Fair Representation Learning for Recommendation: A Mutual Information Perspective

Chen Zhao[1], **Le Wu**[1,2,*], **Pengyang Shao**[1], **Kun Zhang**[1], **Richang Hong**[1,2], **Meng Wang**[1,2]

[1] School of Computer Science and Information Engineering, Hefei University of Technology
[2] Hefei Comprehensive National Science Center
{zhaochen.hfut.lmc, lewu.ustc, shaopymark, zhang1028kun, hongrc.hfut, eric.mengwang}@gmail.com

Abstract

Recommender systems have been widely used in recent years. By exploiting historical user-item interactions, recommender systems can model personalized potential interests of users and have been widely applied to a wide range of scenarios. Despite the impressive performance, most of them may be subject to unwanted biases related to sensitive attributes (e.g., race and gender), leading to unfairness. An intuitive idea to alleviate this problem is to ensure that there is no mutual information between recommendation results and sensitive attributes. However, keeping independence conditions solely achieves fairness improvement while causing an obvious degradation of recommendation accuracy, which is not a desired result. To this end, in this paper, we re-define recommendation fairness with a novel two-fold mutual information objective. In concerned details, we define fairness as mutual information minimization between embeddings and sensitive information, and mutual information maximization between embeddings and non-sensitive information. Then, a flexible **Fair M**utual **I**nformation (**FairMI**) framework is designed to achieve this goal. FairMI first employs a sensitive attribute encoder to capture sensitive information in the data. Then, based on results from the sensitive attribute encoder, an interest encoder is developed to generate sensitive-free embeddings, which is expected to contain rich non-sensitive information of input data. Moreover, we propose novel mutual information (upper/lower) bounds with contrastive information estimation for model optimization. Extensive experiments over two real-world datasets demonstrate the effectiveness of our proposed FairMI in reducing unfairness and improving recommendation accuracy simultaneously.

Introduction

Recommender systems have been widely applied to plenty of open platforms due to its ability of helping users explore their potential interests (Covington, Adams, and Sargin 2016; Koren, Bell, and Volinsky 2009). Plenty of algorithms have been proposed to mine users' preference to items from historical data. Among them, Collaborative Filtering (CF) is one of the representative algorithms due to its relatively-high performance and easy-to-collect user-item behavior data (Koren, Bell, and Volinsky 2009; Rendle et al. 2009; Chen et al. 2020). By learning accurate user and item embeddings from historical user-item interactions, CF-based algorithms have made impressive performance on recommendation accuracy. However, historical interactions can be biased by sensitive attributes (e.g., race and gender). As a consequence, CF-based algorithms may inherit or even amplify these data biases for embedding learning, and finally lead to unfairness in recommendation results (Ekstrand et al. 2018a; Li et al. 2021; Ekstrand et al. 2018b). For example, news recommender systems may recommend news with a clear political bias, manipulating users' options (Wu et al. 2021a; Li et al. 2022). Career recommender systems may disproportionately recommend relatively low-income jobs to female users (Lambrecht and Tucker 2019).

Recently, fairness in recommender systems has gained increasing attention (Li et al. 2022). One of promising directions is to learn fair embeddings from biased user-item interactions, which is also known as embedding fairness (Zemel et al. 2013; Madras et al. 2018). The core idea of embedding fairness is to impose constraints on the independence between learned embeddings and sensitive attributes. Following this principle, numerous works have been proposed, such as adversarial learning based debias (Bose and Hamilton 2019; Wu et al. 2021a), regularization based methods (Yao and Huang 2017). (Wu et al. 2021a) proposed to decompose users' embeddings into bias-aware embeddings and bias-free embeddings with orthogonality regularization. Further, some researchers argued that these models are not satisfactory for fair recommendation (Shao et al. 2022). Most of them assume each instance is independent, however, users and items in recommendation are not independent but tightly correlated. To solve the problem, researchers proposed to filter both user and item embeddings with the adversarial learning of a user-centric graph (Wu et al. 2021c). These models achieve better fairness, however, they also decrease recommendation accuracy. Therefore, how to avoid accuracy reduction when achieving fair recommendations becomes a key challenge for embedding fairness in CF-based recommendations.

As mutual information (MI) is a more general measure of the mutual dependence between two variables, this inspires us to quantitatively describe the recommendation fairness process with MI. In this paper, we propose a **Fair M**utual **I**nformation framework (**FairMI**) for learning fair represen-

*Corresponding Author

tations for recommendation. The proposed model consists of one sensitive attribute encoder, one interest encoder, and a novel two-fold MI based objective from the user side and the item side. Specifically, we first use a sensitive attribute encoder to generate sensitive-aware embeddings that mainly capture the sensitive information (information related to sensitive attributes). Then, based on the results from sensitive attribute encoder, an interest encoder is employed to generate sensitive-free embeddings, which are encouraged to contain as much non-sensitive information as possible (other information not related to sensitive attributes). To realize the function of these two encoders, we first leverage a sensitive attribute prediction target to optimize the sensitive attribute encoder. As for interest encoder, we design a novel two-fold MI based objective which minimizes the MI between sensitive-free embeddings and sensitive-aware embeddings, and maximizes the MI between the user-item interaction data and sensitive-free embeddings conditioned on sensitive-aware embeddings. Along this line, FairMI is able to generate fair embeddings, which preserves as much non-sensitive information as possible while removing sensitive information. Extensive experiments on two real-world datasets clearly demonstrate that FairMI achieves the best recommendation accuracy and fairness performance, compared with published baselines. The major contributions of this paper are listed as follows:

- We propose a novel two-fold MI based objective from both the user side and item side to improve recommendation fairness while avoiding accuracy decrease.
- We propose the **FairMI** framework for embedding fairness in CF-based recommendations, in which the generated sensitive-free embeddings can preserve as much non-sensitive information as possible while removing sensitive information.
- Extensive experiments on two real-world datasets clearly show that our proposed FairMI has the best performance on both recommendation accuracy and fairness.

Related Work

User Fairness in Recommender Systems

As recommender systems are data-driven, they will inevitably inherit data biases related with specific sensitive attributes (e.g., gender, age) and lead to user fairness issues in recommender systems (Ekstrand et al. 2018b). For example, career recommender systems may make decisions that favor user groups with specific sensitive attributes, leading to the deepening of career stereotypes. Therefore, how to define and quantify user fairness in recommender systems becomes a key challenge (Ekstrand et al. 2018a). Researchers have proposed several definitions for fairness (Hardt, Price, and Srebro 2016; Gunawardana and Shani 2009). For example, individual fairness refers to the fact that a model is fair if it makes similar predictions for similar individuals with different values of sensitive attributes (Bechavod, Jung, and Wu 2020). Counterfactual fairness considers fairness by mitigating differences between the factual world and counterfactual world (Kusner et al. 2017). Among all fairness definitions, group fairness has been widely studied due to its relatively-reasonable definition and wide scope of application (Zemel et al. 2013; Hardt, Price, and Srebro 2016). To achieve group fairness in recommender systems, researchers proposed many fairness-aware CF-based models, e.g., regularization based approaches (Yao and Huang 2017), adversarial learning based methods (Bose and Hamilton 2019; Wu et al. 2021c), and re-balancing technologies (Pedreshi, Ruggieri, and Turini 2008). For instance, researchers have proposed five fairness metrics as regularization terms in CF models to balance recommendation results between female users and male users (Yao and Huang 2017). And a composition of filters and adversaries has been proposed to remove correlations between sensitive attributes and recommendations based on adversarial learning (Bose and Hamilton 2019). Researchers further proposed FairGo to measure and removed unfairness in CF models from a graph based perspective (Wu et al. 2021c). However, while these models successfully mitigate unfair recommendation results to some extent, they still suffered from a substantial drop of recommendation accuracy.

Mutual Information Estimation

Mutual Information (MI) is a Shannon entropy-based measurement for the dependence between two random variable. The definition of MI between variables $\mathbf{X}$ and $\mathbf{Y}$ is

$$\mathcal{I}(\mathbf{X};\mathbf{Y}) = \mathcal{H}(X) - \mathcal{H}(X|Y), \tag{1}$$

where $\mathcal{H}(X)$ denotes the entropy of variable $\mathbf{X}$ which quantifies the amount of information to describe $\mathbf{X}$, and $\mathcal{H}(X|Y)$ denotes the conditional entropy which quantifies the amount of information to describe $\mathbf{X}$ given $\mathbf{Y}$ is known. In machine learning, MI as a criterion to encourage or limit the dependence between variables, is widely used in various task (Cheng et al. 2021; Yang et al. 2021; Shuai et al. 2022). Therefore, how to accurately estimate MI is critical. Early works proposed many non-parametric techniques to estimate MI (Paninski and Yajima 2008; Kraskov, Stögbauer, and Grassberger 2004). However, these methods did not scale to the size and dimensionality of datasets. To overcome the problem, recent works (Poole et al. 2019; Nguyen, Wainwright, and Jordan 2010; Belghazi et al. 2018; Oord, Li, and Vinyals 2018; Cheng et al. 2020) utilize the variational bounds with deep neural network to estimate MI of high dimension continuous random variables. For instance, (Oord, Li, and Vinyals 2018) proposed a MI lower bound: InfoNCE, which is derived from Noise Contrastive Estimation (NCE) (Gutmann and Hyvärinen 2010). (Cheng et al. 2020) proposed the Contrastive Log-ratio Upper Bound (CLUB), where MI is estimated by the difference of conditional probabilities between positive and negative sample pairs. It is worth mentioning that both InfoNCE and CLUB bridge MI estimation with contrastive learning (Hjelm et al. 2019).

Mutual Information based Fairness

MI is often used as a mathematical description to quantitatively analyze fair representation tasks and thus portray the trade-off between fairness and accuracy. For instance, (Creager et al. 2019) introduced flexible fair representation learning via disentanglement that decomposes information

from multiple sensitive attributes. (Moyer et al. 2018) modeled their representation learning task as an optimization objective that minimizes the MI between encoding and sensitive variables. (Song et al. 2019) proposed an information-theoretically motivated objective for learning the formulation of maximum expressiveness subject to fairness constraints. (Gupta et al. 2021) proposed to minimize the MI between representations and sensitive attributes via contrastive information estimation. In summary, these models often involve the calculation of MI, and how accurately evaluating the MI between high-dimensional continuous variables plays a crucial role.

Due to the success of MI based fairness studies, we are encouraged to employ MI to learn fair embeddings for recommendation. Note that, previous MI based fairness studies ignore the utility of non-sensitive information for the trade-off between accuracy and fairness. Instead, we utilize the two-fold mutual information objective to encourage the embeddings to capture as much non-sensitive information as possible.

The Proposed Framework

Preliminary

Before formally presenting our proposed framework, we introduce essential notions for CF models. There are usually two entity sets: user set $U(|U| = M)$ and item set $V(|V| = N)$. $\mathbf{R} \in \mathbb{R}^{M \times N}$ denotes the user-item interaction data, where M and N denote the number of users and items, respectively. For implicit feedback, if user u has interacted with item v, then $r_{uv} = 1$; otherwise $r_{uv} = 0$. Interaction data naturally form a user-item bipartite graph, formulated as $\mathcal{G} =< U \cup V, \mathbf{A} >$. The adjacency matrix can be formulated as

$$\mathbf{A} = \begin{bmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{R^T} \end{bmatrix}. \tag{2}$$

Learning high-quality user and item embeddings is the foundation of recommendation systems (Rendle et al. 2009; He et al. 2020). We use $\mathbf{E} = [\mathbf{E}_U, \mathbf{E}_V] \in \mathbb{R}^{(M+N) \times D}$ to represent learned embeddings, where D denotes the dimension of embedding. The target of recommender system is to predict the preference $\hat{r}_{uv}$ of user u to item v. Since neural graph based models can exploit the potential collaborative signal of the bipartite graph structure by aggregating the neighborhood node's embeddings (Chen et al. 2020; He et al. 2020; Wu et al. 2020), and achieve better performance, we select graph based method as the basic encoder.

The Architecture of FairMI

In order to generate fair embeddings for better recommendations, we propose a FairMI framework, which consists of one sensitive attribute encoder, one interest encoder, and a novel two-fold MI based objective. The overall architecture is illustrated in Figure 1. A basic idea is to decompose the embedding $\mathbf{e}$ into a sensitive-aware embedding $\mathbf{e}^s$ and an sensitive-free embedding $\mathbf{e}^z$. The sensitive-aware embedding have rich sensitive information that triggers unfairness, while the sensitive-free embedding contains only non-sensitive information. Next, we will introduce the technical details of each component in our proposed FairMI.

Sensitive Attribute Encoder. We expect the sensitive attribute encoder to capture the information tightly correlated with the particular sensitive attribute. Our intuition of extracting sensitive information is to build a sensitive attribute classifier. As the user-item bipartite graph structure information has been proven to contain abundant sensitive information (Wu et al. 2021c), we utilize the graph information for better classification. Specifically, we use the graph-based encoder (He et al. 2020) to extract the embeddings. The aggregation process can be formulated as

$$\begin{aligned} \mathbf{h}_v^{k+1} &= GCN(\mathbf{h}_v^k, \{\mathbf{h}_u^k : u \in \mathbf{R}_v\}), \\ \mathbf{h}_u^{k+1} &= GCN(\mathbf{h}_u^k, \{\mathbf{h}_v^k : v \in \mathbf{R}_u\}), \end{aligned} \tag{3}$$

where $\mathbf{R}_u$ and $\mathbf{R}_v$ denote neighboring nodes of user u and item v, respectively. $\mathbf{h}_u^k$ and $\mathbf{h}_v^k$ denotes the hidden node representations of user u and item v at the k-th layer (k ranges from 0 to K). The output representations of K-th layer are treated as the learned embeddings: $\mathbf{e}_u^s = \mathbf{h}_u^K, \mathbf{e}_v^s = \mathbf{h}_v^K$. Then, we apply a sensitive attribute classifier $\mathcal{S}$ to predict sensitive attributes, formulated as $\hat{a}_u = \mathcal{S}(\mathbf{e}_u^s)$, where $\hat{a}_u$ is the predicted sensitive attributes of user u, and $\mathcal{S}$ is realized by a one-layer fully-connected-network. Corresponding loss function can be formulated as:

$$\min_{\theta_\mathcal{S}, \mathbf{E}^s} \mathcal{L}_A = -\frac{1}{M} \sum_{u=1}^{M} a_u \log(\hat{a}_u), \tag{4}$$

where $\theta_\mathcal{S}$ denotes parameters of the classifier $\mathcal{S}$. After training model, we apply the GCN encoder to obtain sensitive-aware embeddings $\mathbf{e}_u^s, \mathbf{e}_v^s$ in the inference part.

Interest Encoder. Based on sensitive-aware embeddings $\mathbf{e}_u^s, \mathbf{e}_v^s$, our goal is to generate sensitive-free embeddings $\mathbf{e}_u^z, \mathbf{e}_v^z$ which has no relationships with sensitive-aware embeddings but keep as much other non-sensitive information as possible. We borrow the success of fairness-aware MI based studies (Creager et al. 2019; Song et al. 2019), and propose a novel **two-fold MI based recommendation fairness objective**. We first define $\mathbf{R}_u$ denotes user u's interacted items and $\mathbf{R}_v$ denotes the users who have interacted with item v. Then, for each user u in U, we have the following two conditions.

Condition 1: sensitive-free user embedding $\mathbf{e}_u^z$ should have no MI with sensitive-aware user embedding $\mathbf{e}_u^s$,

Condition 2: sensitive-free user embedding $\mathbf{e}_u^z$ should have maximum MI with user interactions $\mathbf{R}_u$, conditioned on sensitive-aware user embedding $\mathbf{e}_u^s$.

Similarly, for each item v in V, we have the same goal from item side.

Condition 3: sensitive-free item embedding $\mathbf{e}_v^z$ should have no MI with sensitive-aware item embedding $\mathbf{e}_v^s$,

Condition 4: sensitive-free item embedding $\mathbf{e}_v^z$ should have maximum MI with the item interactions $\mathbf{R}_v$, conditioned on sensitive-aware item embedding $\mathbf{e}_v^s$.

Condition 1&3 encourage the sensitive-free embedding not contain information related to the sensitive-aware embedding, which can be formalized as minimizing MI between

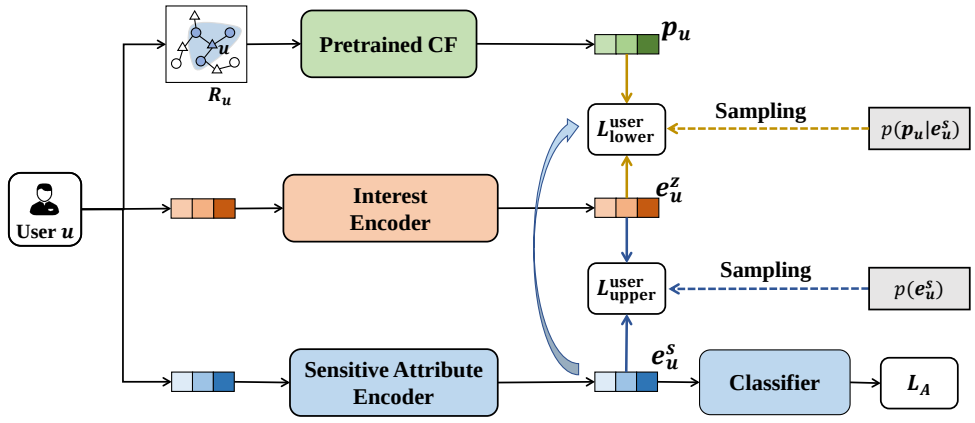

Figure 1: The architecture of our proposed FairMI. Note that, we show the details of FairMI on the user side, and the structure on the item side is similar. The only difference is that we do not train the sensitive encoder and classifier on the item side. Instead, we directly use well-trained sensitive attribute encoder on the user side to get sensitive-aware item embeddings.

sensitive-free embeddings and sensitive-aware embeddings, denoted as $\mathcal{I}(\mathbf{e}^z_u;\mathbf{e}^s_u)$ and $\mathcal{I}(\mathbf{e}^z_v;\mathbf{e}^s_v)$. Condition 2&4 encourage the sensitive-free embedding to encode as much non-sensitive information irrelevant with sensitive-aware information as possible. This can be achieved by maximizing $\mathcal{I}(\mathbf{e}^z_u;\mathbf{R}_u|\mathbf{e}^s_u)$ and $\mathcal{I}(\mathbf{e}^z_v;\mathbf{R}_v|\mathbf{e}^s_v)$. This term encourages sensitive information is not to be leaked into sensitive-free embedding, while improving the amount of non-sensitive information in embedding. In a way, this also weakens the impact of sensitive information and brings both recommendation accuracy and fairness performance improvement (Song et al. 2019; Gupta et al. 2021).

We have to note that although directly maximizing the MI $\mathcal{I}(\mathbf{e}^z_u;\mathbf{R}_u)$ and $\mathcal{I}(\mathbf{e}^z_v;\mathbf{R}_v)$ can also satisfy a similar request of Condition 2&4, this maximization ignores the fact that the information extracted from biased historical data is the combination of sensitive information and non-sensitive information. In consequence, both non-sensitive and sensitive information are enhanced, which is in conflict with minimizing $\mathcal{I}(\mathbf{e}^z,\mathbf{e}^s)$. To compare two formulations of mutual information maximization, we also set $\mathcal{I}(\mathbf{e}^z;\mathbf{R})$ as a special case of our proposed framework in experiments for better evaluation. Here, the overall loss function can be formulated as:

$$\min_{\mathbf{E}^z} \mathcal{L}_{all} = \mathcal{L}_{rec} + \mathcal{L}_{MI}, \tag{5}$$

where $\mathcal{L}_{MI}$ denotes our proposed two-fold MI based loss, which will be introduced in the following part. $\mathcal{L}_{rec}$ can be any recommendation loss for implicit feedback, e.g., BPR loss (Rendle et al. 2009):

$$\mathcal{L}_{rec} = -\sum_{u=1}^{M} \sum_{(v,k)\in\mathcal{D}_u} \ln \sigma(\mathbf{e}^{z\top}_u \mathbf{e}^z_v - \mathbf{e}^{z\top}_u \mathbf{e}^z_k), \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function; $\mathcal{D}_u = \{(v,k)|v \in \mathbf{R}_u \cap k \in V - \mathbf{R}_u\}$.

MI Bounds

As illustrated in Eq.(5), the remaining challenge is how to calculate $\mathcal{L}_{MI}$. To this end, we propose a novel two-fold MI based objective to approximate the loss, which consists of a MI upper bound for minimizing $\frac{1}{M}\sum_{u=1}^{M}\mathcal{I}(\mathbf{e}^z_u;\mathbf{e}^s_u)$ and $\frac{1}{N}\sum_{v=1}^{N}\mathcal{I}(\mathbf{e}^z_v;\mathbf{e}^s_v)$, and a MI lower bound for maximizing $\frac{1}{M}\sum_{u=1}^{M}[\mathcal{I}(\mathbf{e}^z_u;\mathbf{R}_u|\mathbf{e}^s_u)]$ and $\frac{1}{N}\sum_{v=1}^{N}[\mathcal{I}(\mathbf{e}^z_v;\mathbf{R}_v|\mathbf{e}^s_v)]$. The technical details are reported as follows.

MI Upper Bound. In order to effectively satisfy the requirement of sensitive-free embeddings should have no MI with sensitive-aware embeddings, we derive a sample-based MI upper bound in Proposition.1 based on the Contrastive Log-ratio Upper Bound (CLUB)(Cheng et al. 2020). We take the minimization of $\frac{1}{M}\sum_{u=1}^{M}\mathcal{I}(\mathbf{e}^z_u;\mathbf{e}^s_u)$ from user side as an example to introduce the technical details.

Proposition 1. *Given $\mathbf{e}^s_j \sim p(\mathbf{e}^s_u)$, if the conditional distribution $p(\mathbf{e}^s_u|\mathbf{e}^z_u)$ between $\mathbf{e}^s_u$ and $\mathbf{e}^z_u$ is known, then*

$$\mathcal{I}(\mathbf{e}^s_u;\mathbf{e}^z_u) \le \mathbb{E}\left[\log p(\mathbf{e}^s_u|\mathbf{e}^z_u) - \frac{1}{M}\sum_{j=1}^{M}\log p(\mathbf{e}^s_j|\mathbf{e}^z_u)\right]. \tag{7}$$

According to this proposition, the problem turns to calculate the conditional probability $p(\mathbf{e}^s_u|\mathbf{e}^z_u)$. Therefore, we propose to leverage a neural network $q_\phi(\mathbf{e}^s_u|\mathbf{e}^z_u)$ to approximate its value by minimizing their KL-divergence:

$$\min_{q_\phi} \mathbb{D}_{\mathrm{KL}}\left[q_\phi(\mathbf{e}^s_u|\mathbf{e}^z_u)||p(\mathbf{e}^s_u|\mathbf{e}^z_u)\right]. \tag{8}$$

Specifically, we first assume that $q_\phi(\mathbf{e}^s_u|\mathbf{e}^z_u)$ accords with the conditional Gaussian distribution, then log-likelihood maximization is leveraged to update corresponding parameters ϕ. To this end, the MI upper bound can be formulated as follows:

$$\min_{\mathbf{e}_u^z} \mathcal{L}_{\text{upper}}^{\text{user}} = \frac{1}{M}\sum_{u=1}^{M}\left[\log q_\phi(\mathbf{e}_u^s|\mathbf{e}_u^z) - \frac{1}{M}\sum_{j=1}^{M}\log q_\phi(\mathbf{e}_j^s|\mathbf{e}_u^z)\right]. \tag{9}$$

We have to note that the parameters of network $q_\phi(\cdot)$ and $\mathbf{e}^z$ are updated iteratively. Analogously, we can obtain the upper bound from item-side $\mathcal{L}_{\text{upper}}^{\text{item}}$ to minimize $\frac{1}{N}\sum_{v=1}^{N}\mathcal{I}(\mathbf{e}_v^z;\mathbf{e}_v^s)$.

MI Lower Bound. For the requirement of sensitive-free embeddings should have maximize MI with the historical interactions, we develop a novel MI lower bound to realize it. Similarly, we take the conditional MI from the user side $\frac{1}{M}\sum_{u=1}^{M}[\mathcal{I}(\mathbf{e}_u^z;\mathbf{R}_u|\mathbf{e}_u^s)]$ as an example for the technical details introduction.

Due to the high-dimension and sparsity of the user historical interactions, we leverage a pre-trained models (e.g., BPR (Rendle et al. 2009), LightGCN (He et al. 2020)) to generate low-rank embedding $\mathbf{p}_u$ to denote $\mathbf{R}_u$. Then, we employ Conditional InfoNCE (Gupta et al. 2021) to achieve this goal, formulating as the following proposition:

Proposition 2. *Given* $\mathbf{p}_u, \mathbf{e}_u^z, \mathbf{e}_u^s \sim p(\mathbf{p}_u, \mathbf{e}_u^z, \mathbf{e}_u^s)$, $\mathbf{p}_j \sim p(\mathbf{p}_u|\mathbf{e}_u^s)$*, with a score function* f*, we have*

$$\mathcal{I}(\mathbf{e}_u^z;\mathbf{p}_u|\mathbf{e}_u^s) \geq \mathbb{E}\left[\log \frac{\exp f(\mathbf{p}_u,\mathbf{e}_u^z,\mathbf{e}_u^s)}{\frac{1}{M}\sum_{j=1}^{M}\exp f(\mathbf{p}_j,\mathbf{e}_u^z,\mathbf{e}_u^s)}\right]. \tag{10}$$

To calculate this equation, two points should be considered: the score function $f(\cdot)$ and sampling strategy. For the former, we leverage weighted cosine similarity as the score function, which is similar to (Wu et al. 2021a,b):

$$f(\mathbf{p}_u,\mathbf{e}_u^z,\mathbf{e}_u^s) = \text{sim}(\mathbf{p}_u, \mathbf{e}_u^z + \alpha\cdot\mathbf{e}_u^s), \tag{11}$$

where α is the hyper-parameter to control the impact of sensitive-aware embedding. $\text{sim}(\cdot)$ is the cosine similarity function.

For the latter, directly sampling from the conditional distribution is very difficult. Thus, we propose a novel alternative method to realize this goal. Specifically, inspired by (Cheng et al. 2021), we assume the sensitive embedding $\mathbf{e}_u^s$ have the potential bias directions. For example, for the sensitive attribute "gender", there are two potential bias directions:{"*male*", "*female*"}, any sensitive embedding correlated with topic "gender" belong to one of the potential bias directions. Therefore, we develop $\pi(\mathbf{e}_i^s,\mathbf{e}_j^s)$ to measure whether two sensitive embeddings belong to the same or opposite bias direction. If $\pi(\mathbf{e}_i^s,\mathbf{e}_j^s) > 0$, $\mathbf{e}_i^s$ and $\mathbf{e}_j^s$ have the same direction, and vice versa. To this end, $\{\mathbf{p}_j \mid \pi(\mathbf{e}_i^s,\mathbf{e}_j^s) > 0\}$ can be used to realize the conditional sampling. And our proposed lower bound can be formulate as follows:

Table 1: Statistics of the datasets.

Dataset	#Users	#Items	#Interactions	Density
Movielens-1m	6,040	3,952	1,000,209	4.19%
Lastfm-360K	48,386	21,711	2,045,305	0.19%

$$\max_{\mathbf{e}_u^z} \mathcal{L}_{\text{lower}}^{\text{user}} = \frac{1}{M}\sum_{u=1}^{M}\left[\log \frac{\exp(\text{sim}(\mathbf{p}_u, w(\mathbf{e}_u^z,\mathbf{e}_u^s,\alpha)))}{\frac{1}{M}\sum_{j=1}^{M}\exp(\text{sim}(\mathbf{p}_j, w(\mathbf{e}_u^z,\mathbf{e}_u^s,\alpha)))}\right], \tag{12}$$

where $w(\mathbf{e}_u^z,\mathbf{e}_u^s,\alpha) = \mathbf{e}_u^z + \alpha\cdot\mathbf{e}_u^s$ for easy understanding. During implementation, we set $\pi(\cdot,\cdot)$ to be the Pearsons Correlation Coefficient and $\alpha = 0.1$. Similarly, the lower bound from item side can also be achieved by maximizing $\mathcal{L}_{\text{lower}}^{\text{item}}$. Finally, the two-fold MI based loss can be formulated as follows:

$$\mathcal{L}_{MI} = \beta(\mathcal{L}_{\text{upper}}^{\text{user}} + \mathcal{L}_{\text{upper}}^{\text{item}}) - \gamma(\mathcal{L}_{\text{lower}}^{\text{user}} + \mathcal{L}_{\text{lower}}^{\text{item}}). \tag{13}$$

Experiments

Experimental Settings

Datasets. We conduct experiments on two datasets: MovieLens-1M (Harper and Konstan 2015) and Lastfm-360K (Celma Herrada et al. 2009). On MovieLens-1M, We split the historical records into training set and test set with the ratio of 8:2, and 10% of the test set is used as validation. In order to turn MovieLens-1M into implicit dataset, similar to previous studies (Islam et al. 2021), we consider items with user ratings greater than 0 as positive feedback. For Lastfm-360K dataset, we first split the training sets and test sets by 7:3, and 10% of the test set is used as validation, then we remove users with less than 20 interaction records, and at the same time remove entries with less than 20 plays. In addition, we treat **gender** as the sensitive attribute. The statistics are recorded in Table 1.

Evaluation Metrics. We focus on the recommendation accuracy and fairness performance. For recommendation accuracy, we apply two widely used ranking metrics: RECALL (Gunawardana and Shani 2009) and NDCG (Järvelin and Kekäläinen 2017). Larger values of these two metrics mean better recommendation accuracy. For fairness performance, inspired by two widely-adopted group fairness metrics, i.e., *Demographic Parity (DP)* (Zemel et al. 2013) and *Equalized of Opportunity (EO)* (Hardt, Price, and Srebro 2016), we derive fairness metrics for recommendation, it estimates the group preferences to all items, formulated as:

$$\forall v \in V, f_{G_0}^v = \frac{\sum_{u\in G_0}\mathbf{I}_{v\in TopK_u}}{|G_0|}, f_{G_1}^v = \frac{\sum_{u\in G_1}\mathbf{I}_{v\in TopK_u}}{|G_1|},$$
$$\mathbf{f}_{G_0} = [f_{G_0}^1,..,f_{G_0}^v,...,f_{G_0}^N], \mathbf{f}_{G_1} = [f_{G_1}^1,..,f_{G_1}^v,...,f_{G_1}^N], \tag{14}$$

G_0 and G_1 denote the user group with different sensitive attributes, i.e. $a_u = 0$ and $a_u = 1$ respectively. $TopK_u$ is Top-K ranked items for user u. $\mathbf{I} \in \mathbb{R}^N$, if item v is in the set $TopK_u$, then $\mathbf{I}_v = 1$, otherwise $\mathbf{I}_v = 0$. Then

Table 2: Recommendation accuracy and fairness performance on MovieLens-1M.

Model	K	NDCG@K ↑			RECALL@K ↑			DP@K ↓			EO@K ↓		
		10	20	30	10	20	30	10	20	30	10	20	30
BPRMF	Base	0.1943	0.2537	0.2926	0.1437	0.2280	0.2916	0.2854	0.2572	0.2412	0.3580	0.3316	0.3122
	DP	0.1899	0.2490	0.2878	0.1409	0.2240	0.2875	0.2187	0.1870	0.1684	0.3231	0.2944	0.2793
	Adv	0.1900	0.2485	0.2866	0.1404	0.2230	0.2858	0.1684	0.1363	0.1214	0.2736	0.2499	0.2363
	FairRec	0.1896	0.2485	0.2860	0.1407	0.2236	0.2847	0.1656	0.1317	0.1191	0.2714	0.2451	0.2334
	FairMI*	**0.2022**	**0.2607**	**0.3005**	0.1487	**0.2326**	0.2957	0.1501	0.1285	0.1180	0.2406	0.2161	0.2111
	FairMI	**0.2022**	0.2606	0.2997	**0.1491**	0.2324	**0.2959**	**0.1381**	**0.1179**	**0.1110**	**0.2233**	**0.2038**	**0.1959**
LightGCN	Base	0.2025	0.2671	0.3075	0.1523	0.2449	0.3114	0.2937	0.2626	0.2452	0.3621	0.3325	0.3120
	DP	0.1981	0.2603	0.3008	0.1481	0.2363	0.3028	0.2297	0.1924	0.1745	0.3247	0.2955	0.2811
	Adv	0.1970	0.2579	0.2982	0.1474	0.2346	0.3009	0.1517	0.1183	0.1029	0.2646	0.2338	0.2216
	FairRec	0.1950	0.2561	0.2955	0.1472	0.2339	0.2986	0.1536	0.1193	0.1042	0.2590	0.2283	0.2243
	FairGo	0.1822	0.2373	0.2741	0.1336	0.2108	0.2710	0.2728	0.2436	0.2275	0.3382	0.3101	0.2921
	FairGNN	0.1964	0.2569	0.2963	0.1466	0.2323	0.2969	0.1472	0.1181	0.1045	0.2608	0.2320	0.2221
	FairMI*	**0.2128**	**0.2754**	**0.3151**	0.1581	0.2473	0.3121	0.1597	0.1340	0.1242	0.2426	0.2243	0.2151
	FairMI	**0.2128**	0.2752	0.3148	**0.1586**	**0.2477**	**0.3124**	**0.1337**	**0.1111**	**0.1004**	**0.2228**	**0.2006**	**0.1979**

Table 3: Recommendation accuracy and fairness performance on Lastfm-360K.

Model	K	NDCG@K ↑			RECALL@K ↑			DP@K ↓			EO@K ↓		
		10	20	30	10	20	30	10	20	30	10	20	30
BPRMF	Base	**0.1959**	**0.2449**	**0.2743**	**0.1564**	**0.2372**	**0.2943**	0.2664	0.2480	0.2376	0.3345	0.3201	0.3122
	DP	0.1922	0.2403	0.2693	0.1530	0.2327	0.2901	0.2230	0.2037	0.1929	0.3161	0.3008	0.2941
	Adv	0.1902	0.2384	0.2672	0.1520	0.2312	0.2874	0.1464	0.1248	0.1155	0.2702	0.2617	0.2582
	FairRec	0.1876	0.2353	0.2644	0.1502	0.2288	0.2853	0.1484	0.1279	0.1169	0.2712	0.2697	0.2592
	FairMI*	0.1931	0.2408	0.2701	0.1533	0.2322	0.2879	0.1480	0.1290	0.1230	0.2526	0.2506	0.2507
	FairMI	0.1932	0.2409	0.2700	0.1535	0.2320	0.2885	**0.1318**	**0.1210**	**0.1153**	**0.2405**	**0.2408**	**0.2430**
LightGCN	Base	0.1971	0.2463	0.2762	0.1572	**0.2381**	**0.2964**	0.2860	0.2673	0.2569	0.3508	0.3332	0.3247
	DP	0.1965	0.2447	0.2752	0.1554	0.2353	0.2943	0.2453	0.2269	0.2173	0.3297	0.3140	0.3077
	Adv	0.1898	0.2377	0.2671	0.1515	0.2303	0.2874	0.1382	0.1273	0.1163	0.2682	0.2599	0.2576
	FairRec	0.1892	0.2375	0.2627	0.1505	0.2276	0.2837	0.1397	0.1295	0.1157	0.2700	0.2607	0.2596
	FairGo	0.1693	0.2115	0.2389	0.1371	0.2065	0.2627	0.2626	0.2450	0.2338	0.3282	0.3124	0.3059
	FairGNN	0.1879	0.2358	0.2651	0.1501	0.2290	0.2859	0.1372	0.1210	0.1171	0.2690	0.2609	0.2592
	FairMI*	0.1970	0.2453	0.2749	0.1565	0.2359	0.2937	0.1352	0.1239	0.1177	0.2486	0.2482	0.2484
	FairMI	**0.1976**	**0.2464**	**0.2766**	**0.1576**	0.2373	0.2945	**0.1312**	**0.1199**	**0.1152**	**0.2402**	**0.2399**	**0.2382**

Table 4: Ablation study with K=10.

Model		NDCG@K ↑	RECALL@K ↑	DP@K ↓	EO@K ↓
BPRMF	Base	0.1943	0.1437	0.2854	0.3580
	w/o U	**0.2031**	**0.1498**	0.1757	0.2652
	w/o L	0.1887	0.1390	0.1588	0.2599
	FairMI	0.2022	0.1491	**0.1381**	**0.2233**
LightGCN	Base	0.2025	0.1523	0.2937	0.3621
	w/o U	**0.2139**	**0.1595**	0.2102	0.2870
	w/o L	0.1959	0.1470	0.1427	0.2551
	FairMI	0.2128	0.1586	**0.1337**	**0.2228**

we take Jensen–Shannon divergence ($JSD(\cdot,\cdot)$) to compare two groups:

$$DP@K = JSD(\mathbf{f}_{G_0}, \mathbf{f}_{G_1}). \tag{15}$$

We also derive *EO* which requires similar prediction results across different groups conditional on the real preferences of users, formulated as:

$$\forall v \in V, d^v_{G_0} = \frac{\sum_{u \in G_0} \mathbf{I}_{v \in \mathbf{R}^t_u \cap TopK_u}}{|G_0|}, d^v_{G_1} = \frac{\sum_{u \in G_1} \mathbf{I}_{v \in \mathbf{R}^t_u \cap TopK_u}}{|G_1|},$$
$$\mathbf{d}_{G_0} = [d^1_{G_0}, .., d^v_{G_0}, ..., d^N_{G_0}], \mathbf{d}_{G_1} = [d^1_{G_1}, .., d^v_{G_1}, ..., d^N_{G_1}], \tag{16}$$

$$EO@K = JSD(\mathbf{d}_{G_0}, \mathbf{d}_{G_1}), \tag{17}$$

where $\mathbf{R}^t_u$ denotes the items that user u clicked on in the test data. Note that, smaller values of *DP* and *EO* mean better fairness performance.

Baseline. We apply FairMI on the basis of two CF models: BPRMF (Rendle et al. 2009) and LightGCN (He et al. 2020). To verify the effectiveness of FairMI, we compare with a number of recent approaches, including regularization based methods and adversarial based methods. DP (Yao and Huang 2017) as a regularization based models, by adding different statistical fairness regularization terms to CF models. We also compare with various adversarial based methods. In particular, Adv utilizes adversarial learning to reduce relevance between sensitive attributes and embeddings (Bose and Hamilton 2019). Compared to Adv, FairGo (Wu et al. 2021c) builds a more sophisticated adversarial learning which considers the unfairness hidden in user sub-graphs. FairRec (Wu et al. 2021a) is a fairness-aware framework with decomposed adversarial learning and orthogonality regularization. FairGNN (Dai and Wang 2021) adversarially learns fair graph representations with limited sensitive attributes, they apply the sensitive attribute estimator to make up for the missing sensitive attributes to improve fairness performance, as items do not have sensitive attributes, CF scene can also be treated as limited sensitive attributes. Besides, to prove the effectiveness of conditional MI, we further propose FairMI*, a variant of FairMI. Specifically, we replace the conditional MI ($\mathcal{I}(\mathbf{e}^z_u, \mathbf{R}_u|\mathbf{e}^s_u)$ and $\mathcal{I}(\mathbf{e}^z_v, \mathbf{R}_v|\mathbf{e}^s_v)$) with ($\mathcal{I}(\mathbf{e}^z_u, \mathbf{R}_u)$ and $\mathcal{I}(\mathbf{e}^z_v, \mathbf{R}_v)$), and use the InfoNCE (Oord, Li, and Vinyals 2018) to optimize it.

Implementation Details. The experiments are imple-

mented with Pytorch-1.7.0 on 1 NVIDIA TITAN-RTX GPU. For sensitive attribute encoder, we use LightGCN as backbone and a one-layer fully-connected network as the attribute predictor. For interest encoder, we use the conditional Gaussian distribution to parameterize the $q_\phi(\mathbf{e}^s|\mathbf{e}^z) = \mathcal{N}(\mathbf{e}^s|\boldsymbol{\mu}_\phi(\mathbf{e}^z), \boldsymbol{\Sigma}_\phi(\mathbf{e}^z))$, where mean $\boldsymbol{\mu}_\phi(\cdot)$ and variance $\boldsymbol{\Sigma}_\phi(\cdot)$ are two-layer fully-connected networks with the activation function is Tanh$(\cdot)$. In order for $g_\theta(\cdot)$ to extract rich semantic information from $\mathcal{G}$, we pre-train BPRMF or LightGCN by Eq.(6) (If the interest encoder's backbone is BPRMF, we parameterize $g_\theta(\cdot)$ as it, and vice versa). We set the embedding size as $D = 64$, the mini-batch size is set to 2048 for Movielens-1M and 4096 for Lastfm-360K; and choose the Adam optimizer with the initial learning rate equaling 0.001. The weight factors β and γ of $\mathcal{L}_{\text{upper}}$ and $\mathcal{L}_{\text{lower}}$ are searched in the range of $\{10^{-4}, 10^{-3}, 10^{-2}, 5 \times 10^{-2}, 10^{-1}\}$.

Overall Performance

We have several observations from overall results (Table 2 and Table 3). Specifically, we emphasize the best results in **bold**, and underline to highlight the second best results. First, compared to the base models (BPRMF and LightGCN), FairMI effectively improves the fairness performance (DP@K and EO@K) while ensure the recommendation accuracy (NDCG and RECALL). In some case (Movielens-1M), the recommendation accuracy even exceeds the based models, this may due to the base model is limited by performance bottleneck in capturing non-sensitive information. Second, compared to various fairness-aware models (DP, Adv, FairGo, FairRec, FairGNN), FairMI achieve the best trade-off between recommendation accuracy and fairness performance on both two real-world datasets. We also find that other fairness-aware models all achieves an appreciable fairness performance, but suffer from a clear drop in recommendation performance. The reason is that these methods only focus on reducing the sensitive information in embeddings without explicitly maximizing the relations between embeddings and non-sensitive information. Finally, compared to FairMI, FairMI* has a similar recommendation accuracy, but a worse fairness performance. The results prove that by conditioning mutual information on sensitive information, the sensitive-free embedding can easily retain non-sensitive information and further reduce the effect of sensitive information.

Model Analyses

Ablation Studies. We conduct ablation studies on MovieLens-1M to verify the effectiveness of different modules. Specifically, we evaluate recommendation accuracy and fairness (under TopK=10) and select BPRMF and LightGCN as the base models. The experimental results are presented in Table.4, where "w/o L" and "w/o U" denotes FairMI without the MI lower bound (Eq.(12)) and FairMI without the MI upper bound (Eq.(9)), respectively. We have several observations Table.4. First, "w/o L" is a bit worse than FairMI on fairness performance (DP@K, EO@K), and "w/o L" causes an obvious reduction of recommendation

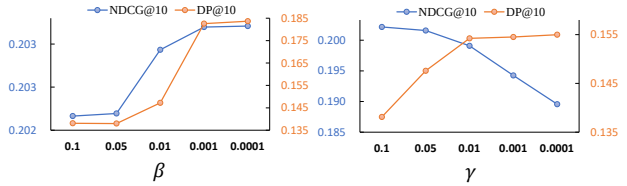

Figure 2: FairMI based on BPRMF with varying β and γ

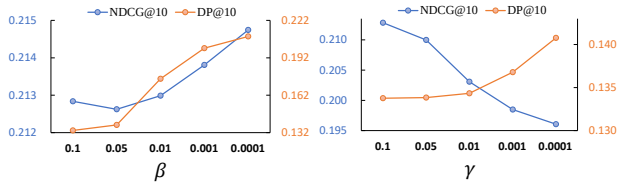

Figure 3: FairMI based on LightGCN with varying β and γ

accuracy (NDCG@K, RECALL@K). The results proves that maximizing the lower bound of conditional MI can effectively improves fairness performance, and significantly improves accuracy performances. Second, we observe that "w/o U" has slightly better performance on recommendation accuracy but heavily worse fairness performance than FairMI. This proves that minimizing MI between sensitive-free embeddings and sensitive-aware embeddings can significantly improve fairness performance while cause a little bit decrease on accuracy. Finally, FairMI achieves the best trade-off between recommendation accuracy and fairness performance. The experimental results show that the combination of the MI lower and upper bounds is a complementary process and demonstrate the need for the two-fold constraint.

Parameter Sensitivity Analysis. We conduct additional experiments on Movielens-1M to verify the impact of different β and γ. We can observe the obvious trade-off effects from Fig.2 and Fig.3. Firstly, with the reduction of β, this implies the relaxation of MI upper bound, sensitive-free embeddings will contain more sensitive information, which exacerbates the unfairness of recommendation results, but leads to a rebound in recommendation performance. Secondly, along with the reduction of γ, sensitive information will be more easily leaked from the user-item interaction data to sensitive-free embedding, leading to a reduction in fairness performance. And reduction of non-sensitive information in sensitive-free embedding leads to a decrease in recommendation accuracy.

Conclusion

In this paper, we discussed that existing studies on recommendation fairness failed to capture rich non-sensitive information, leading to an obvious decrease on recommendation accuracy. Therefore, we proposed FairMI, a novel MI based framework, which consisted of a sensitive attribute encoder to generate sensitive-aware embeddings, an interest encoder to generate sensitive-free embeddings, and a novel two-fold MI based objective to guide the optimization of embeddings. To realize the MI based goal, we further utilized novel MI upper/lower bounds to minimize/maximize MI. The extensive experiments showed the effectiveness of our proposed FairMI. In the future, we would like to extend the proposed framework to multiple sensitive attributes.

Acknowledges

This work was supported in part by grants from the National Key Research and Development Program of China (No.2021ZD0111802), the CCF-AFSG Research Fund (No.CCF-AFSG RF20210006), National Natural Science Foundation of China (No.62006066, No.61972125), and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

References

Bechavod, Y.; Jung, C.; and Wu, S. Z. 2020. Metric-Free Individual Fairness in Online Learning. *Advances in Neural Information Processing Systems*, 33: 11214–11225.

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, D. 2018. Mutual Information Neural Estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, 531–540.

Bose, A.; and Hamilton, W. 2019. Compositional Fairness Constraints for Graph Embeddings. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 715–724.

Celma Herrada, Ò.; et al. 2009. *Music recommendation and discovery in the long tail*. Universitat Pompeu Fabra.

Chen, L.; Wu, L.; Hong, R.; Zhang, K.; and Wang, M. 2020. Revisiting Graph Based Collaborative Filtering: A Linear Residual Graph Convolutional Network Approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34: 27–34.

Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; and Carin, L. 2020. CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, 1779–1788.

Cheng, P.; Hao, W.; Yuan, S.; Si, S.; and Carin, L. 2021. FairFil: Contrastive Neural Debiasing Method for Pretrained Text Encoders. In *International Conference on Learning Representations*.

Covington, P.; Adams, J.; and Sargin, E. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, 191–198.

Creager, E.; Madras, D.; Jacobsen, J.-H.; Weis, M.; Swersky, K.; Pitassi, T.; and Zemel, R. 2019. Flexibly Fair Representation Learning by Disentanglement. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 1436–1445.

Dai, E.; and Wang, S. 2021. Say No to the Discrimination: Learning Fair Graph Neural Networks with Limited Sensitive Attribute Information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 680–688.

Ekstrand, M. D.; Tian, M.; Azpiazu, I. M.; Ekstrand, J. D.; Anuyah, O.; McNeill, D.; and Pera, M. S. 2018a. All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, 172–186.

Ekstrand, M. D.; Tian, M.; Kazi, M. R. I.; Mehrpouyan, H.; and Kluver, D. 2018b. Exploring Author Gender in Book Rating and Recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, 242–250.

Gunawardana, A.; and Shani, G. 2009. A Survey of Accuracy Evaluation Metrics of Recommendation Tasks. *Journal of Machine Learning Research*, 10: 2935–2962.

Gupta, U.; Ferber, A. M.; Dilkina, B.; and Ver Steeg, G. 2021. Controllable Guarantees for Fair Outcomes via Contrastive Information Estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35: 7610–7619.

Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, 297–304.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29: 3315–3323.

Harper, F. M.; and Konstan, J. A. 2015. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5: 1–19.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 639–648.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

Islam, R.; Keya, K. N.; Zeng, Z.; Pan, S.; and Foulds, J. 2021. Debiasing Career Recommendations with Neural Fair Collaborative Filtering. In *Proceedings of the Web Conference 2021*, 3779–3790.

Järvelin, K.; and Kekäläinen, J. 2017. IR Evaluation Methods for Retrieving Highly Relevant Documents. *SIGIR Forum*, 51: 243–250.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42: 30–37.

Kraskov, A.; Stögbauer, H.; and Grassberger, P. 2004. Estimating mutual information. *Physical Review E*, 69: 066138.

Kusner, M.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4069–4079.

Lambrecht, A.; and Tucker, C. 2019. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, 65: 2966–2981.

Li, Y.; Chen, H.; Fu, Z.; Ge, Y.; and Zhang, Y. 2021. User-Oriented Fairness in Recommendation. In *Proceedings of the Web Conference 2021*, 624–632.

Li, Y.; Chen, H.; Xu, S.; Ge, Y.; Tan, J.; Liu, S.; and Zhang, Y. 2022. Fairness in Recommendation: A Survey. *arXiv preprint arXiv:2205.13619*.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning Adversarially Fair and Transferable Representations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, 3384–3393.

Moyer, D.; Gao, S.; Brekelmans, R.; Steeg, G. V.; and Galstyan, A. 2018. Invariant Representations without Adversarial Training. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 9102–9111.

Nguyen, X.; Wainwright, M. J.; and Jordan, M. I. 2010. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56: 5847–5861.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Paninski, L.; and Yajima, M. 2008. Undersmoothed kernel entropy estimators. *IEEE Transactions on Information Theory*, 54: 4384–4388.

Pedreshi, D.; Ruggieri, S.; and Turini, F. 2008. Discrimination-Aware Data Mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 560–568.

Poole, B.; Ozair, S.; Van Den Oord, A.; Alemi, A.; and Tucker, G. 2019. On Variational Bounds of Mutual Information. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 5171–5180.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 452–461.

Shao, P.; Wu, L.; Chen, L.; Zhang, K.; and Wang, M. 2022. FairCF: Fairness-aware Collaborative Filtering. *Science China Information Sciences*, 65: 127–141.

Shuai, J.; Zhang, K.; Wu, L.; Sun, P.; Hong, R.; Wang, M.; and Li, Y. 2022. A Review-Aware Graph Contrastive Learning Framework for Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1283–1293.

Song, J.; Kalluri, P.; Grover, A.; Zhao, S.; and Ermon, S. 2019. Learning Controllable Fair Representations. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, 2164–2173.

Wu, C.; Wu, F.; Wang, X.; Huang, Y.; and Xie, X. 2021a. Fairness-aware News Recommendation with Decomposed Adversarial Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35: 4462–4469.

Wu, J.; Wang, X.; Feng, F.; He, X.; Chen, L.; Lian, J.; and Xie, X. 2021b. Self-Supervised Graph Learning for Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 726–735.

Wu, L.; Chen, L.; Shao, P.; Hong, R.; Wang, X.; and Wang, M. 2021c. Learning Fair Representations for Recommendation: A Graph-Based Perspective. In *Proceedings of the Web Conference 2021*, 2198–2208.

Wu, L.; Yang, Y.; Zhang, K.; Hong, R.; Fu, Y.; and Wang, M. 2020. Joint Item Recommendation and Attribute Inference: An Adaptive Graph Convolutional Network Approach. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 679–688.

Yang, Y.; Wu, L.; Hong, R.; Zhang, K.; and Wang, M. 2021. Enhanced Graph Learning for Collaborative Filtering via Mutual Information Maximization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 71–80.

Yao, S.; and Huang, B. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2925–2934.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, 325–333.