i))r_i^rz_i^c \sum_{z} z_i^c p(Z)
\]

(12c)

\[
= \sigma(f_c(u, i))\sigma(f_l(u, i))r_i^rz_i^c E(Z | r)
\]

(12d)

where \(E(Z | r)\) is the expectation of \(Z| r\) which is a constant and does not change the recommendation results. As Fig. 4 shows, we use (11) to train the matching functions \(f_c(u, i)\) and \(f_l(u, i)\), and model the influence of the popularity \(z_i\) and quality \(r_i\) on \(L\) at the training stage. Thereafter, we remove the influence of the confounder \(z_i\) to alleviate the two unfairness issues, i.e., estimating the deconfounded prediction \(s\) to rank item candidates

\[
s(u, i, r_i) = \sigma(f_c(u, i))\sigma(f_l(u, i))r_i^rz_i^c.
\]

(13)

### III. EXPERIMENTS

#### A. Experimental Settings

1) **Datasets**: We conduct experiments on two datasets.

1) Kwai: This dataset is released by Kuaishou User Interest Modeling Challenge, which contains two-week data of click/like/follow interactions between users and videos. In this work, we utilize the click and like interactions where like is considered as the postclick behavior. We take 10-core settings for clicks and 5-core settings for likes since likes are relatively sparse.

2) Tmall: This is the dataset released in IJCAI16 challenge, which is collected from Tmall, the largest business-to-consumer E-Commerce website in China. It records two types of user behaviors (i.e., click and buy) within the time period from 1 June 2015 to 30 November 2015. Buy is treated as the postclick behavior. Due to its large scale, we use 20-core settings for clicks and 10-core settings for the buys.

We summarize the statistics of datasets in Table I. For the experiments, we first sort the user–item interactions by timestamps, and then split the datasets into train, valid and test set with the ratio of 80%, 10%, and 10%.

---

3https://www.kuaishou.com/activity/uimc.

4https://tianchi.aliyun.com/dataset/dataDetail?dataId=53.
2) **Baselines:** We compare our proposed MBD framework with the following baselines.

1) Random: This method simply recommends the randomly sampled items for all users.

2) MF [19]: This is the basic matrix factorization model.

3) ESMM [1] is a multitask learning method for multi-behavior recommendation. It learns the CTR and CVR prediction models simultaneously and uses the CTR and CTCVR losses for training. In this work, we take the CTCVR score as the postclick behavior prediction score. In addition, we use MF to make the CTR prediction and multilayer perceptron (MLP) to make the CVR prediction, where the hidden dimensions are 128 and 64; and the activation function is sigmoid.

4) Multi-IPW-DR [7] uses multitask inverse propensity weighting (multi-IPW) estimator and multitask doubly robust (multi-DR) estimator to mitigate selection bias and data sparsity in multibehavior recommendation. We tune the IPW bound percentage $\tau$ in the range of $\{0.5, 0.55, 0.6, \ldots, 0.95\}$. **Multi-IPW** is the method that only uses multitask IPW estimator.

5) AdFair [21] uses an adversarial training framework to generate feature-independent user embeddings for recommendation and achieve personalized fairness for users. We use ESMM as its user–item matching model and take the popularity as the sensitive feature of items. We use ESMM as its backbone model. We tune the adversarial parameter $\lambda$ in the range of $\{0.01, 0.1, \ldots, 100\}$.

6) FER [22] proposes a conceptual and computational framework that applies fairness constraints on the item rankings. It uses the average relevance of each document (normalized between 0 and 1) with the query as the utility. In the experiments, we use the user–item relevance as the utility of the item.

7) DCR [23] proposes a deconfounding causal recommendation (DCR) framework to estimate the causal effect of content features to remove the backdoor path behind user–item matching, with a MoE architecture to speed up the inference process.

8) MACR [12], which is model-agnostic using counterfactual reasoning method for eliminating popularity bias. In this study, we use ESMM as its backbone model. We tune the hyperparameters $\lambda$ and $\beta$ in the range of $\{0.001, 0.01, \ldots, 1\}$, which control the weights of prediction loss, and tune the parameter $c$ for inference in the range of $\{0.01, 0.05, 0.1, \ldots, 1\}$.

9) PDA [3], which use causal intervention for mitigating popularity bias and use the adjusted popularity in inference to get better recommendation accuracy.

10) ESMM-RI and ESMM-RT are two varieties of ESMM by considering item quality in recommendation. ESMM-RI utilizes the well-trained ESMM but multiplies the postclick conversion ratio with CTCVR prediction score during the inference stage. This will directly leverage the information of item quality for recommendations. Different from ESMM-RI, ESMM-RT multiplies the postclick conversion ratio with the CTCVR prediction score in both training and inference stages.

3) **Hyperparameters:** For a fair comparison, all methods are tuned on validation data. We optimize all models by Adam optimizer with the batch size of 4096 and learning rate of 0.01 for all the experiments. $L_2$ regularization coefficient is searched from the range of $\{0.001, \ldots, 10^{-6}\}$. Besides, we adopt the early stopping strategy that stops training if top-50 Recall plus Hit Ratio on the validation dataset does not increase for ten epochs. For the parameter $\gamma_c$ and $\gamma_z$, we tune them in the range of $\{0.1, 0.2, \ldots, 2\}$ with step 0.1. Larger $\gamma_z$ and $\gamma_c$ mean larger impact of the popularity and quality of items.

4) **Metrics:** To measure the recommendation performance, we adopt three widely used evaluation metrics: Hit Ratio, Recall, and NDCG. All metrics are computed by the all-ranking protocol—all items that are not interacted by a user are the candidates. We take the recommendation accuracy of postclick behavior as the evaluation results of different models. Because the postclick behaviors are so sparse in datasets and the quantity of items is numerous that the difference of models’ performance on top-5 or top-10 is not significant. Therefore, we report the recommendation results of $K = 50$ and $K = 100$.

### B. Performance on Accuracy

In this section, we will study the recommendation performance on accuracy of our proposed method compared to other baselines. Table II shows top-$K$ recommendation results evaluated by reliable postclick behaviors on two datasets, and the main observations are as follows.

1) The proposed MBD achieves the best performance and outperforms all the baselines on the two datasets. Compared to ESMM-RT, the Hit@50 improvements on two datasets are 7.45% and 6.41%, and the Recall@50 improvements are 11.7% and 9.06%, respectively. Our MBD framework improves the recommendation accuracy through mitigating the two unfairness. By alleviating the exposure unfairness, we remove the interference of popularity which hinders the model to learn the true interests of users. Furthermore, by alleviating the quality unfairness, the interference of low-quality but high popularity items on the user’s choice is removed. In the end, we can better fit the user’s true interests and improve the recommendation accuracy. That verifies the effectiveness of MBD because it takes into account the item quality and removes the bad effect of popularity bias through two interventions.

2) MF only performs better than the random method. That is because MF only uses the postclick historical data for
model learning and suffers from the data sparsity issue [1], which hinders the model to learn user interests well.

3) ESMM performs better than MF, MACR, and AdFair. It uses multitask learning strategy to learn both CTR and CVR tasks simultaneously, which effectively solves the data sparsity issue and enables the model to recommend more satisfying items.

4) Multi-IPW-DR performs worse than multi-IPW because the imputation model aims to estimate the error for the missing data in the whole exposure space. However, in this work, we assume that the unclicked data in the exposure space are not available for training. Therefore, the imputation model does not work well. In addition, the performance of multi-IPW is unstable. It performs better than ESMM on Kwai but worse on Tmall. This may be attributed to the high model variance of the propensity score, which decreases the robustness of IPW methods.

5) AdFair performs worse than ESMM, which is because AdFair makes a tradeoff between recommendation accuracy and fairness through adversarial learning. Besides, the model is easy to focus on fairness at the expense of large accuracy because it blindly masks the items’ features without considering the causal relationships between different factors in the recommendation.

6) FER performs worse than ESMM-RI, ESMM-RT, and PDA on both datasets because the model constrains the click-through rate for the two groups to be proportional to their average utility without considering the quality of items and the relations of multiple behaviors.

7) DCR has better performance than ESMM-RI because it incorporates mixture of experts (MoE) method in its model design. The overall results still show the better performance of our proposed method than DCR. It might because that the DCR model only focus on click-through rate in recommendation and ignores the quality of item and the relation of multiple behaviors.

8) MACR has inferior performance than ESMM on two datasets. The reason is that it removes the natural direct effect [2] of items and users for mitigating the influence of item popularity and user conformity without considering other important factors like item quality. This may hinder the model from learning accurate user interests from biased interactions.

9) PDA shows better results than MACR, multi-IPW, and AdFair. By doing intervention on items and adjusting popularity bias during the inference stage, PDA gets better recommendation accuracy. Nevertheless, it still performs worse than ESMM-RT and MBD because of ignoring the relations between item quality and popularity.

10) Our proposed variants ESMM-RT and ESMM-RI surpass ESMM, MACR, and AdFair because these two methods take into account the item quality. ESMM-RI performs worse than ESMM-RT because it does not utilize the postclick conversion ratio for training and only multiplies it with the prediction score for inference, which may hurt some well-learned user interests during training.

C. Performance on Fairness

In this section, we conduct an analysis of recommendation results for each kind of unfairness to verify the effectiveness of our proposed method on mitigating the two unfairness issues.

1) Analysis on the Exposure Unfairness: In this section, we divide all users’ top-50 recommendations into several groups by two steps. 1) We divide the items into several groups by their postclick conversion ratios. The items in the same ratio threshold make a group. 2) We further separate each group into subgroups by items’ received clicks. Similarly, each subgroup has the same quantity of recommendation frequency to reduce the massive difference of items’ distribution between the subgroups.

We calculate the absolute difference of exposure between every two subgroups by (3) and take the average as the

### TABLE II

**RECOMMENDATION PERFORMANCE EVALUATED BY POSTCLICK BEHAVIORS ON THE TWO DATASETS**

| Model   | Kwai | Tmall |
|---------|------|-------|
|         | H@50 | H@100 | R@50 | R@100 | N@50 | N@100 | H@50 | H@100 | R@50 | R@100 | N@50 | N@100 |
| Random  | 0.0166 | 0.0296 | 0.0122 | 0.0235 | 0.0042 | 0.0078 | 0.0050 | 0.0101 | 0.0019 | 0.0027 | 0.0006 | 0.0009 |
| MF      | 0.0427 | 0.0838 | 0.0215 | 0.0462 | 0.0085 | 0.0115 | 0.0086 | 0.1078 | 0.0431 | 0.0535 | 0.0312 | 0.0340 |
| ESMM    | 0.0577 | 0.0933 | 0.0280 | 0.0504 | 0.0093 | 0.0135 | 0.1176 | 0.1463 | 0.0617 | 0.0787 | 0.0428 | 0.0448 |
| Multi-IPW | 0.0625 | 0.1155 | 0.0318 | 0.0642 | 0.0105 | 0.0166 | 0.1133 | 0.1487 | 0.0584 | 0.0801 | 0.0422 | 0.0475 |
| Multi-IPW-DR | 0.0561 | 0.1044 | 0.0281 | 0.0557 | 0.0092 | 0.0146 | 0.1079 | 0.1442 | 0.0546 | 0.0787 | 0.0389 | 0.0469 |
| AdFair  | 0.0569 | 0.0981 | 0.0289 | 0.0514 | 0.0087 | 0.0130 | 0.1144 | 0.1444 | 0.0606 | 0.0795 | 0.0433 | 0.0470 |
| FER     | 0.0624 | 0.1031 | 0.0313 | 0.0615 | 0.0101 | 0.0159 | 0.1144 | 0.1484 | 0.0616 | 0.0712 | 0.0427 | 0.0480 |
| DCR     | 0.0735 | 0.1251 | 0.0389 | 0.0644 | 0.0121 | 0.0183 | 0.1198 | 0.1495 | 0.0633 | 0.0794 | 0.0477 | 0.0480 |
| MACR    | 0.0496 | 0.0822 | 0.0252 | 0.0421 | 0.0082 | 0.0117 | 0.1143 | 0.1453 | 0.0578 | 0.0763 | 0.0439 | 0.0482 |
| PDA     | 0.0640 | 0.1168 | 0.0350 | 0.0684 | 0.0127 | 0.0178 | 0.0120 | 0.1511 | 0.0640 | 0.0828 | 0.0455 | 0.0491 |
| ESMM-RT | 0.0625 | 0.1153 | 0.0303 | 0.0634 | 0.0101 | 0.0162 | 0.1196 | 0.1495 | 0.0630 | 0.0817 | 0.0450 | 0.0479 |
| ESMM-RT | 0.0846 | 0.1281 | 0.0477 | 0.0721 | 0.0215 | 0.0319 | 0.1233 | 0.1532 | 0.0640 | 0.0839 | 0.0465 | 0.0508 |
| MBD     | 0.0909 | 0.1451 | 0.0533 | 0.0859 | 0.0153 | 0.0214 | 0.1312 | 0.1654 | 0.0698 | 0.0933 | 0.0484 | 0.0536 |

Note: H, R, and N denote Hit Ratio, Recall, and NDCG, respectively. The best results are highlighted in bold.
disparity of that group. The disparities of all groups are shown in Table III. If the disparity is closer to zero, the exposure of items is less relevant to the item’s popularity, which denotes fairer recommendations for the group. Besides, in Fig. 6, we examine the subgroup variance of the exposure in each group, which intuitively shows the exposure difference between the subgroups.

From Table III and Fig. 6, we find that the Random method achieves the lowest variance because it is not affected by item popularity in the historical data. Furthermore, MBD framework gets better performance as compared to other methods. It validates that MBD mitigates the disparity in the subgroups with different popularity and treats them equally. The IPW method exhibits high exposure variance in some groups due to its unstable propensity estimation.

To examine the recommendation results with more details, we visualize the exposure of each subgroup in the group with the largest item number, which is shown in Fig. 8.

1) The curve of the random method is the most horizontal one, which means that the item popularity has no effect on the recommendation results.

2) The curve of MBD is the smoothest one except the random method, which implies that the exposure of items hardly changes with popularity. It verifies the effectiveness of MBD in mitigating the exposure unfairness.

3) For ESMM, the exposure of subgroups significantly increases with clicks, which reflects that the items with high popularity are receiving more exposure. This is consistent with the analysis of the exposure unfairness.

4) Multi-IPW, AdFair, MACR, and PDA are all helpful to mitigate the influence of popularity bias. However, they still can be improved by explicitly considering the item quality for fairness.

5) In Fig. 8, the fairness improvements of MBD on Kwai are better than those on Tmall because the user behaviors on the E-commerce platform are less affected by popularity bias. As shown in Fig. 5, the postclick conversion ratio is lower on Tmall than Kwai, which means that the user behaviors are more demand-oriented on Tmall, i.e., users are more likely to click what they want to buy on the e-commerce platform. Therefore, the users’ interactions on Tmall are less affected by popularity bias than Kwai, and thus these debiasing methods have smaller advantages on Kwai.

2) Analysis on the Quality Unfairness: We separate all users’ top-50 recommendations into several groups by two steps. 1) We divide the items into groups by their received postclicks. The thresholds are chosen to ensure that the items in each group are relatively even. The distributions of postclicks are shown in Fig. 7. 2) We separate each group into several subgroups by items’ postclick conversion ratios, which indicates that the items in each subgroup have similar quality. In each group, we calculate the disparity between every two subgroups by (4) and take the average as the disparity of that group. The exposure disparities regarding item quality are shown in Table IV. Besides, as in Fig. 9, we examine the variance of the exposure divided by the ratio in each group.

From Table IV and Fig. 9, we observe that our proposed MBD framework achieves the lowest disparity and variance as compared to other methods, which illustrates that MBD will match the corresponding exposure based on the item quality. In addition, AdFair, MACR, and PDA perform better than ESMM, indicating that they mitigate the quality unfairness to some extent by eliminating the influence of popularity. However, they do not achieve superior performance due to ignoring the influence between the item quality and popularity.

To see more details in the recommendation results, we examine the results of each subgroup in the group with the maximal item number. As shown in Fig. 10, x-axis denotes the postclick ratio of each subgroup while y-axis is the value of exposure divided by the postclick conversion ratio [illustrated in (4)]. From Fig. 10, we can find the following.

1) The curve of MBD is the flattest one, which shows that the items are more likely to receive exposure according to their quality. This justifies the effectiveness of MBD in mitigating the quality unfairness.

2) On the curve of ESMM, the exposure/ratio value over subgroups decreases with ratio, which means that for items with similar postclick behaviors, the ones with low qualities have received more exposure to such items because of the high popularity.

3) Multi-IPW, AdFair, MACR, and PDA mitigate the influence of popularity bias because their curves are more horizontal as compared to the one of ESMM.

3) Ablation Study: In this section, we will disable two interventions separately and verify the effectiveness of each intervention. If we only conduct the intervention on $I$, the prediction probability will multiply the average popularity of the same-quality items in the inference stage, as shown in

$$P(L|U, do(I), Q) = \sum_Z P(L|U, I, Q, Z, C = 1) \times P(C = 1|U, I) P(Z|Q) = \sum_Z \sigma(f_l(u, i)) r_i^z \sigma(f_c(u, i)) p(Z|Q)$$

$$= \sigma(f_c(u, i)) p(Z|Q)$$

(14a)
### TABLE III

| Model  | Kwai                      | Tmall                      |
|--------|---------------------------|----------------------------|
|        | [0.1, 0.14)               | [0.1, 0.14)               |
|        | [0.14, 0.18)              | [0.18, 0.22)              |
|        | [0.22, 0.26)              | [0.26, 0.30)              |
|        | [0.30, 0.34)              | [0.34, 0.38)              |
| Random | 0.0013                    | 0.0016                    |
| ESMM   | 0.0185                    | 0.0205                    |
| Multi-IPW | 0.0125                    | 0.0110                    |
| AdFair | 0.0091                    | 0.0102                    |
| FER    | 0.0131                    | 0.0116                    |
| MACR   | 0.0171                    | 0.0124                    |
| PDA    | 0.0020                    | 0.0065                    |
| ESMM-RT| 0.0021                    | 0.0017                    |
| MBD    | 0.0018                    | 0.0046                    |

Note: The best results except Random are highlighted in bold. Smaller scores indicate better performance.

---

If we only intervene on $Q$, the prediction probability will multiply the popularity of the item in inference, as shown in

$$P(L|U, I, do(Q))$$

$$= \sum_Z P(L|U, I, Q, Z, C = 1)$$

$$\times P(C = 1|U, I)P(Z|I)$$

$$= \sum_Z \sigma(f_l(u, i)) \gamma_l z_l^{\gamma_l} \sigma(f_c(u, i)) p(Z|I)$$

$$= \sigma(f_c(u, i)) \sigma(f_l(u, i)) r_l^{\gamma_l} \sum_Z z_l^{\gamma_l} p(Z|I).$$

---

**IV. RELATED WORKS**

In this work, we explore how to mitigate popularity bias in multibehavior recommendation, which is relevant to four research topics of concern: popularity bias, multiple behaviors in recommendation, causal recommendation, and fairness in recommendation.
TABLE IV

| Model  | Kwai | Tmall |
|--------|------|-------|
|        | 5-6  | 7-8   | 9-10 | 11-15 | 16-17 | 18-21 | 22-25 | 26-29 | 30-40 |
| ESMM   | 0.4512 | 0.5079 | 0.6263 | 0.5250 | 0.5832 |
| Multi-IPW | 0.3860 | 0.5025 | 0.5421 | 0.3789 | 0.5524 |
| FER    | 0.4387 | 0.4846 | 0.5094 | 0.3989 | 0.4414 |
| MACR   | 0.3438 | 0.3868 | 0.5613 | 0.4269 | 0.3926 |
| PDA    | 0.4840 | 0.4845 | 0.5670 | 0.5025 | 0.4184 |
| ESMM-RT | 0.2248 | 0.2023 | 0.2442 | 0.1906 | 0.2597 |

Note: The best results are highlighted in bold.

Table V

| Model  | Ratio | Unfairness 1 | Unfairness 2 |
|--------|-------|--------------|--------------|
|        |       | (0.1,0.14) | (0.14,0.18) | (0.18,0.22) | (0.22,0.26) | (0.26,0.30) | Average |
| ESMM-RT | 0.0021 | 0.0053 | 0.0211 | 0.0193 | 0.0162 | 0.0128 |
| ESMM-doQ | 0.0024 | 0.0072 | 0.0119 | 0.0179 | 0.0147 | 0.0108 |
| ESMM-doI | 0.0019 | 0.0042 | 0.0116 | 0.0179 | 0.0142 | 0.0096 |
| MBD    | 0.0018 | 0.0046 | 0.0109 | 0.0171 | 0.0084 | 0.0085 |

A. Popularity Bias

Popularity bias has long been recognized as an important issue in recommender systems [17], [24], [25], [26], [27], [28]. The most widely used method is IPS [7], [29], [30], [31]. However, there is no guaranteed accuracy for estimating propensity scores. The other method is to eliminate the popularity bias through causal inference. PDA [3] examines the deviation caused by popularity from the perspective of causality. It finds the bias amplification caused by the influence of popularity bias on item exposure. However, it does not consider the relations between popularity and other factors, such as item quality. [32] proposes two strategies to neutralize popularity bias. The first one isolates popularity bias in one embedding direction and neutralizes the popularity direction posttraining. The other one encourages all embedding directions to be disentangled and popularity neutral. CD$^2$AN [33] disentangles item property representation and popularity representation from item property embedding and learns the comprehensive long-tail item property representation by introducing unexposed items and user behavior sequences into the item tower. CauSeR [6] is a session-based recommendation method, which considers popularity bias in two stages: data generation stage and training stage. It performs deconfounded training and causal inference to remove the biases introduced during the training and simultaneously models conformity bias. However, the above causal inference methods focus on the popularity bias on one user behavior and may cause unfairness due to blindly eliminating the popularity.

B. Multibehavior Recommendation

Recommender systems learn user interests from various user behaviors and making a better recommendation based on user multiple behaviors is also a significant issue [1], [34], [35]. Unified graph can be constructed to represent multibehavior data. In the graph, it learns behavior strength by user–item propagation layer and captures behavior semantics by...
item-item propagation layer [36]. By considering the user’s click and like behavior at the same time, these works [1], [8] alleviate the problem of data sparsity of user’s like behavior through multitask training. There is also work that incorporates the IPS method in multibehavior recommendation. Multi-IPW-DR [7] considers the user’s click, and like behavior at the same time. It proposes to take the probability of click as the tendency score to adjust the final postclick prediction result and use the multi-DR method to solve the sampling deviation. However, the IPS method often has high instability because of the high variance of the propensity score.

C. Causal Recommendation

Causal inference has been widely used in recommendation [37], [38]. PDA [3] leverages popularity by causal intervention that removes the confounding popularity bias in model training and adjusts the recommendation score with desired popularity bias in inference. However, it only focuses on one user behavior and ignores the unfairness problem caused by popularity. The deconfounding technique is used in linear models to learn real interest influenced by unobserved confounders [39] and backdoor adjustment [5] is used to alleviate bias amplification on major groups of items.

There are works incorporating counterfactual inference in recommendation. Wei et al. [12], Zou et al. [40] uses multitask learning to estimate the contribution of each cause and performs counterfactual inference to remove the effect of item popularity during testing. Counterfactual inference is also used to address the clickbait issue, which estimates the direct effect of exposure features in the prediction and removes it from recommendation scores [41].

Besides, some works learn causal embeddings to get unbiased representations for users and items. By separating user and item embeddings for interest and conformity respectively, each embedding can capture only one cause by training with cause-specific data [42]. Bias-free uniform data can be used to guide the model to learn unbiased embedding, forcing the model to discard item popularity [43]. However, obtaining such uniform data needs to randomly expose items to users, which may hurt user experience and the data is usually of a small scale which makes the learning less stable.

D. Fairness in Recommendation

With the increasing attention on the fairness in recommendation [22], [44], [45], [46], [47], there have been numerous attempts to define notions of fairness [48], [49], [50], [51], [52]. Generally speaking, they define it from two perspectives: individual fairness and group fairness. Individual fairness denotes that similar individuals (e.g., users or items) should receive similar treatments (e.g., exposure or clicks) [53], [54]. Besides, group fairness indicates that decisions should be balanced for different groups distinguished by sensitive attributes like gender or race [55], [56], [57]. User-oriented fairness is considered in recommendation by requiring the active and inactive user groups to be treated similarly [58]. To achieve the fairness in group, adversarial learning can be introduced to enforce fairness constraints on graph embeddings, which ensures that the learned representations do not correlate with certain attributes, such as age or gender [59]. FairGAN [60] proposes a generative adversarial network (GAN)-based learning algorithm mapping the exposure fairness issue to the problem of negative preferences in implicit feedback data. Counterfactual fair recommendations can also be achieved through adversary learning by generating feature-independent user embeddings for recommendation [21]. Some other methods like statistical parity-based measures [61] compute the difference in the distribution of different groups for different prefixes of the ranking. [57] proposes a reinforcement learning framework to deal with the changing group labels of items to achieve long-term fairness in recommendation. However, these fairness-related works usually make the tradeoff between ranking accuracy and fairness, and the reason possibly lies in that they blindly neglect the features of users or items.

V. CONCLUSION

In this work, we studied how to mitigate popularity bias in multibehavior recommendation. Firstly, we formulated two unfairness issues caused by popularity bias in multibehavior recommendation, and inspected their underlying reasons from a causal view. Thereafter, we proposed a MBD framework to alleviate two kinds of unfairness problems, which leverages backdoor adjustment to block two backdoor paths. We conducted extensive experiments on two real-world datasets, which justify the effectiveness of MBD on mitigating the unfairness issues and improving the accuracy.

Instead of blindly erasing the impact of popularity, this work takes the initial step to consider the effect of item quality and solve the unfairness problems caused by popularity bias in multibehavior recommendation. In the future, there are many research directions for further exploration. 1) The analysis of popularity bias on multiple behaviors with more complex relations. For example, comment, favorite and share are three kinds of behaviors, which can be incorporated to explore their relations with popularity bias and unfairness. 2) The relations between the multiple user behaviors might be useful to reduce various biases in recommendation or information retrieval, such as selection bias and position bias. 3) It is promising to incorporate the model-agnostic MBD framework into more recommender methods to alleviate the unfairness issues, especially under the multibehavior recommendation settings.

REFERENCES

[1] X. Ma et al., “Entire space multi-task model: An effective approach for estimating post-click conversion rate,” in Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2018, pp. 1137–1140.
[2] J. Pearl, Causality. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[3] Y. Zhang et al., “Causal intervention for leveraging popularity bias in recommendation,” in Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2021, pp. 11–20.
[4] H. Steck, “Calibrated recommendations,” in Proc. 12th ACM Conf. Recommender Syst., 2018, pp. 154–162.
[5] W. Wang, F. Feng, X. He, X. Wang, and T.-S. Chua, “Deconfounded recommendation for alleviating bias amplification,” in Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2021, pp. 1717–1725.
[6] P. Gupta, A. Sharma, P. Malhotra, L. Víg, and G. Shroff, “Causer: Causal session-based recommendations for handling popularity bias,” in Proc. 30th ACM Int. Conf. Inf. Knowl. Manage., 2021, pp. 3048–3052.

[7] W. Zhang et al., “Large-scale causal approaches to debiasing post-click conversion rate estimation with multi-task learning,” in Proc. Web Conf., 2020, pp. 2775–2778.

[8] H. Wen et al., “Entire space multi-task modeling via post-click behavioral decomposition for conversion rate prediction,” in Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2020, pp. 2377–2386.

[9] Y. Saito, S. Yaginuma, Y. Nishino, H. Sakata, and K. Nakata, “Unbiased recommendation learning from missing-not-at-random implicit feedback,” in Proc. Int. Conf. Web Search Data Mining, New York, NY, USA: ACM, 2020, pp. 501–509.

[10] Y. Liu, K. Ge, X. Zhang, and L. Lin, “Real-time attention based look-alike model for recommender system,” in Proc. 25th ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2019, pp. 2765–2773.

[11] M. Alnma, J. Lehmann, G. Shaked, F. Silvestri, and G. Tolomei, “Promoting positive post-click experience for in-stream yahoom zenith users,” in Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2015, pp. 1929–1938.

[12] T. Wei, F. Feng, J. Chen, Z. Wu, J. Yi, and X. He, “Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system,” in Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2021, pp. 1601–1608.

[13] M. B. Abdullah, “On a robust correlation coefficient,” J. Roy. Statist. Soc. D (The Statist.), vol. 39, no. 4, pp. 455–460, 1990.

[14] J. Hauke and T. Kossowski, “Comparison of values of pearson’s and spearman’s correlation coefficients on the same sets of data,” Quaestiones Geographiae, vol. 30, no. 2, p. 87, 2011.

[15] K. Järvelin and J. Kekäläinen, “Cumulated gain-based evaluation of ir techniques,” ACM Trans. Inf. Syst. (TOIS), vol. 20, no. 4, pp. 422–446, 2002.

[16] M. Morik, A. Singh, J. Hong, and T. Joachims, “Controlling fairness and bias in dynamic learning-to-rank,” in Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2020, pp. 429–438.

[17] F. Germano, V. Gómez, and G. Le Mens, “The low-get-richer: A surprising consequence of popularity-based rankings?” in Proc. World Wide Web Conf., 2019, pp. 2764–2770.

[18] M. Richardson, E. Domińowska, and R. Ragno, “Predicting clicks: Estimating the click-through rate for new ads,” in Proc. 16th Int. Conf. World Wide Web, 2007, pp. 521–530.

[19] Y. Koren, R. Bell, and C. Volinsky, “Matrix factorization techniques for recommender systems,” Computer, vol. 42, no. 8, pp. 30–37, Aug. 2009.

[20] L. Al Shalabi, Z. Shaaban, and B. Kasasbeh, “Data mining: A preprocessing engine,” J. Comput. Sci., vol. 2, no. 9, pp. 735–739, 2006.

[21] Y. Li, H. Chen, S. Xu, Y. Ge, and Y. Zhang, “Towards personalized recommendations based on optimal notion,” in Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2021, pp. 1054–1063.

[22] A. Singh and T. Joachims, “Fairness of exposure in rankings,” in Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2018, pp. 2219–2228.

[23] X. He et al., “Addressing confounding feature issue with causal recommendation,” ACM Trans. Inf. Syst., vol. 41, no. 3, pp. 1–23, 2023.

[24] H. Abdollahpouri, R. Burke, and B. Mobasher, “Controlling popularity bias in learning-to-rank recommendation,” in Proc. 11th ACM Conf. Recommender Syst., 2017, pp. 42–46.

[25] Y.-J. Park and A. Tuzhilin, “The long tail of recommender systems and how to leverage it,” in Proc. ACM Conf. Recommender Syst., 2008, pp. 11–18.

[26] Z. Zhu, Y. He, X. Zhao, Y. Zhang, J. Wang, and J. Caverlee, “Popularity-opportunity bias in collaborative filtering,” in Proc. 14th ACM Int. Conf. Web Search Data Mining, 2021, pp. 85–93.

[27] Y. Ge et al., “Understanding echo chambers in e-commerce recommender systems,” in Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2020, pp. 2261–2270.

[28] Z. Zhu, Y. He, Y. Zhang, C. Xiao, T. Zhong, and F. Zhou, “Exploring indirect entity relations for knowledge graph enhanced recommender system,” Expert Syst. Appl., vol. 213, 2023, Art. no. 118984.

[29] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims, “Recommendations as treatments: Debiasing learning and evaluation,” in Proc. 53rd Int. Conf. Mach. Learn., 2016, pp. 1670–1679.

[30] Z. Zhu, Y. He, Y. Zhang, and J. Caverlee, “Unbiased implicit recommendation and propensity estimation via combinatorial joint learning,” in Proc. 14th ACM Conf. Recommender Syst., 2020, pp. 551–556.
[56] I. Zliobaite, “On the relation between accuracy and fairness in binary classification,” in Proc. 2nd Workshop Fairness, Accountability, Transparency in Mach. Learn. (FATML) ICML, vol. 15, 2015.

[57] Y. Ge et al., “Towards long-term fairness in recommendation,” in Proc. 14th ACM Int. Conf. Web Search Data Mining, 2021, pp. 445–453.

[58] Y. Li, H. Chen, Z. Fu, Y. Ge, and Y. Zhang, “User-oriented fairness in recommendation,” in Proc. Web Conf., 2021, pp. 624–632.

[59] A. Bose and W. Hamilton, “Compositional fairness constraints for graph embeddings,” in Proc. Int. Conf. Mach. Learn., PMLR, 2019, pp. 715–724.

[60] J. Li, Y. Ren, and K. Deng, “Fairgan: Gans-based fairness-aware learning for recommendations with implicit feedback,” in Proc. ACM Web Conf., 2022, pp. 297–307.

[61] K. Yang and J. Stoyanovich, “Measuring fairness in ranked outputs,” in Proc. 29th Int. Conf. Scientific Statist. Database Manage., 2017, pp. 1–6.

Xi Wang received the B.Sc. degree in information and computational science and the M.Sc. degree in system engineering from Sino-French Engineer School, Beihang University, Beijing, China. His research activities mainly focused on machine learning, artificial intelligence, and smart education.

Wenge Rong received the B.Sc. degree from Nanjing University of Science and Technology, Nanjing, China, in 1996, the M.Sc. degree from Queen Mary College, University of London, London, U.K., in 2003, and the Ph.D. degree from the University of Reading, Reading, U.K., in 2010.

Currently, he is a Professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. He has many years of working experience as a senior software engineer in numerous research projects and commercial software products. His area of research covers machine learning, natural language processing, and information management.

Fuli Feng is a Professor with the University of Science and Technology of China, Hefei, China. His research interests include information retrieval, data mining, and multimedia analytics, particularly in machine learning techniques and applications such as causal inference, graph neural networks, adversarial learning, multsource learning, recommender systems, and text mining. He has about 100 publications published in top conferences and journals.

He received the Best Paper Honorable Mention in SIGIR 2021 and the Best Poster Award in WWW 2018. Moreover, he has served as the AC or SPC/PC-member for top-tier conferences and the invited reviewer for prestigious journals.

Chuantao Yin received the Ph.D. degree in computer science from École Centrale de Lyon, Lyon, France, in 2010.

Currently, he is a Professor with Sino-French Engineer School, Beihang University, Beijing, China. His research activities mainly focused on machine learning, artificial intelligence, and smart education.

Zhang Xiong is a Professor with the School of Computer Science and Engineering, Beihang University, Beijing, China, and the Director of the Advanced Computer Application Research Engineering Center, National Educational Ministry of China, Beijing, China. He has published over 250 referred papers in international journals and conference proceedings. His research interests span from smart cities, knowledge management, and information systems.

Prof. Xiong has won a National Science and Technology Progress Award.