Path-Specific Counterfactual Fairness for Recommender Systems

Yaochen Zhu
University of Virginia
uqp4qh@virginia.edu

Jing Ma
University of Virginia
jm3mr@virginia.edu

Liang Wu
LinkedIn Inc.
liawu@linkedin.com

Qi Guo
LinkedIn Inc.
qguo@linkedin.com

Liangjie Hong
LinkedIn Inc.
liahong@linkedin.com

Jundong Li
University of Virginia
jundong@virginia.edu

ABSTRACT

Recommender systems (RSs) have become an indispensable part of online platforms. With the growing concerns of algorithmic fairness, RSs are not only expected to deliver high-quality personalized content, but are also demanded not to discriminate against users based on their demographic information. However, existing RSs could capture undesirable correlations between sensitive features and observed user behaviors, leading to biased recommendations. Most fair RSs tackle this problem by completely blocking the influences of sensitive features on recommendations. But since sensitive features may also affect user interests in a fair manner (e.g., race on culture-based preferences), indiscriminately eliminating all the influences of sensitive features inevitably degenerate the recommendations quality and necessary diversities. To address this challenge, we propose a path-specific fair RS (PSF-RS) for recommendations. Specifically, we summarize all fair and unfair correlations between sensitive features and observed ratings into two latent proxy mediators, where the concept of path-specific bias (PS-Bias) is defined based on path-specific counterfactual inference. Inspired by Pearl's minimal change principle, we address the PS-Bias by minimally transforming the biased factual world into a hypothetically fair world, where a fair RS model can be learned accordingly by solving a constrained optimization problem. For the technical part, we propose a feasible implementation of PSF-RS, i.e., PSF-VAE, with weakly-supervised variational inference, which robustly infers the latent mediators such that unfairness can be mitigated while necessary recommendation diversities can be maximally preserved simultaneously. Experiments conducted on semi-simulated and real-world datasets demonstrate the effectiveness of PSF-RS.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Mathematics of computing** → **Causal networks**.

KEYWORDS

Path-Specific Fairness; Recommender System; Variational Inference

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599462

ACM Reference Format:
Yaochen Zhu, Jing Ma, Liang Wu, Qi Guo, Liangjie Hong, and Jundong Li. 2023. Path-Specific Counterfactual Fairness for Recommender Systems. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3580305.3599462

1 INTRODUCTION

As content grows exponentially on the web, recommender systems (RSs) are becoming increasingly critical in modern online service platforms [51]. RSs capture user interests based on their historical behaviors [17, 40], profiles [11, 54], and the content of items they have interacted with [48, 53], aiming to automatically deliver new items tailored to users' personalized interests. Nevertheless, the observed user behaviors may be unfairly correlated with certain sensitive user features, such as gender, race, and age, which can be unintentionally captured by the RSs and perpetuate into future recommendations [25]. Consequently, users may find the recommended items offensive, especially when people's concerns for discrimination have grown substantially over time [9, 10, 34, 37].

In recent years, considerable efforts have been devoted to promoting fairness of RSs from both academia and industry [44]. From the industry's perspective, several platforms are beginning to provide interfaces to encourage users to report potentially unfair recommendations when using the platform [12, 23]. Meanwhile, researchers are investigating new approaches to incorporate fairness-aware mechanisms into RSs (i.e., fair RSs) to avoid discrimination. Early fair RSs mainly rely on statistical parity to evaluate the fairness of recommendations. For instance, demographic parity demands the same positive rate (e.g., the probability of recommending an item) for different user groups. However, recent research demonstrates that statistical parity may not be adequate to reason with fairness, as different causal relations between sensitive features and outcomes may result in divergent conclusions [22]. For example, in the Berkeley admission dataset, the lower admission rate of female applicants is because females tend to apply for difficult departments [3], and naively increasing the acceptance of female applicants to achieve statistical parity may be unfair to male applicants. Therefore, causality-aware fairness gains more attention, where causal models are established with domain knowledge to reason with the causal influence of sensitive features on the observed outcomes and prevent it from negatively influencing future decisions [26].

Existing causality-aware fair RSs mainly seek to eliminate all causal effects of sensitive features on recommendations, e.g., by constraining the user latent variables learned from observed ratings to be independent of sensitive features via strategies such as adversarial training [43] or maximum mean discrepancy minimization

[30]. However, a dilemma for these methods is that, most of these features may also influence user interest in a fair manner. Take race as an example. Indeed, race can be associated with various negative social stereotypes, and recommendations based on these stereotypes can be offensive to users. However, race can also determine users' cultural background [42], such as accustomed tablewares, etc., and recommending chopsticks to East-Asian users is rarely considered offensive for online shopping platforms. Consequently, indiscriminately eliminating all the causal influence of race on recommendations may degenerate the cultural diversity critical for personalization. Another widely acknowledged example is from Pearl [39], which states that the education level of job applicants should not affect job recommendations based on negative stereotypes, but may indirectly influence the decision via certain job-related applicant features correlated with education level, such as skills. Therefore, a better strategy to achieve fair RS is path-specific causal analysis, where only unfair correlations between sensitive features and observed ratings are eliminated in recommendations.

However, the problem remains difficult because of the following multifaceted challenges. First, a prerequisite for most path-specific causal inference algorithms is the prior knowledge of the causal model, where factors that lead to fair or unfair correlations between sensitive features and outcomes are known and measured in advance [8, 21, 36, 46]. However, this assumption does not hold for RSs, as factors that causally determine the observed user behaviors are usually latent, which makes it difficult to judge whether or not they mediate the fair influences of sensitive features and can be generalized to other users. In addition, although recent awareness of fair RS from the industry has made it possible to collect potential unfair recommendations based on users' feedback to facilitate the identification of unfair latent mediators of sensitive features, such observations are usually extremely sparse, and it is difficult to ensure fairness for users with sparse or no known unfair items (i.e., path-specific fairness for RS suffers from cold-start issues [27]).

To address the aforementioned challenges, we propose a novel path-specific fair RS (PSF-RS) for recommendations. We first establish a causal graph to reason with the causal generation process of the biased observed ratings, assuming that the fair and unfair correlations between sensitive features and the observed ratings can be summarized into two latent proxy mediators. We then define the concept of path-specific bias (PS-Bias) based on path-specific counterfactual analysis on the causal graph, where we demonstrate that naive RSs can be unfair even if they do not explicitly use users' sensitive features for recommendations. To remedy the bias, inspired by Pearl's minimal change principle [39], we minimally transform the biased factual world into a hypothetically fair world with zero PS-Bias, where a fair RS model can be learned accordingly by solving a constrained optimization problem. We demonstrate that although existing fair RSs can also achieve zero PS-Bias, their modification of the biased factual world is not minimal, which destroys causal structures necessary for the diversities in recommendations. In contrast, PSF-RS eliminates the PS-Bias while maximally preserving the fair influences of sensitive features simultaneously. For the technical part, we propose a feasible implementation of PSF-RS, i.e., PSF-VAE, with weakly-supervised variational inference, where the latent proxy mediators of sensitive features can be inferred for all users with weak supervisions from the extremely sparse known unfair items. The contribution of this paper can be summarized as:

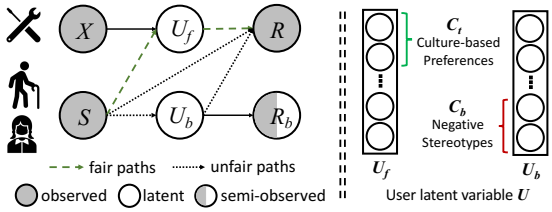


Figure 1: Causal graph that depicts the generation process of the observed ratings R and semi-observed unfair items R_b.

- To the best of our knowledge, we are the first to investigate path-specific fairness for RSs to ensure fairness while maximally preserving the necessary diversities in recommendations.
- Theoretically, a novel path-specific fair RS (PSF-RS) is proposed based on latent mediation analysis and path-specific counterfactual analysis, which minimally alters the biased factual world into a hypothetically fair world, where a fair RS can be learned accordingly by solving a constrained optimization problem.
- A feasible implementation of PSF-RS, i.e., PSF-VAE, is proposed based on weakly-supervised variational inference, where the fairness of recommendations can be generalized to users with sparse or no observed unfair item recommendations.

2 THEORETICAL ANALYSIS

2.1 Task Formulation

The focus of this paper is on fairness of recommendations with implicit feedback [18]. Consider a dataset $\mathcal{D} = \{(\mathbf{r}_i, \mathbf{s}_i, \mathbf{x}_i)\}_{i=1}^{I}$ of I users, where $\mathbf{r}_i \in \{0,1\}^J$ is a binary vector indicating whether user i has interacted with each of the J items, $\mathbf{s}_i \in \mathbb{R}^{K_s}$ denotes the sensitive user features such as race, gender, etc., and $\mathbf{x}_i \in \mathbb{R}^{K_x}$ denotes the non-sensitive user features that are not causally dependent on $\mathbf{s}_i$. Features $\mathbf{s}_i$ are sensitive in that carelessly basing recommendations on them may result in discrimination. In addition, due to the increasing awareness of fair RS from the industry, for a subset of users, we also collect certain items that each may consider unfair if these items are explicitly recommended (e.g., through self-reported unfair recommendations). We use another binary vector $\mathbf{r}_{b,i'} \in \{0,1\}^J$ to indicate the known unfair items for user i'. $\mathbf{r}_{b,i'}$ is extremely sparse and is unavailable for the majority of the users[1].

Observing the dilemma that sensitive features can both unfairly correlate with the observed ratings and causally influence user interests, the purpose of this paper is to design a path-specific fair RS that maximally eliminates the former while maximally preserving the latter, such that fairness can be achieved while necessary diversities in recommendations can be maximally preserved simultaneously.

2.2 Causal Model and Assumptions

Throughout this paper, we assume that the causal graph that generates the observed biased ratings R and the semi-observed unfair items R_b can be represented by Fig. 1, where the edges denote the direction of causal influences. The details are introduced as follows.

[1]In the remainder, the subscripts i and i' would be omitted if no ambiguity exists. The capital non-boldface symbols R, S, X, R_b are used to denote the random vectors.

2.2.1 ***User Fair Latent Variable.*** Most existing probabilistic RSs aggregate the hidden factors that causally determine the observed user behaviors R into the user latent variable U [18, 24, 28], which is usually assumed to be causally influenced by user features S and X [26]. Existing fair RSs consider all the variation of U due to S as unfair and indiscriminately eliminate them when making new recommendations. However, we postulate that for each user, we can find $U_f \in \mathbb{R}^{K_f}$ contained in U that mediates the fair influence of S on R (or has no causal relations with S). We name U_f the user fair latent variable. U_f has the property of being **resolving**[2] for S in that any influence of S on R **mediated** by U_f should be preserved to facilitate necessary diversities in recommendations. For example, sensitive feature *race* can determine a user's *cultural preference* C_t (could be several dimensions of U), which is a crucial factor that determines users' personalized interest. Therefore, C_t should be subsumed in U_f such that the causal influence of S on R mediated by C_t, which can be denoted by a **causal path** $S \rightarrow C_t \rightarrow R$, is allowed to be captured by RSs to promote culture-tailored recommendations.

2.2.2 ***User Bias Latent Variable: The Proxy Mediator.*** In addition, we use the user bias latent variable $U_b \in \mathbb{R}^{K_b}$ to summarize the remaining variations of U due to S, which captures the unfair correlations between sensitive features S and the observed ratings R in the collected data. The unfair influence of S mainly lies in two-fold. From the users' perspective, sensitive features S can determine some social stereotypes C_b (which could be some other dimensions of U) associated with certain demographic groups. Although some users may behave just according to the stereotypes (which leads to another causal path from S to R, i.e., $S \rightarrow C_b \rightarrow R$), we should not generalize them to other users with the same sensitive features. In addition, the unfair influence of S can also be attributed to the previous RS, where items unfairly associated with certain demographic groups may be overly exposed to these users that bias their behaviors [29]. Formally, the assumption that describes the unfair correlations between S and R can be summarized as follows:

Assumption 1. *The unfair correlations between S and R are composed of (1) the direct effect of S on R; (2) all indirect mediated effects of S on R* ***not resolved*** *by U_f, where the latter is assumed to be able to be summarized by a one-step* ***latent proxy mediator*** *$U_b \in \mathbb{R}^{K_b}$.*

The above assumption of unfair correlations between S and R is based on the *skeptical view* of Kilbertus et al. [21], which states that all potential influences of sensitive features on outcomes should be assumed as discriminatory unless they can be justified by a resolving mediator, which is the user fair latent variable U_f in our case. We summarize all indirect unfair influences of S into a user bias latent variable U_b because it is intractable to enumerate and measure all unfair mediators of sensitive features (e.g., all discriminatory stereotypes). One sufficient condition that allows such a substitution is that U_b **blocks** every mediated unfair path between S and R while **unblocking** every fair path resolved by U_f. This could be the case where all unfair mediators of S causally determine U_b and through which influence R, which is a common assumption in latent mediation analysis [1, 7]. Since our primary task is to analyze the fair and unfair influences of sensitive features S on the observed ratings R, other exogenous variables that causally determine U_f and U_b are omitted and summarized into their uncertainties.

[2]For readers without much background knowledge in causal inference, we provide simple and intuitive definitions for the terms highlighted in **bold** in Appendix A.

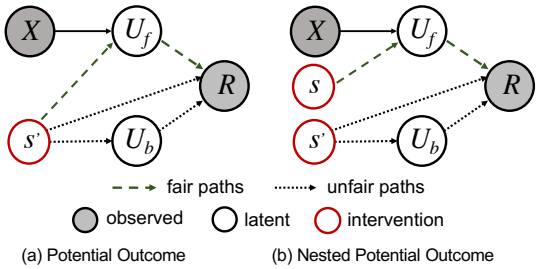


Figure 2: Comparisons between potential outcome that sets sensitive features S to $\mathbf{s}'$ and nested potential outcome that sets S to different values along different causal paths.

2.2.3 ***Path-Specific Counterfactuals.*** After introducing the latent factors U_f and U_b that mediate the fair and unfair influences of sensitive features S on observed ratings R and the causal graph in Fig. 1, we are ready to define the unfairness inherent in the dataset $\mathcal{D}$, which is a crucial first step toward achieving fairness in RSs.

According to the causal graph in Fig. 1, we can represent the variation of R due to S (with fixed X) in $\mathcal{D}$ with the distribution $p(R|S,X)$, which is governed by latent mediators U_f, U_b as follows:

$$p(R|S,X) = \mathbb{E}_{p(U_f|S,X),p(U_b|S)}\left[p(R|U_f,U_b)\right], \tag{1}$$

where $\mathcal{F} = \{p(R|U_f,U_b), p(U_f|S,X), p(U_b|S)\}$ are the **structural equations** associated with the causal graph. However, we should note that not all variations of R due to S encapsulated in $p(R|S,X)$ are discriminatory, as the causal influences of S mediated by U_f, e.g., the cultural-based preferences ($S \rightarrow C_t \rightarrow R$), are crucial manifestations of diversity and personalization in user interests.

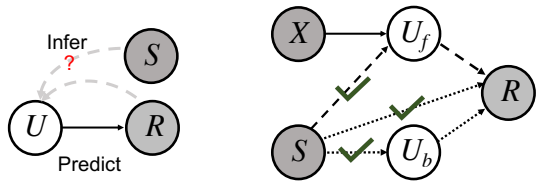
To address the above challenge, we measure the unfair variation of R due to S with path-specific counterfactual inference [22], where we determine how ratings R will change if users' sensitive features S are set to a counterfactual value $\mathbf{s}'$ along the unfair paths $S \rightarrow U_b \rightarrow R$ and $S \rightarrow R$, while maintaining its factual value $\mathbf{s}$ along the fair path $S \rightarrow U_f \rightarrow R$. To achieve this objective, it is necessary to introduce the Nested Potential Outcome (NPO) defined as follows:

Definition 2.1. We use the ***Nested Potential Outcome*** *(NPO) $R_{S\leftarrow\mathbf{s}'}(U_{f,S\leftarrow\mathbf{s}}, U_{b,S\leftarrow\mathbf{s}'})$ to denote the random variable of user ratings where user sensitive features S are set to $\mathbf{s}'$ on the unfair paths $S \rightarrow R$ and $S \rightarrow U_b \rightarrow R$ and to $\mathbf{s}$ on the fair path $S \rightarrow U_f \rightarrow R$.*

The NPO $R_{S\leftarrow\mathbf{s}'}(U_{f,S\leftarrow\mathbf{s}}, U_{b,S\leftarrow\mathbf{s}'})$ can be intuitively represented by an intervened causal graph in Fig. 2-(b). However, the unconditional NPO reasons with the intervention conducted upon the whole population, whose factual sensitive features S do not necessarily equal $\mathbf{s}$. Therefore, to constrain the NPO to users with factual sensitive feature $S = \mathbf{s}$ (and non-sensitive features $X = \mathbf{x}$) such that the fair influence of $S = \mathbf{s}$ on R is excluded from the unfairness measurement, we condition it on $X = \mathbf{x}$ and $S = \mathbf{s}$ as follows:

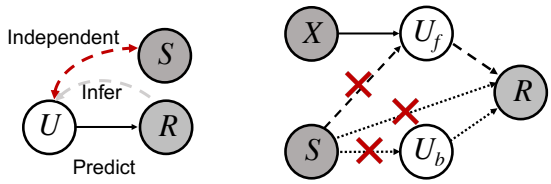
$$R_{S\leftarrow\mathbf{s}'}(U_{f,S\leftarrow\mathbf{s}}, U_{b,S\leftarrow\mathbf{s}'})|X = \mathbf{x}, S = \mathbf{s}. \tag{2}$$

The conditional NPO described in Eq. (2) essentially reasons with the observed ratings of hypothetical users whose sensitive features



(a) Assumption of naïve RS (b) How S influences predictions

Figure 3: Naive RS that infers U from R and (possibly) S for rating predictions. If the inference is accurate, all influences of S on R are allowed in future recommendations.



(a) Assumption of existing fair RS (b) How S influences predictions

Figure 4: Existing fair RS that constrains the inferred U to be independent of S. If the constraint is satisfied, both fair and unfair influences of S are blocked in recommendations.

are in a "superposition" state: Their sensitive features S preserve the factual value $S = \mathbf{s}$ along the fair path $S \to U_f \to R$ while having the counterfactual value $S = \mathbf{s}'$ along the unfair paths $S \to U_b \to R$ and $S \to R$. This allows the theoretical analysis of path-specific bias/fairness of different RS models in the following subsections.

2.3 Unfairness of Naive RSs

Based on the conditional NPO, we are now ready to formally analyze the unfairness of naive RSs whose rating predictions are *consistent* with the causal mechanisms that generate the biased observed ratings. We show that even if these models do not directly use sensitive features S for recommendations, they can still capture the unfair correlations between S and R and make biased recommendations.

2.3.1 ***Path-Specific Bias for Naive RSs.*** Naive RSs assume that the observed ratings R are generated from user latent variables U via generative distribution $p_{naive}(R|U)$[3], where $p_{naive}(R|U)$ and U can be obtained by maximizing the log-likelihood $\mathcal{L}$ of the observed ratings R (and possibly with the support of user features S and X) via factorization [35] or variational inference [28]. The inferred U and the generative distribution $p_{naive}(R|U)$ are then used to predict new ratings for recommendations (Fig. 3-(a)). If the learned generative and inference distributions of the naive RSs are accurate, U captures all latent factors that causally influence the observed user behaviors R, i.e., $U = \{U_f, U_b\}$ (or its bijective), and $p_{naive}(R|U)$ is consistent with the causal mechanism that generates the observed ratings, i.e., $p(R|U_f, U_b)$. Therefore, the unfairness of the naive RSs can be quantified by the path-specific effects of S on R through the unfair paths on the factual causal graph, which can be defined as:

$$\begin{aligned} PSBias(\mathbf{x}, \mathbf{s}, \mathbf{s}') = &\mathbb{E}\left[R_{S\leftarrow\mathbf{s}'}\left(U_{f,S\leftarrow\mathbf{s}}, U_{b,S\leftarrow\mathbf{s}'}\right)\middle|X=\mathbf{x}, S=\mathbf{s}\right] \\ &- \mathbb{E}\left[R_{S\leftarrow\mathbf{s}}\left(U_{f,S\leftarrow\mathbf{s}}, U_{b,S\leftarrow\mathbf{s}}\right)\middle|X=\mathbf{x}, S=\mathbf{s}\right]. \end{aligned} \tag{3}$$

Intuitively, for users with factual features $X = \mathbf{x}$ and $S = \mathbf{s}$, path-specific bias $PSBias(\mathbf{x}, \mathbf{s}, \mathbf{s}')$ defined in Eq. (3) denotes the difference of rating predictions from naive RSs if their sensitive features S change to $\mathbf{s}'$ along the unfair paths $S \to R$ and $S \to U_b \to R$, while S is held unchanged along the fair path $S \to U_f \to R$, and the non-sensitive features X are held unchanged along all the paths. $PSBias(\mathbf{x}, \mathbf{s}, \mathbf{s}')$ won't be zero for naive RSs if causal path $S \to U_b \to R$ is not trivial, but the claim is not self-evident from Eq. (3), and we show how to calculate $PSBias(\mathbf{x}, \mathbf{s}, \mathbf{s}')$ in the next subsection.

[3]We use p_{model} to represent the distributions assumed by an RS model, which should be distinguished with the structural causal equations p (with no subscription) in $\mathcal{F}$ that describe the causal generative process of the biased observed ratings.

2.3.2 ***Calculation of PS-Bias.*** It is generally intractable to calculate $PSBias(\mathbf{x}, \mathbf{s}, \mathbf{s}')$ because it contains NPOs that reason with hypothetical users with counterfactual sensitive features $S = \mathbf{s}'$ along the unfair paths. However, with the Sequential Ignorability Assumption commonly used in causal mediation analysis [19], the first counterfactual term in Eq. (3) can be calculated as follows:

$$\begin{aligned} &\mathbb{E}\left[R_{S\leftarrow\mathbf{s}'}\left(U_{f,S\leftarrow\mathbf{s}}, U_{b,S\leftarrow\mathbf{s}'}\right)\middle|X=\mathbf{x}, S=\mathbf{s}\right] \\ &= \int_{\mathbf{r},\mathbf{u}_f,\mathbf{u}'_b} p(\mathbf{r}|\mathbf{s}', \mathbf{u}_f, \mathbf{u}'_b)\cdot p(\mathbf{u}_f|\mathbf{s}, \mathbf{x})\cdot p(\mathbf{u}'_b|\mathbf{s}')\cdot \mathbf{r} \\ &= \int_{\mathbf{r},\mathbf{u}_f,\mathbf{u}'_b} p(\mathbf{r}|\mathbf{u}_f, \mathbf{u}'_b(\mathbf{s}'))\cdot p(\mathbf{u}_f|\mathbf{s}, \mathbf{x})\cdot p(\mathbf{u}'_b|\mathbf{s}')\cdot \mathbf{r}, \end{aligned} \tag{4}$$

where in the final step, we summarize the direct unfair influence of sensitive features S on ratings R into U_b for simplicity. The rigorous proof can be referred to in Appendix B.1. Similarly, the second factual term in Eq. (3) can be calculated as follows:

$$\begin{aligned} &\mathbb{E}\left[R_{S\leftarrow\mathbf{s}}\left(U_{f,S\leftarrow\mathbf{s}}, U_{b,S\leftarrow\mathbf{s}}\right)=\mathbf{r}\middle|X=\mathbf{x}, S=\mathbf{s}\right] \\ &= \int_{\mathbf{r},\mathbf{u}_f,\mathbf{u}_b} p(\mathbf{r}|\mathbf{u}_f, \mathbf{u}_b)\cdot p(\mathbf{u}_f|\mathbf{s}, \mathbf{x})\cdot p(\mathbf{u}_b|\mathbf{s})\cdot \mathbf{r}, \end{aligned} \tag{5}$$

where Eqs. (4) and (5) can be plugged into Eq. (3) to calculate the $PSBias(\mathbf{x}, \mathbf{s}, \mathbf{s}')$. Clearly, $PSBias(\mathbf{x}, \mathbf{s}, \mathbf{s}')$ for naive RSs cannot be zero, because sensitive features S can unfairly influence the observed ratings R via the user bias latent variable U_b, which makes the $p(U_b|S)$ and $p(R|U_f, U_b)$ terms in Eqs. (4) and (5) non-trivial.

2.4 Minimal Change Principle and Over-Fairness of Existing Fair RSs

To remedy the bias, existing fair RSs impose constraints upon the naive RSs. An exemplar strategy is to maximize the log-likelihood $\mathcal{L}$ of the observed ratings in $\mathcal{D}$, i.e., $\mathcal{D}_R$, while constraining the inferred user latent variables U to be independent of the sensitive features S (see Fig. 4-(a)). This can be formulated as follows:

$$\max_{U,\, p_{ef}} \ \mathcal{L}\left(p_{ef}(R \mid U); \mathcal{D}_R\right) \ s.t., U \perp\!\!\!\perp S. \tag{6}$$

The constraint can be implemented via strategies such as adversarial training [26] or maximum mean discrepancy (MMD) minimization [30]. To satisfy such a constraint, the causal mechanisms $p(U_f|S, X)$ and $p(U_b|S)$ that underlie the generation of the observed ratings must be altered into $p_{ef}(U_f|X)$, $p_{ef}(U_b)$ by dropping the dependence on S, which can be represented by a new causal graph illustrated in Fig. 4-(b) (with causal edges marked by × removed).

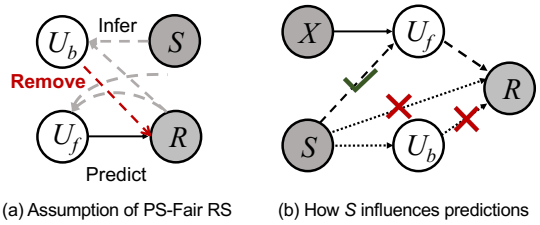


Figure 5: PSF-RS that minimally changes the biased factual world represented by Fig. 1 into a hypothetically fair world, where a PS-Fair RS model can be learned accordingly.

We can prove that existing fair RSs can eliminate the PS-Bias if the constraint is tight, such that U and S are strictly independent (see Appendix B.2 and B.3 for details). However, it can also lead to over-fairness issues, where the causal structure $p(U_f|S,X)$ that denotes the fair influences of S on R mediated by U_f is destroyed. Therefore, necessary diversities in recommendations due to the fair influence of sensitive features (e.g., cultural diversity) can be undesirably lost. Essentially, the independence constraint of existing fair RSs is against the Minimal Change Principle of Pearl [39], which states that counterfactuals (i.e., a fair rating generation model) should be reasoned with by minimally adjusting the factual world (i.e., the causal model that generates biased observed ratings).

2.5 Path-Specific Fairness for RSs

To address the over-fairness drawbacks of existing fair RSs, we propose a path-specific fair RS, i.e., PSF-RS, that minimally alters the biased factual world (represented by the causal graph in Fig. 1) into a hypothetically fair world, and based on it generates new ratings for recommendations. Specifically, we aim to find a counterfactual distribution $p_{psf}(R|U_f,U_b)$ close to the factual distribution $p(R|U_f,U_b)$ that causally generates the biased observed ratings (measured by KL-divergence), while inducing a new causal model with zero $PSBias^*(\mathbf{x},\mathbf{s},\mathbf{s}')$[4], where other factual causal mechanisms in $\mathcal{F}$, i.e., $p(U_f|S,X)$ and $p(U_b|S)$, remain unchanged.

Assuming for now that the latent mediators U_f and U_b are known for each user (where the inference of U_f and U_b with weak supervision in R_b will be thoroughly discussed in the next section), since the observed ratings R in the dataset $\mathcal{D}$ are generated according to $p(R|U_f,U_b)$, the minimization of the KL between $p_{psf}(R|U_f,U_b)$ and $p(R|U_f,U_b)$ is equivalent to the maximization of the likelihood $\mathcal{L}$ of the observed ratings in $\mathcal{D}$. Therefore, the objective of PSF-RS can be formulated as a constrained optimization problem as follows:

$$\max_{p_{psf}} \mathcal{L}\left(p_{psf}(R \mid U_f,U_b);\mathcal{D}_R\right) \; s.t.,\; PSBias^*(\mathbf{x},\mathbf{s},\mathbf{s}') = 0,\; \forall \mathbf{x},\mathbf{s},\mathbf{s}'. \tag{7}$$

The constraint essentially restricts the family of RS models that we can use for recommendations into the ones that induce a new causal model with zero $PSBias^*(\mathbf{x},\mathbf{s},\mathbf{s}')$. The simplest distribution family that satisfies the constraint is the one that uses only U_f to generate recommendations, i.e., $p_{psf}(R \mid U_f)$ (see Appendix B.4 for the proof of zero PS-Bias for the PSF-RS). The newly-induced causal graph that changes $p(R|U_f,U_b)$ to $p_{psf}(R|U_f)$ while keeping $p(U_f|S,X)$ and $p(U_b|S)$ intact is shown in Fig. 5-(b) for reference.

[4]we use * to distinguish the PS-Bias of new causal model induced by PSF-RS from the PS-Bias of naive RSs that recommend according to the biased factual causal model.

3 PS-FAIR VARIATIONAL AUTO-ENCODER

Previous sections have demonstrated PSF-RS's theoretical advantage of achieving path-specific fairness while maximally preserving the necessary diversities in recommendations. However, its practical implementation still faces two challenges as follows:

- First, since both fair and unfair mediators of S, i.e., U_f and U_b, are latent, the objective of PSF-RS in Eq. (7) cannot be directly optimized to obtain the PS-Fair rating predictor $p_{psf}(R|U_f)$.
- In addition, although the known unfair items R_b, i.e., another indirect causal effect of S mediated by U_b, can be used to infer U_b and distinguish it from U_f, R_b is extremely sparse and is only partially observable for a small subset of users.

To address the aforementioned challenges, we propose a novel semi-supervised deep generative model called path-specific fair variational auto-encoder (PSF-VAE) as the implementation of PSF-RS. Specifically, in the *factual* modeling step, PSF-VAE infers U_f and U_b from the biased observational ratings R in the dataset $\mathcal{D}$ via deep neural networks (DNNs), where user features S and X are used as extra covariates and R_b as additional weak supervision signals. Then, in the *counterfactual* reasoning step, U_b that explains away the unfair influences of S is eliminated according to Eq. (7), and U_f that maximally preserves the fair influence of S and other aspects of user interests is utilized to generate new recommendations.

3.1 Factual Generative Process

The factual generative process of PSF-VAE is consistent with the causal model in Fig. 1, such that latent mediators U_f and U_b can be properly inferred from the biased observational data. PSF-VAE starts by generating for each user the user fair and bias latent mediators U_f and U_b from Gaussian priors $p_\theta(U_f|S,X)$ and $p_\theta(U_b|S)$ as

$$\mathbf{u}_f \sim \mathcal{N}(f_{uf}([\mathbf{s}||\mathbf{x}]), \mathbf{I}_{K_f}),\; \mathbf{u}_b \sim \mathcal{N}(f_{ub}(\mathbf{s}), \mathbf{I}_{K_b}), \tag{8}$$

where f_{uf} and f_{ub} are two functions, $[\cdot||\cdot]$ represents vector concatenation, and θ denotes the trainable parameters associated with the generative network, respectively. Then, for the small subset of users with known unfair items $\mathbf{r}_b$, $\mathbf{r}_b$ are generated from $\mathbf{u}_b$ via $p_\theta(R_b|U_b)$ parameterized as the following Bernoulli distribution,

$$\mathbf{r}_b \sim Bernoulli(MLP_b(\mathbf{u}_b)), \tag{9}$$

where MLP_b is a multi-layer perceptron (MLP) with sigmoid final layer activation ($sigmoid(\mathbf{x}) = 1/(1+e^{-\mathbf{x}})$). Finally, the observed ratings $\mathbf{r}$ are generated from both $\mathbf{u}_f$ and $\mathbf{u}_b$ via $p_\theta(R|U_f,U_b)$ parameterized as the following multinomial distribution,

$$\mathbf{r} \sim Multi(MLP_r([\mathbf{u}_f||\mathbf{u}_b]), N), \tag{10}$$

where MLP_r is another MLP with softmax final layer activation, i.e., $[softmax(\mathbf{x})]_i = e^{\mathbf{x}_i}/\sum_j e^{\mathbf{x}_j}$; N is the number of interacted items.

3.2 Weakly-Supervised Variational Inference

Given that the (factual) generative distributions of both R and R_b are parameterized by DNNs, and R_b is only partially observable for a small subset of users, the true posterior distributions of the latent variables, i.e., $p_\theta(U_f|R,S,X)$ and $p_\theta(U_b|R_b,R,S)$, are intractable.

Therefore, we resort to variational inference [4, 28], where we introduce tractable distribution families of U_f and U_b parameterized by DNNs with trainable parameters ϕ, i.e., $q_\phi(U_f|\cdot)$ and $q_\phi(U_b|\cdot)$, and in q_ϕ find the distributions closest to the true but intractable posteriors measured by KL-divergence as the approximations.

The variational posterior for U_f, i.e., $q_\phi(U_f|R,S,X)$, is straightforward. However, for U_b, we eschew the normally-adopted variational posterior $q_\phi(U_b|R_b,R,S)$ but use $q_\phi(U_b|R,S)$ with R_b omitted instead, such that the inference of U_b does not depend on the partially observed R_b. Therefore, it can be generalized to users with no observed unfair items. Under such circumstances, if R and S contain sufficient information of R_b, which can be guaranteed since both R and R_b are under the unfair causal influence of S mediated by U_b, weak supervision signals in R_b from the subset of users with observed unfair items can still guide the training of the inference network $q_\phi(U_b|R,S)$ to provide good variational approximations.

3.3 Evidence Lower Bound

The minimization of the KL-divergence between variational and true posterior distributions is equivalent to the maximization of the evidence lower bound (ELBO) as (proofs see Appendix B.5)[5]:

$$\begin{aligned} ELBO = &\mathbb{E}_{q_\phi(U_f,U_b|\cdot)}[\ln p_\theta(R|U_f,U_b)] + \mathbb{E}_{q_\phi(U_b|\cdot)}[p_\theta(R_b|U_b)] \\ &- \mathbb{KL}[q_\phi(U_f|R,S,X)||p_\theta(U_f|S,X)] - \mathbb{KL}[q_\phi(U_b|R,S) \,||\, p_\theta(U_b|S)], \end{aligned} \tag{11}$$

which is a lower bound of the model evidence $\ln p_\theta(R,R_b|S,X)$. In Eq. (11), the first two terms are the expected log-likelihood of R and R_b given the latent mediators U_f and U_b, which encourage U_f and U_b to best explain the observed biased ratings (where the bias in R is explained-away from U_f by U_b), and the last two terms are the KL-divergence between the variational posteriors and the priors.

For users with no observed unfair items R_b, the second expected log-likelihood term $\mathbb{E}_{q_\phi(U_b|R,S)}[p_\theta(R_b|U_b)]$ is dropped from the ELBO, and we only use the observed ratings R and the user sensitive features S to infer the corresponding user bias latent variable U_b via the variational posterior $q_\phi(U_b|R,S)$. For these users, when maximizing the first term of the ELBO, i.e., $\mathbb{E}_{q_\phi(U_f,U_b|\cdot)}[\ln p_\theta(R|U_f,U_b)]$, the inferred U_b **can still help explain away** the unfair influence of S on R, such that U_f can focus exclusively on capturing the fair user interests that are generalizable to future recommendations.

3.4 Disentanglement via Adversarial Training

Before introducing $p_{psf}(R|U_f)$ that minimally changes the biased factual world into a hypothetically fair world to make fair recommendations, we note that the theoretical PS-Fairness of PSF-RS requires a correctly specified inference model (as Eq. (7) requires known U_f and U_b). Especially, we need to ensure $U_f \perp\!\!\!\perp U_b|S$, which prevents U_f from **directly** depending on U_b, such that the unfair information of S cannot be leaked to U_f. Since the true posteriors of U_f and U_b are not guaranteed to be in the variational family q_ϕ, the unfair information of S in U_b may be leaked to U_f due to potential mis-specification of the inference model, especially when supervision signals in R_b are available only for a subset of users.

[5]In practice, we further simplify the ELBO by dropping the dependence of the priors of U_f and U_b on S and X, i.e., $\mathbf{u}_f \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{K_f})$, $\mathbf{u}_b \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{K_b})$. In addition, we first optimize the U_b-specific terms in the ELBO, and then fix U_b and learn other terms.

We utilize an adversarial training-based strategy [14] to ensure the conditional independence of U_f and U_b given S in case of inference model mis-specification. Following [2], we first parameterize a discriminator model p_d that predicts U_b from U_f and S as:

$$p_d(U_b|U_f,S) = \mathcal{N}(MLP_d([U_f||S]), \mathbf{I}_{K_d}). \tag{12}$$

Then, concurrent with the maximization of the ELBO in Eq. (11), U_f and U_b obtained from variational posteriors q_ϕ are used to train the discriminator p_d. Specifically, we fix $\hat{q}_\phi(U_b|R,S)$, sample $\hat{\mathbf{u}}_b$ from it and train the discriminator $p_d(U_b|U_f,S)$ to best predict $\hat{\mathbf{u}}_b$ from U_f and S. Meanwhile, we constrain the inference model of U_f, i.e., $q_\phi(U_f|R,S,X)$, to fool the discriminator. The above process can be formulated as a GAN-like mini-max game as follows:

$$\min_{q_\phi} \max_{p_d} \mathbb{E}_{q_\phi(U_f|R,S,X)}[\ln p_d(\hat{\mathbf{u}}_b|U_f,S)], \;\; \hat{\mathbf{u}}_b \sim \hat{q}_\phi(U_b|R,S). \tag{13}$$

With a sufficient capacity of the discriminator p_d, Li et al. [26] showed that $U_f \perp\!\!\!\perp U_b|S$ holds when the equilibrium of Eq. (13) is achieved. Therefore, the direct dependence of U_f on U_b that leads to the leak of unfair information of S can be further mitigated.

3.5 PS-Fair Rating Predictions

Finally, we introduce $p_{psf}(R|U_f)$, the counterfactual rating generator that minimally modifies the biased factual world while ensuring path-specific fairness and necessary diversities in recommendations. Specifically, after optimizing the "factual step" of PSF-VAE via Eqs. (11) and (13), we fix $q_\phi(U_f|R,S,X)$ and obtain the user fair latent variables $\hat{\mathbf{u}}_f$ as the posterior mean. Then the PS-Fair rating predictor $p_{psf}(R|U_f)$ can be obtained by optimizing Eq. (7) with the inferred $\hat{\mathbf{u}}_f$ and the observed ratings $\mathbf{r}$. Specifically, we parameterize $p_{psf}(R|U_f)$ as the following multinomial distribution,

$$\mathbf{r} \sim Multi(MLP_{psf}(\hat{\mathbf{u}}_f), N), \tag{14}$$

where MLP_{psf} is another MLP with softmax as the last layer activation. Finally, the multinomial probabilities of all previously uninteracted items can be obtained via $p_{psf}(R|U_f)$, which are then ranked such that M most relevant ones are fetched for recommendations.

4 EXPERIMENTS

In this section, we present the extensive experiments conducted on two semi-simulated datasets and one real-world dataset to demonstrate the effectiveness of the proposed PSF-VAE, with an emphasis on answering the following three research questions[6]:

- **RQ1.** How well can PSF-VAE achieve fairness compared with different RS methods with and without fairness constraints?
- **RQ2.** How well can PSF-VAE preserve necessary fair influences of sensitive features compared with existing fair RS algorithms?
- **RQ3.** How does the number of users with known unfair items R_b influences the fairness performance of PSF-VAE?

4.1 Datasets

It is difficult to directly evaluate PSF-VAE on real-world datasets, as the true fair and unfair causal effects of sensitive features on the observed ratings cannot be identified from the datasets. Therefore, we first establish semi-simulated datasets with known causal mechanisms between sensitive features and rating observations. We then

[6]Codes are available at https://github.com/yaochenzhu/PSF-VAE.

Table 1: Statistics of the semi-simulated (ML-1M and AM-VG) and the real-world (LinkedIn) datasets. #Int. stands for the number of observed interactions. Sps. (R) and Sps. (R_b) denote the sparsity of observed ratings, unfair items, respectively.

Dataset	#Int.	#Users	#Items	Sps. (R)	Sps. (R_b)
ML-1M	993,504	6,000	3,706	95.53%	99.76%
AM-VG	127,741	7,253	4,338	99.60%	99.93%
LinkedIn	1,055,241	8,896	5,931	98.01%	99.62%

introduce a real-world dataset collected from LinkedIn[7], where for a subset of users, their negative feedback on recommendations (i.e., explicit dismissals of Ads) is treated as the proxy of unfair items.

4.1.1 ***Semi-Simulated Dataset.*** The semi-simulated datasets are established based on the widely-used MovieLens-1M (ML-1M) [16] and Amazon Videogames (AM-VG) datasets [33]. For each dataset, we train a Multi-VAE model [28] on the binarized ratings, where the decoder $f_{gen} = MLP_{gen}(\mathbf{u})$ maps the user latent variable $U \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$ to the multinomial parameters $\tilde{R}$ of the ratings R. The latent dimension K is fixed to 200 as [28]. We then assume that the first K_f and the remaining $K_b = K - K_f$ dimensions of U, which we denote as U_f and U_b, mediate the fair and unfair influences of sensitive features S on the observed ratings R, respectively. In the simulation, for each user, we first generate a confounder $\mathbf{c} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{K_f})$ that simultaneously affects $\mathbf{u}_f$ and $\mathbf{u}_b$, where user sensitive features $\mathbf{s}$ are derived from $\mathbf{c}$ by $PCA(\mathbf{c}, K_s)$. The fair and unfair latent mediators $\mathbf{u}_f$ and $\mathbf{u}_b$ are then generated as follows:

$$\mathbf{u}_f = \lambda_f \mathbf{c} + \sqrt{(1-\lambda_f^2)}\boldsymbol{\epsilon}_f; \;\; \mathbf{u}_b = \lambda_b Redim(\mathbf{c}, K_b) + \sqrt{(1-\lambda_b^2)}\boldsymbol{\epsilon}_b,$$

where the exogenous variables $\boldsymbol{\epsilon}_f \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{K_f})$, $\boldsymbol{\epsilon}_b \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{K_b})$, the function $Redim$ reduces the dimension of $\mathbf{c}$ to K_b through random selection, and the coefficients λ_f and λ_b determine the noise level of $\mathbf{u}_f$ and $\mathbf{u}_b$, which are empirically fixed as 0.9 and 0.9, respectively.

The observed ratings are generated from $\mathbf{u}_f$ and $\mathbf{u}_b$ by first calculating the multinomial parameters $\tilde{\mathbf{r}} = f_{gen}([\mathbf{u}_f || \mathbf{u}_b])$, where the top $100 \times p_r\%$ (ranked among all users) are selected as the rating observations $\mathbf{r}$. p_r is set to be the same as the original datasets. The unfair items $\mathbf{r}_b$ are simulated with the sub-network f^b_{gen} in f_{gen} that corresponds to $\mathbf{u}_b$[8]. Similarly, we obtain the multinomial parameters $\tilde{\mathbf{r}}_b = f^b_{gen}(\mathbf{u}_b)$, where the top $100 \times p_b\%$ are selected as the unfair items. p_b is determined such that the ratio of the average number of observed ratings and unfair items is the same as the real-world dataset introduced later. We do not simulate non-sensitive features $\mathbf{x}$ because the sequential ignorability assumption automatically holds with the above data generation process.

4.1.2 ***Real-World Dataset.*** In addition, we collect a real-world dataset from LinkedIn for job recommendations, where ratings R denote users' interactions with the job Ads. We use the data where users actively dismissed the recommended jobs as substitutes for the unfair items R_b. User sensitive features S include age, gender, and education level, all of which can influence the job recommendation in a fair manner. For example, age can determine the experience and seniority of the users, whereas education level can determine their knowledge and skills. To avoid privacy issues in user data collection, we train a generative model (VAE) to encode the raw data into a joint distribution $p_{gen}(R, R_b, S)$ where S is embedded into a 50-dimensional continuous vector, and we generate anonymized data from p_{gen} accordingly for the experiments to protect privacy [52]. The statistics of the datasets are summarized in Table 1.

4.2 Experimental Settings

4.2.1 ***Setups.*** In our experiments, we randomly split the users into train, validation, and test sets based on the ratio of 8:1:1 [28]. For each user, 20% of the observed ratings are held out for evaluation. For the ML-1M and AM-VG datasets, the simulated unfair items $\mathbf{r}_b$ for $100 \times (1 - c_r)\%$ of the training and validation users are masked out as zero (where c_r is set to 0.3 as with the LinkedIn dataset), while $\mathbf{r}_b$ for all test users are used to obtain unbiased evaluations of the fairness of different methods. In our experiments, we first fix the simulated dimension of U_b, i.e., K_b, to 50 in the ML-1M and AM-VG datasets to compare the recommendation performance and fairness across different methods. We then simulate the datasets with varied K_b to further demonstrate the robustness of PSF-VAE to different levels of unfair correlations between observed ratings and sensitive features. Finally, we show the sensitivity of PSF-VAE to the percentage of users with observed unfair items. All reported results are averaged over ten random splits of the datasets.

4.2.2 ***Evaluation Metrics.*** We evaluate different RSs from two aspects: *recommendation performance* and *fairness*. The recommendation performance is measured by two widely-used ranking-based metrics: Recall (R@M) and truncated normalized discounted cumulative gain (N@M)[9]. Fairness is measured by the hit rate of top M items on unfair items (HiR@M). For the semi-simulated datasets, the true unfair items $\mathcal{J}_{b,i}$ are available for all test users, while for the LinkedIn dataset, we can only calculate HiR @M for test users with observed unfair items. In our experiments, we find that M generally does not affect the relative performance of different methods. Therefore, we set M to 20 for Recall and 100 for NDCG as with [28], and set M to 10 for HiR due to the sparsity of observed unfair items.

4.2.3 ***Model Selection.*** During the training stage, we monitor the composite metric $Met_{rf}(i)$ = R@20(i) + N@100(i) - HiR@10(i) on validation users with known unfair items and $Met_r(i)$ = R@20(i) + N@100(i) on validation users with no observed unfair items, and calculate the weighted average of Met_{rf} and Met_r, i.e., $\hat{Met}$, over all validation users. We then select the model with the largest $\hat{Met}$ and report the recommendation and fairness metrics on test users.

4.3 Comparisons with Baselines

4.3.1 ***Baseline Descriptions.*** To answer **RQs 1** and **2**, we compare the proposed PSF-VAE with various state-of-the-art RSs with/without fairness-aware mechanisms. The main baselines included for comparisons can be categorized into four classes as follows:

- **Unawareness.** RSs with unawareness use only seemingly non-sensitive information (i.e., observed ratings and non-sensitive

[7] https://www.linkedin.com/.

[8] If we denote $f_{gen}(\mathbf{u})$ as $\tilde{f}_{gen}(\mathbf{W}\mathbf{u} + \mathbf{b})$, the subnetwork can be obtained by $f^b_{gen} = \tilde{f}_{gen}(\mathbf{W}_{:,K-K_b:K}\mathbf{u}_b + \mathbf{b})$, where $\mathbf{W}_{:,K-K_b:K}$ selects the last K_b columns of $\mathbf{W}$.

[9] We also use the recommendation quality (i.e., R@M and N@M) as an indirect measure of RSs' ability to preserve the fair influences of sensitive features on ratings.

Table 2: Comparison between PSF-VAE and various baselines. ↑ denotes the larger the better, while ↓ denotes the opposite.

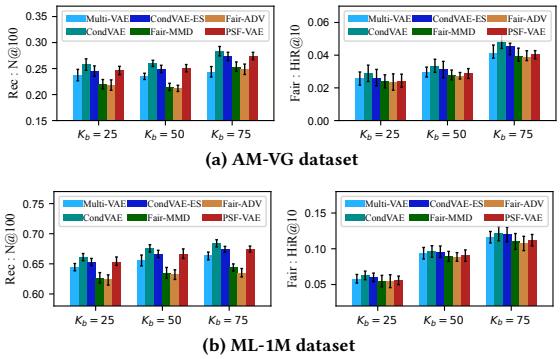
AM-VG	Rec: R@20 ↑	Rec: N@100 ↑	Fair: HiR@10 ↓
Multi-VAE	0.2454 ± 0.0130	0.2350 ± 0.0093	0.0297 ± 0.0030
CondVAE	**0.2780** ± 0.0103	**0.2599** ± 0.0058	0.0315 ± 0.0045
CondVAE-ES	0.2686 ± 0.0115	0.2493 ± 0.0061	0.0302 ± 0.0053
Fair-MMD	0.2304 ± 0.0118	0.2147 ± 0.0094	0.0279 ± 0.0025
Fair-ADV	0.2285 ± 0.0081	0.2119 ± 0.0076	**0.0274** ± 0.0020
PSF-NN	0.2702 ± 0.0124	0.2549 ± 0.0095	0.0310 ± 0.0029
PSF-VAE	0.2691 ± 0.0104	0.2507 ± 0.0075	0.0288 ± 0.0032

ML-1M	Rec: R@20 ↑	Rec: N@100 ↑	Fair: HiR@10 ↓
Multi-VAE	0.5493 ± 0.0133	0.6556 ± 0.0064	0.0938 ± 0.0075
CondVAE	**0.5689** ± 0.0145	**0.6757** ± 0.0065	0.0953 ± 0.0077
CondVAE-ES	0.5615 ± 0.0151	0.6665 ± 0.0069	0.0949 ± 0.0080
Fair-MMD	0.5312 ± 0.0119	0.6350 ± 0.0069	0.0893 ± 0.0074
Fair-ADV	0.5304 ± 0.0129	0.6348 ± 0.0060	**0.0886** ± 0.0063
PSF-NN	0.5654 ± 0.0104	0.6701 ± 0.0051	0.0942 ± 0.0040
PSF-VAE	0.5601 ± 0.0148	0.6668 ± 0.0070	0.0904 ± 0.0084

LinkedIn	Rec: R@20 ↑	Rec: N@100 ↑	Fair: HiR@10 ↓
Multi-VAE	0.1665 ± 0.0043	0.2553 ± 0.0046	0.0703 ± 0.0034
CondVAE	**0.2056** ± 0.0037	**0.3042** ± 0.0031	0.0718 ± 0.0037
CondVAE-ES	0.1991 ± 0.0047	0.2965 ± 0.0036	0.0705 ± 0.0023
Fair-MMD	0.1579 ± 0.0054	0.2398 ± 0.0066	0.0608 ± 0.0040
Fair-ADV	0.1573 ± 0.0062	0.2372 ± 0.0070	**0.0591** ± 0.0034
PSF-NN	0.2032 ± 0.0024	0.3005 ± 0.0028	0.0709 ± 0.0023
PSF-VAE	0.2024 ± 0.0045	0.2987 ± 0.0034	0.0647 ± 0.0029

features) for recommendations. In this regard, the Unawareness counterpart of PSF-VAE is the vanilla **Multi-VAE** [28].

- **Naive.** Naive RSs explicitly utilize the sensitive features S for recommendations. In our case, it can be implemented as a generalized Multi-VAE where the rating inputs are augmented with the sensitive features S. The augmentation is implemented as with the user conditional Multi-VAE (**CondVAE**) in [38].
- **Total Fairness.** RSs with total fairness block all the effects of sensitive features S on recommendations. Built upon the Unawareness model (i.e., Multi-VAE), the inferred user latent variables U are constrained to be disentangled from the user sensitive features S while fitting on the observed ratings R. We consider the following two disentanglement strategies:
  - **Fair-ADV**. Fair-ADV constrains the user latent variables of Multi-VAE to be independent with sensitive features S via adversarial training; details can be referred to in [26].
  - **Fair-MMD**. Fair-MMD minimizes the maximum mean discrepancy (MMD) of user latent variables given sensitive features S in Multi-VAE [30]. Specifically, we randomly select one dimension of S and binarize it for the minimization.
- **PS-Fairness.** We consider the following naive PS-Fair strategy for RSs, i.e., **PSF-NN**, where for each user, we calculate the similarities with all users with available R_b measured by sensitive features. Then we select the N closest neighbors, get the top K unfair items, and remove them if they appear in the list.



(a) AM-VG dataset

(b) ML-1M dataset

Figure 6: Comparison between PSF-VAE and baselines with different dimension of U_b, *i.e.*, K_b, for the simulated datasets.

Finally, since a simple strategy to improve the fairness over the Naive model is through underfitting on the observed ratings R, we design an early-stop baseline, **CondVAE-ES**, which has the closest N@100 on the validation users with PSF-VAE, to demonstrate the fairness improvement of PSF-VAE is not due to simple underfitting.

4.3.2 ***Comparison Results.*** The comparison between PSF-VAE and various baselines is shown in Table 2. The best results (compared across four classes) are shown in **bold**, and the runner-ups are underlined. In summary, we have the following observations: **(1)** By utilizing all information in sensitive features for recommendations, CondVAE has the best recommendation performance and the worst fairness. **(2)** By simply ignoring the sensitive features, the Unawareness model (Multi-VAE) has improved fairness over the Naive model, while the recommendation performance is decreased simultaneously. **(3)** RSs with Total Fairness further improve the fairness over Multi-VAE, since the correlations between sensitive features and observed ratings are removed from user latent variables. However, since the fair influences of sensitive features are indiscriminately discarded, they also have the worst recommendation performance. **(4)** Although PSF-NN achieves better fairness than CondVAE, the improvement is not significant. The reason could be that the nearest-neighbor strategy is too crude to model the complicated unfair influences of sensitive features on observed ratings. **(5)** PSF-VAE has much better recommendation performance than the Total Fairness models and better fairness than the Naive and Unawareness models, because PSF-VAE only blocks the unfair influence of sensitive features on ratings, while their fair effects on user interests are maximally preserved for recommendations.

In addition, we set the simulated dimension of U_b, i.e., K_b, to different values in the AM-VG and ML-1M datasets to change the relative strengths of fair and unfair causal influences of sensitive features on the observed ratings and repeat the experiments in Fig. 6, which further demonstrates that PSF-VAE achieves a better balance between the recommendation performance and fairness.

4.4 Ablation Study

In this section, we compare the proposed PSF-VAE with the following variants as the ablation study to further verify its effectiveness.

- **PSF-VAE-nLat** removes the user bias variable U_b and directly constrains the user latent variables U in Multi-VAE to be independent of the observed unfair items R_b via adversarial training.

Table 3: Comparisons between different variants of PSF-VAE.

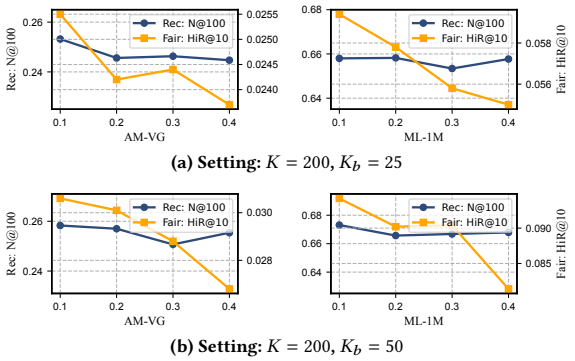
AM-VG	Rec: R@20 ↑	Rec: N@100 ↑	Fair: HiR@10 ↓
PSF-VAE-nLat	0.2276 ± 0.0080	0.2102 ± 0.0045	0.0270 ± 0.0022
PSF-VAE-nWSL	0.2729 ± 0.0084	0.2543 ± 0.0046	0.0299 ± 0.0021
PSF-VAE-nADV	0.2721 ± 0.0093	0.2528 ± 0.0061	0.0297 ± 0.0029
PSF-VAE-Mask	0.2624 ± 0.0096	0.2463 ± 0.0074	0.0291 ± 0.0031
PSF-VAE	0.2691 ± 0.0104	0.2507 ± 0.0075	0.0288 ± 0.0032
ML-1M	**Rec: R@20 ↑**	**Rec: N@100 ↑**	**Fair: HiR@10 ↓**
PSF-VAE-nLat	0.5163 ± 0.0152	0.6246 ± 0.0073	0.0869 ± 0.0083
PSF-VAE-nWSL	0.5647 ± 0.0135	0.6691 ± 0.0069	0.0932 ± 0.0081
PSF-VAE-nADV	0.5630 ± 0.0149	0.6687 ± 0.0075	0.0925 ± 0.0072
PSF-VAE-Mask	0.5577 ± 0.0132	0.6659 ± 0.0063	0.0911 ± 0.0068
PSF-VAE	0.5601 ± 0.0148	0.6668 ± 0.0070	0.0904 ± 0.0084
LinkedIn	**Rec: R@20 ↑**	**Rec: N@100 ↑**	**Fair: HiR@10 ↓**
PSF-VAE-nLat	0.1868 ± 0.0048	0.2832 ± 0.0035	0.0614 ± 0.0033
PSF-VAE-nWSL	0.2047 ± 0.0041	0.3009 ± 0.0032	0.0675 ± 0.0035
PSF-VAE-nADV	0.2032 ± 0.0046	0.3004 ± 0.0040	0.0660 ± 0.0039
PSF-VAE-Mask	0.2016 ± 0.0039	0.2969 ± 0.0051	0.0654 ± 0.0044
PSF-VAE	0.2024 ± 0.0045	0.2987 ± 0.0034	0.0647 ± 0.0029

- **PSF-VAE-nWSL** removes the weakly-supervised learning module of PSF-VAE, i.e., when fitting on the biased observed ratings R as Eq. (11), we only introduce the user bias latent variable U_b for the subset of users with observed unfair items R_b.
- **PSF-VAE-nADV** removes the adversarial training module in PSF-VAE that ensures the conditional independence between latent mediators U_f and U_b given user sensitive features S.
- **PSF-VAE-Mask** trains the same generative and inference networks as PSF-VAE. However, instead of learning a new model $p_{psf}(R|U_f)$, it masks out the weights in $p_\theta(R|U_f, U_b)$ that correspond to U_b, which leads to a new distribution $p_{masked(\theta)}(R|U_f)$, and uses $p_{masked(\theta)}$ to make the recommendations.

From Table 3 we can find that, PSF-VAE-nLat has the worst recommendation performance among all the variants, which shows that directly conducting adversarial training on the observed unfair items $\mathbf{r}_b$ is not stable, as $\mathbf{r}_b$ are high dimensional sparse vectors. In addition, PSF-VAE-nWSL, PSF-VAE-nADV, and PSF-VAE-Mask have worse fairness compared with PSF-VAE, with comparable recommendation performance. The results further validate the effectiveness of the weakly supervised learning and adversarial training modules of PSF-VAE to promote PS-Fairness in recommendations.

4.5 Sensitivity Analysis

To answer **RQ 3**, we vary the mask rate of users with known unfair items in the simulated datasets, i.e., $1-c_r$, and plot the relations with recommendation performance and fairness in Fig. 7. From Fig. 7 we can find that, the fairness of PSF-VAE generally improves with the increase of c_r, with slight negative influences on recommendation performance. This indicates that although PSF-VAE can perform well with small c_r, encouraging more users to provide feedback on unfair items can further promote PS-Fairness in recommendations.



(a) Setting: $K = 200, K_b = 25$

(b) Setting: $K = 200, K_b = 50$

Figure 7: Sensitivity of PSF-VAE with different percentages of users with observed unfair items in AM-VG, ML-1M datasets.

5 RELATED WORK

Fair RSs. Traditional fair RSs mainly rely on statistic parity to ensure the fairness of recommendations for users, with metrics such as demographical parity, equalized odds, etc. [5, 15, 45, 57, 58]. However, recent research indicates that the statistical discrepancy between the outcomes of different user groups may be well explained by some important non-sensitive factors [20, 49, 50], and algorithms that indiscriminately enforce statistical parity may still be biased against certain user groups or individuals [22, 31].

Causal RSs. Through a causal lens [6, 32, 56], user-oriented unfairness can be viewed as a non-confounder-induced bias due to the undesirable causal effects of sensitive features on observed user ratings [47, 55]. Existing causality-aware fair RSs treat all causal effects of sensitive features as unfair and remove them indiscriminately [30]. In contrast, PSF-RS preserves the fair influences of sensitive features on recommendations by identifying the fair and unfair latent mediators of sensitive features, where fairness can be achieved with the diversity of recommendations maximally preserved.

6 CONCLUSIONS

In this paper, we propose a path-specific fair recommender system (PSF-RS) to address the unfairness in recommendations while maximally preserving the fair influences of sensitive features on user interest. Specifically, PSF-RS summarizes all fair and unfair correlations between sensitive features and observed user ratings into two latent proxy mediators, which can be disentangled with weakly supervised variational inference based on the extremely sparse observed unfair items. To address the bias, we minimally alter the biased factual world into a hypothetically fair world, where a fair RS is learned accordingly by solving a constrained optimization problem. Extensive experiments show the effectiveness of PSF-RS.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation (NSF) under grants (IIS-2006844, IIS-2144209, IIS-2223769, CNS-2154962, and BCS-2228534), the Commonwealth Cyber Initiative Awards (VV-1Q23-007 and HV-2Q23-003), the JP Morgan Chase Faculty Research Award, the Cisco Faculty Research Award, the Jefferson Lab Subcontract 23-D0163, the UVA 3Cavaliers Seed Grant, and the 4-VA Collaborative Research Grant.

REFERENCES

[1] Jeffrey M Albert, Cuiyu Geng, and Suchitra Nelson. 2016. Causal mediation analysis with a latent mediator. *Biometrical Journal* 58, 3 (2016), 535–548.

[2] Alexis Bellot and Mihaela van der Schaar. 2019. Conditional independence testing using generative adversarial networks. In *NeurIPS*, Vol. 32.

[3] Peter J Bickel, Eugene A Hammel, and J William O'Connell. 1975. Sex bias in graduate admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science* 187, 4175 (1975), 398–404.

[4] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* 112, 518 (2017), 859–877.

[5] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *ICDMW*. 13–18.

[6] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2020. Bias and debias in recommender system: A survey and future directions. *arXiv preprint* (2020).

[7] Lu Cheng, Ruocheng Guo, and Huan Liu. 2022. Causal mediation analysis with hidden confounders. In *WSDM*. 113–122.

[8] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *AAAI*, Vol. 33. 7801–7808.

[9] Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li. 2023. Fairness in graph mining: A survey. *IEEE TKDE* (2023).

[10] Yingqiang Ge, Shuchang Liu, Zuohui Fu, Juntao Tan, Zelong Li, Shuyuan Xu, Yunqi Li, Yikun Xian, and Yongfeng Zhang. 2022. A survey on trustworthy recommender systems. *arXiv preprint arXiv:2207.12515* (2022).

[11] Xue Geng, Hanwang Zhang, Jingwen Bian, and Tat-Seng Chua. 2015. Learning image and user features for recommendation in social networks. In *ICCV*. 4274–4282.

[12] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search and recommendation systems with application to LinkedIn talent search. In *SIGKDD*. 2221–2231.

[13] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[15] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *NeurIPS*, Vol. 29.

[16] F Maxwell Harper and Joseph A Konstan. 2015. The MovieLens datasets: History and context. *ACM TIIS* 5, 4 (2015), 1–19.

[17] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *WWW*. 173–182.

[18] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *ICDM*. 263–272.

[19] Kosuke Imai, Luke Keele, and Dustin Tingley. 2010. A general approach to causal mediation analysis. *Psychological Methods* 15, 4 (2010), 309.

[20] Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. 2019. Fairness in algorithmic decision making: An excursion through the lens of causality. In *WWW*. 2907–2914.

[21] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *NeurIPS*.

[22] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *NeurIPS*.

[23] G Roshan Lal, Sahin Cem Geyik, and Krishnaram Kenthapadi. 2020. Fairness-aware online personalization. *arXiv preprint arXiv:2007.15270* (2020).

[24] Xiaopeng Li and James She. 2017. Collaborative variational autoencoder for recommender systems. In *SIGKDD*. 305–314.

[25] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented fairness in recommendation. In *WWW*. 624–632.

[26] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2021. Towards personalized fairness based on causal notion. In *SIGIR*. 1054–1063.

[27] Yunqi Li, Dingxian Wang, Hanxiong Chen, and Yongfeng Zhang. 2023. Transferable fairness for cold-start recommendation. *arXiv preprint arXiv:2301.10665* (2023).

[28] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *WWW*. 689–698.

[29] Dugang Liu, Pengxiang Cheng, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2020. A general knowledge distillation framework for counterfactual recommendation via uniform data. In *SIGIR*. 831–840.

[30] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S Zemel. 2016. The variational fair autoencoder. In *ICLR*.

[31] Jing Ma, Ruocheng Guo, Mengting Wan, Longqi Yang, Aidong Zhang, and Jundong Li. 2022. Learning fair node representations with graph counterfactual fairness. In *WSDM*. 695–703.

[32] Jing Ma, Ruocheng Guo, Aidong Zhang, and Jundong Li. 2021. Multi-cause effect estimation with disentangled confounder representation. In *IJCAI*. 2790–2796.

[33] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *SIGIR*. 43–52.

[34] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM CSUR* 54, 6 (2021), 1–35.

[35] Andriy Mnih and Russ R Salakhutdinov. 2007. Probabilistic matrix factorization. In *NeurIPS*.

[36] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *AAAI*, Vol. 32.

[37] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[38] Bo Pang, Min Yang, and Chongjun Wang. 2019. A novel top-N recommendation approach based on conditional variational auto-encoder. In *PAKDD*. 357–368.

[39] Judea Pearl. 2009. *Causality*. Cambridge university press.

[40] Xubin Ren, Lianghao Xia, Jiashu Zhao, Dawei Yin, and Chao Huang. 2023. Disentangled contrastive collaborative filtering. In *SIGIR*.

[41] Donald B Rubin. 1980. Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* 75, 371 (1980), 591–593.

[42] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. 2018. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval* 7, 2 (2018), 95–116.

[43] Christina Wadsworth, Francesca Vera, and Chris Piech. 2018. Achieving fairness through adversarial learning: An application to recidivism prediction. *arXiv preprint arXiv:1807.00199* (2018).

[44] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2022. A survey on the fairness of recommender systems. *JACM* (2022).

[45] Tianxin Wei and Jingrui He. 2022. Comprehensive fair meta-learned recommender system. In *ACM SIGKDD*. 1989–1999.

[46] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. 2019. PC-fairness: A unified framework for measuring causality-based fairness. In *NeurIPS*.

[47] Shuyuan Xu, Jianchao Ji, Yunqi Li, Yingqiang Ge, Juntao Tan, and Yongfeng Zhang. 2023. Causal inference for recommendation: Foundations, methods and applications. *arXiv preprint arXiv:2301.04016* (2023).

[48] Jing Yi, Yaochen Zhu, Jiayi Xie, and Zhenzhong Chen. 2021. Cross-modal variational auto-encoder for content-based micro-video background music recommendation. *IEEE TMM* (2021).

[49] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making—The causal explanation formula. In *AAAI*, Vol. 32.

[50] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. A causal framework for discovering and removing direct and indirect discrimination. *arXiv preprint arXiv:1611.07509* (2016).

[51] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM CSUR* 52, 1 (2019), 1–38.

[52] Xinyang Zhang, Shouling Ji, and Ting Wang. 2018. Differentially private releasing via deep generative model. *arXiv preprint arXiv:1801.01594* (2018).

[53] Yaochen Zhu and Zhenzhong Chen. 2022. Mutually-regularized dual collaborative variational auto-encoder for recommendation systems. In *WWW*. 2379–2387.

[54] Yaochen Zhu and Zhenzhong Chen. 2023. Variational bandwidth auto-encoder for hybrid recommender systems. *IEEE TKDE* 35, 5 (2023), 5371–5385.

[55] Yaochen Zhu, Jing Ma, and Jundong Li. 2023. Causal inference in recommender systems: A survey of strategies for bias mitigation, explanation, and generalization. *arXiv preprint arXiv:2301.00910* (2023).

[56] Yaochen Zhu, Jing Yi, Jiayi Xie, and Zhenzhong Chen. 2022. Deep causal reasoning for recommendations. *arXiv preprint arXiv:2201.02088* (2022).

[57] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-aware tensor-based recommendation. In *CIKM*. 1153–1162.

[58] Ziwei Zhu, Jingu Kim, Trung Nguyen, Aish Fenton, and James Caverlee. 2021. Fairness among new items in cold start recommender systems. In *SIGIR*. 767–776.

A DEFINITION OF CAUSAL CONCEPTS

Causal Graph. A causal graph $G = (\mathcal{V}, \mathcal{E})$ is a directed acyclic graph that describes the causal relationships among the variables of interests, where $\mathcal{V}$ is the set of nodes (which represent random variables in this paper), and $\mathcal{E}$ is the set of edges, respectively. Specifically, a directed edge from variable X to variable Y indicates that X has a causal influence on Y.

Structural Equations. Each causal graph $G = (\mathcal{V}, \mathcal{E})$ can be associated with a set of structural equations $\mathcal{F} = \{p(X|Pa(X)) \mid X \in \mathcal{V}\}$, where $p(X|Pa(X))$ quantifies the causal influence of the parents nodes of X, i.e., $Pa(X)$, on X.

Causal Path. A causal path P between variables X and Y is a sequence of edges (from X to Y) in $\mathcal{E}$ such that each edge starts with the node that ends the previous edge. A directed causal path is a causal path whose edges point in the same direction.

Mediator/Mediate. In a directed causal path P between X and Y, e.g., $X \to M \to Y$, any intermediate node M is a mediator, where the causal effects of X on Y are mediated by M.

Block/Unblock. If conditioning on $M = m$ blocks the causal path P between X and Y, no dependence (both causal and non-causal ones) can be passed from X to Y along the path P when M is known (see [13] for a formal definition). Otherwise, we say that conditioning on $M = m$ unblocks the causal path P.

Intervention. Given a causal graph G, we can conduct interventions on a variable X, which means that we set X to a value x regardless of its observed values as well as the values of its parents $Pa(X)$. If unspecified, the intervention is conducted upon the whole population, but we can also conduct the intervention conditional on $C = c$, which means that we set $X = x$ on the sub-population specified by the conditions.

Potential Outcome. Potential outcomes can be used to formalize the definition of interventions. Specifically, we define the potential outcome $Y_{X\leftarrow x}(i)$ as the value of Y for unit i had X been x. Based on $Y_{X\leftarrow x}(i)$, we can further define the potential outcome **random variable** $Y_{X\leftarrow x}$ to denote the unconditional intervention that set $X = x$ uniformly upon the population. Furthermore, the conditional potential outcome random variable $Y_{X\leftarrow x}|C = c$ can be used to denote the intervention conducted upon the sub-population specified by the condition $C = c$.

Counterfactuals. For $Y_{X\leftarrow x}|C = c$, when $C = X$ and $c = x'$, the conditional potential random variable $Y_{X\leftarrow x}|X = x'$ can be used to define the counterfactual distribution of Y had X for the units with the factual value of $X = x'$ been set to a counterfactual value x. The above analysis also applies to the **Nested Potential Outcome** introduced in Definition 2.1.

B THEORETICAL ANALYSIS

B.1 Proof of Identification of PS-Bias in Eq. (3)

Assumption 2. ***Sequential Ignorability [19].***

Step 1. *We assume that given X, the sensitive features S are ignorable for the mediators U_f, U_b and user ratings R as follows:*

$$U_{f,S\leftarrow\mathbf{s}}, U_{b,S\leftarrow\mathbf{s}'},\ R_{S\leftarrow\mathbf{s}',U_f\leftarrow\mathbf{u}_f,U_b\leftarrow\mathbf{u}'_b} \perp\!\!\!\perp S|X. \tag{15}$$

Step 2. *We also assume that given X, the post-interventional mediators $U_{f,S\leftarrow\mathbf{s}}$, $U_{b,S\leftarrow\mathbf{s}'}$ are ignorable for the user ratings R as follows:*

$$R_{S\leftarrow\mathbf{s}',U_f\leftarrow\mathbf{u}_f,U_b\leftarrow\mathbf{u}'_b} \perp\!\!\!\perp U_{f,S\leftarrow\mathbf{s}}, U_{b,S\leftarrow\mathbf{s}'}|X. \tag{16}$$

The difference between the potential outcome $R_{S\leftarrow\mathbf{s}',U_f\leftarrow\mathbf{u}_f,U_b\leftarrow\mathbf{u}'_b}$ and the nested potential outcome $R_{S\leftarrow\mathbf{s}'}(U_{f,S\leftarrow\mathbf{s}}, U_{b,S\leftarrow\mathbf{s}'})$ lies in the fact that the former directly sets the mediators U_f and U_b to the values $\mathbf{u}_f$ and $\mathbf{u}'_b$, whereas the latter conducts interventions on S by setting S to $\mathbf{s}$ and $\mathbf{s}'$ and let them influence U_f and U_b.

The sequential ignorability assumption holds for the causal graph specified in Fig. 1, because there are no unobserved confounders for the causal paths $S \to U_f$, $S \to U_b$ and $S \to R$ (and thus Eq. (15) holds) and $U_f \to R$ and $U_b \to R$ (and thus Eq. (16) holds).

B.1.1 ***Proof.*** Based on the sequential ignorability assumption defined above, Eq. (4) can be proved with six steps as follows:

$$\begin{aligned}
&\mathbb{E}\left[R_{S\leftarrow\mathbf{s}'}\left(U_{f,S\leftarrow\mathbf{s}}, U_{b,S\leftarrow\mathbf{s}'}\right) = \mathbf{r}\middle|X = \mathbf{x}, S = \mathbf{s}\right] \\
\overset{(a)}{=} &\int_{\mathbf{r},\mathbf{u}_f,\mathbf{u}'_b} p\left(R_{S\leftarrow\mathbf{s}'}\left(U_{f,S\leftarrow\mathbf{s}}, U_{b,S\leftarrow\mathbf{s}'}\right) = \mathbf{r}\middle|X = \mathbf{x}, S = \mathbf{s}, U_{f,S\leftarrow\mathbf{s}} = \mathbf{u}_f,\right. \\
&\qquad \left. U_{b,S\leftarrow\mathbf{s}'} = \mathbf{u}'_b\right) \cdot p\left(U_{f,S\leftarrow\mathbf{s}} = \mathbf{u}_f\middle|X = \mathbf{x}, S = \mathbf{s}\right) \cdot \\
&\qquad p\left(U_{b,S\leftarrow\mathbf{s}'} = \mathbf{u}'_b\middle|X = \mathbf{x}, S = \mathbf{s}\right) \cdot \mathbf{r} \\
\overset{(b)}{=} &\int_{\mathbf{r},\mathbf{u}_f,\mathbf{u}'_b} p\left(R_{S\leftarrow\mathbf{s}',U_f\leftarrow\mathbf{u}_f,U_b\leftarrow\mathbf{u}'_b} = \mathbf{r}\middle|X = \mathbf{x}, S = \mathbf{s}, U_{f,S\leftarrow\mathbf{s}} = \mathbf{u}_f,\right. \\
&\qquad \left. U_{b,S\leftarrow\mathbf{s}'} = \mathbf{u}'_b\right) \cdot p\left(U_{f,S\leftarrow\mathbf{s}} = \mathbf{u}_f\middle|X = \mathbf{x}, S = \mathbf{s}\right) \cdot \\
&\qquad p\left(U_{b,S\leftarrow\mathbf{s}'} = \mathbf{u}'_b\middle|X = \mathbf{x}, S = \mathbf{s}\right) \cdot \mathbf{r} \\
\overset{(c)}{=} &\int_{\mathbf{r},\mathbf{u}_f,\mathbf{u}'_b} p\left(R_{S\leftarrow\mathbf{s}',U_f\leftarrow\mathbf{u}_f,U_b\leftarrow\mathbf{u}'_b} = \mathbf{r}\middle|X = \mathbf{x}, S = \mathbf{s}\right) \cdot \mathbf{r} \cdot \\
&\qquad p\left(U_{f,S\leftarrow\mathbf{s}} = \mathbf{u}_f\middle|X = \mathbf{x}, S = \mathbf{s}\right) \cdot p\left(U_{b,S\leftarrow\mathbf{s}'} = \mathbf{u}'_b\middle|X = \mathbf{x}, S = \mathbf{s}\right) \\
\overset{(d)}{=} &\int_{\mathbf{r},\mathbf{u}_f,\mathbf{u}'_b} p\left(R_{S\leftarrow\mathbf{s}',U_f\leftarrow\mathbf{u}_f,U_b\leftarrow\mathbf{u}'_b} = \mathbf{r}\middle|X = \mathbf{x}\right) \cdot \\
&\qquad p\left(U_{f,S\leftarrow\mathbf{s}} = \mathbf{u}_f\middle|X = \mathbf{x}\right) \cdot p\left(U_{b,S\leftarrow\mathbf{s}'} = \mathbf{u}'_b\middle|X = \mathbf{x}\right) \cdot \mathbf{r} \\
\overset{(e)}{=} &\int_{\mathbf{r},\mathbf{u}_f,\mathbf{u}'_b} p(\mathbf{r}|\mathbf{u}_f, \mathbf{u}'_b, \mathbf{s}', \mathbf{x}) \cdot p(\mathbf{u}_f|\mathbf{s}, \mathbf{x}) \cdot p(\mathbf{u}'_b|\mathbf{s}', \mathbf{x}) \cdot p(\mathbf{x}) \cdot \mathbf{r} \\
\overset{(f)}{=} &\int_{\mathbf{r},\mathbf{u}_f,\mathbf{u}'_b} p(\mathbf{r}|\mathbf{u}_f, \mathbf{u}'_b(\mathbf{s}')) \cdot p(\mathbf{u}_f|\mathbf{s}, \mathbf{x}) \cdot p(\mathbf{u}'_b|\mathbf{s}') \cdot p(\mathbf{x}) \cdot \mathbf{r}.
\end{aligned} \tag{17}$$

Step (a) is based on the total probability theory; step (b) is based on the consistency rule of counterfactuals [41]; step (c) is based on the second step of sequential ignorability; steps (d)(e) are based on the first step of sequential ignorability; and step (f) is based on the conditional independence assumptions implied by the causal graph in Fig. 1. Similar procedures can be used to prove the identification of Eq. (5), where Eq. (3) can be calculated as Eq. (4) - Eq. (5).

B.2 PS-Bias for RS Models with Constraints

In section 2.3, we have introduced the PS-Bias of the naive RSs that predict new ratings according to the exact causal mechanism that generates the biased observed ratings. This section generalizes the PS-Bias for RS models with extra constraints, which serves as the basis for proving the PS-Bias for existing fair RSs and PSF-RS.

We note that the causal mechanism that generates the observed ratings is composed of three structural equations: $\mathcal{F} = \{p(R|U_f, U_b), p(U_f|S, X), p(U_b|S)\}$, which induces the causal graph in Fig. 1 by setting the variables on the RHS of $p \in \mathcal{F}$ as the parents and the variable on the LHS as the child. An RS model with extra constraints **can be viewed as** generating ratings in two steps: **(1)** Certain structural equations p in $\mathcal{F}$ are *minimally* changed to p_{model} according to the constraints (where the irrelevant ones remain intact). We use $\mathcal{F}_{model}$ to denote the new set of structural equations, which induces a new causal graph (e.g., Figs. 3-(b) and 4-(b)). **(2)** Ratings are generated according to the newly-induced causal model. Therefore, PS-Bias for an RS with constraints can be calculated as the path-specific effects of sensitive features S on ratings R along the unfair paths of the **newly-induced causal model**.

B.3 Proof of Zero PS-Bias for Existing Fair RSs

B.3.1 ***Further Analysis.*** Existing fair RSs constrain the user latent variables U to be independent of the user sensitive features S as Eq. (6). To satisfy such a constraint, we need to change at least two structural equations in $\mathcal{F}$, i.e., $p(U_f|S, X)$, $p(U_b|S)$ into $p_{ef}(U_f|X)$, $p_{ef}(U_b)$ (although in practice, when maximizing the likelihood of observed ratings, $p(R|U_f, U_b)$ will also be changed into $p_{ef}(R|U_f, U_b)$ since the distributions of U_f, U_b are altered), where the causal structure $p(U_f|S, X)$ necessary for recommendation diversity is inevitably lost. We use $PSBias^{**}(\mathbf{x}, \mathbf{s}, \mathbf{s}')$ to denote the PS-Bias of the altered causal model induced by existing fair RSs.

B.3.2 ***Proof.*** $PSBias^{**}(\mathbf{x}, \mathbf{s}, \mathbf{s}')$ can be calculated by substituting the three p_{ef} terms introduced above for the p terms in Eq. (3). After the substitution, the first expectation term becomes

$$\begin{aligned}&\mathbb{E}^{**}\left[R_{S\leftarrow\mathbf{s}'}\left(U_{f,S\leftarrow\mathbf{s}}, U_{b,S\leftarrow\mathbf{s}'}\right)\middle|X=\mathbf{x}, S=\mathbf{s}\right]\\&\quad=\int_{\mathbf{r},\mathbf{u}_f,\mathbf{u}_b'} p_{ef}(\mathbf{r}|\mathbf{u}_f, \mathbf{u}_b')\cdot p_{ef}(\mathbf{u}_f|\mathbf{x})\cdot p_{ef}(\mathbf{u}_b')\cdot\mathbf{r}.\end{aligned}\tag{18}$$

Similarly, the second expectation term becomes

$$\begin{aligned}&\mathbb{E}^{**}\left[R_{S\leftarrow\mathbf{s}}\left(U_{f,S\leftarrow\mathbf{s}}, U_{b,S\leftarrow\mathbf{s}}\right)=\mathbf{r}\middle|X=\mathbf{x}, S=\mathbf{s}\right]\\&\quad=\int_{\mathbf{r},\mathbf{u}_f,\mathbf{u}_b} p_{ef}(\mathbf{r}|\mathbf{u}_f, \mathbf{u}_b)\cdot p_{ef}(\mathbf{u}_f|\mathbf{x})\cdot p_{ef}(\mathbf{u}_b)\cdot\mathbf{r}.\end{aligned}\tag{19}$$

Since $PSBias^{**}(\mathbf{x}, \mathbf{s}, \mathbf{s}')$ = Eq. (18) - Eq. (19), the equality of Eqs. (18) and (19) proves that $PSBias^{**}(\mathbf{x}, \mathbf{s}, \mathbf{s}') = 0$ for existing fair RSs.

B.4 Proof of Zero PS-Bias for PSF-RS

In the hypothetically fair world induced by the proposed PSF-RS, $p_{psf}(R|U_f)$ is substituted for $p(R|U_f, U_b)$ in $\mathcal{F}$ while other causal mechanisms invariant to the RS remain unchanged. Similarly, the first expectation term in $PSBias^{*}(\mathbf{x}, \mathbf{s}, \mathbf{s}')$ can be calculated as

$$\begin{aligned}&\mathbb{E}^{*}\left[R_{S\leftarrow\mathbf{s}'}\left(U_{f,S\leftarrow\mathbf{s}}, U_{b,S\leftarrow\mathbf{s}'}\right)\middle|X=\mathbf{x}, S=\mathbf{s}\right]\\&\quad=\int_{\mathbf{r},\mathbf{u}_f,\mathbf{u}_b'} p_{psf}(\mathbf{r}|\mathbf{u}_f)\cdot p(\mathbf{u}_f|\mathbf{s}, \mathbf{x})\cdot p(\mathbf{u}_b'|\mathbf{s}')\cdot\mathbf{r}\\&\quad=\int_{\mathbf{u}_b'} p(\mathbf{u}_b'|\mathbf{s}')\int_{\mathbf{r},\mathbf{u}_f} p_{psf}(\mathbf{r}|\mathbf{u}_f)\cdot p(\mathbf{u}_f|\mathbf{s}, \mathbf{x})\cdot\mathbf{r}\\&\quad=\int_{\mathbf{r},\mathbf{u}_f} p_{psf}(\mathbf{r}|\mathbf{u}_f)\cdot p(\mathbf{u}_f|\mathbf{s}, \mathbf{x})\cdot\mathbf{r}.\end{aligned}\tag{20}$$

Furthermore, the second expectation term becomes

$$\begin{aligned}&\mathbb{E}^{*}\left[R_{S\leftarrow\mathbf{s}}\left(U_{f,S\leftarrow\mathbf{s}}, U_{b,S\leftarrow\mathbf{s}}\right)=\mathbf{r}\middle|X=\mathbf{x}, S=\mathbf{s}\right]\\&\quad=\int_{\mathbf{r},\mathbf{u}_f,\mathbf{u}_b} p_{psf}(\mathbf{r}|\mathbf{u}_f)\cdot p(\mathbf{u}_f|\mathbf{s}, \mathbf{x})\cdot p(\mathbf{u}_b|\mathbf{s})\cdot\mathbf{r}\\&\quad=\int_{\mathbf{r},\mathbf{u}_f} p_{psf}(\mathbf{r}|\mathbf{u}_f)\cdot p(\mathbf{u}_f|\mathbf{s}, \mathbf{x})\cdot\mathbf{r}.\end{aligned}\tag{21}$$

Since $PSBias^{*}(\mathbf{x}, \mathbf{s}, \mathbf{s}')$ = Eq. (20) - Eq. (21), the equality of the RHS of Eqs. (20) and (21) proves that $PSBias^{*}(\mathbf{x}, \mathbf{s}, \mathbf{s}') = 0$ for PSF-RS.

B.5 Proof of ELBO for PSF-VAE

In this section, we prove the ELBO of PSF-VAE in Eq. (11) as follows:

$$\begin{aligned}&\ln p_\theta(R,\ R_b|S, X) = \ln\int_{U_f,U_b} p_\theta(R, R_b, U_f, U_b|S, X)\\&\quad=\ln\int_{U_f,U_b} q_\phi(U_f, U_b|R, S, X)\cdot\frac{p_\theta(R, R_b, U_f, U_b|S, X)}{q_\phi(U_f, U_b|R, S, X)}\\&\quad\overset{(a)}{\geq}\int_{U_f,U_b} q_\phi(U_f, U_b|R, S, X)\cdot\ln\frac{p_\theta(R, R_b, U_f, U_b|S, X)}{q_\phi(U_f, U_b|R, S, X)}\\&\quad=\int_{U_f,U_b} q_\phi(U_f, U_b|R, S, X)\cdot\ln\frac{p_\theta(U_f, U_b|S, X)\cdot p_\theta(R, R_b|U_f, U_b)}{q_\phi(U_f, U_b|R, S, X)}\\&\quad=\mathbb{E}_{q_\phi(U_f,U_b|R,S,X)}\left[\ln\frac{p_\theta(U_f, U_b|S, X)\cdot p_\theta(R, R_b|U_f, U_b)}{q_\phi(U_f, U_b|R, S, X)}\right]\\&\quad=\mathbb{E}_{q_\phi(U_f,U_b|R,S,X)}\left[\ln p_\theta(R, R_b|U_f, U_b)\right]+\\&\qquad\mathbb{E}_{q_\phi(U_f,U_b|R,S,X)}\left[\frac{p_\theta(U_f, U_b|S, X)}{q_\phi(U_f, U_b|R, S, X)}\right]\\&\quad=\mathbb{E}_{q_\phi(U_f,U_b|R,S,X)}\left[\ln p_\theta(R|U_f, U_b)\right]+\mathbb{E}_{q_\phi(U_b|R,S)}\left[\ln p_\theta(R_b|U_b)\right]\\&\quad-\mathbb{KL}[q_\phi(U_f|R, S, X)\ ||\ p_\theta(U_f|S, X)]-\mathbb{KL}[q_\phi(U_b|R, S)\ ||\ p_\theta(U_b|S)],\end{aligned}\tag{22}$$

where step (a) is the application of Jensen's inequality, and the final step is based on the conditional independence assumptions implied by the causal graph in Fig. 1, which leads to the ELBO in Eq. (11).

We can further show that the difference between the ELBO and the log evidence $\ln p_\theta(R, R_b|S, X)$ is exactly the KL-divergence between variational posterior $q_\phi(U_f, U_b|R, S, X) = q_\phi(U_f|R, S, X)\times q_\phi(U_b|R, S)$ and the true posterior $p_\theta(U_f, U_b|R, R_b, S, X)$. To prove this, we can add the KL term to the RHS of (a) in Eq. (22) as follows:

$$\begin{aligned}&(a)+\mathbb{KL}[q_\phi(U_f, U_b|R, S, X)\ ||\ p_\theta(U_f, U_b|R, R_b, S, X)]\\&=\mathbb{E}_{q_\phi(U_f,U_b|R,S,X)}\left[\ln\frac{p_\theta(R, R_b, U_f, U_b|S, X)}{q_\phi(U_f, U_b|R, S, X)}\cdot\frac{q_\phi(U_f, U_b|R, S, X)}{p_\theta(U_f, U_b|R, R_b, S, X)}\right]\\&=\mathbb{E}_{q_\phi(U_f,U_b|R,S,X)}\left[\ln\frac{p_\theta(R, R_b, U_f, U_b|S, X)}{p_\theta(U_f, U_b|R, R_b, S, X)}\right]\\&=\mathbb{E}_{q_\phi(U_f,U_b|R,S,X)}\left[\ln p_\theta(R, R_b|S, X)\right]=\ln p_\theta(R, R_b|S, X),\end{aligned}\tag{23}$$

where the RHS of Eq. (23) is the log evidence $\ln p_\theta(R, R_b|S, X)$. This further proves our claim that minimizing the KL divergence between the variational posteriors defined by PSF-VAE and the true posteriors is equivalent to maximizing the ELBO as Eq. (11).