Popularity Bias Is Not Always Evil: Disentangling Benign and Harmful Bias for Recommendation

Zihao Zhao, Jiawei Chen, Sheng Zhou, Xiangnan He, Xuezhi Cao, Fuzheng Zhang, Wei Wu

Abstract—Recommender system usually suffers from severe *popularity bias* — the collected interaction data usually exhibits quite imbalanced or even long-tailed distribution over items. Such skewed distribution may result from the users' *conformity* to the group, which deviates from reflecting users' true preference. Existing efforts for tackling this issue mainly focus on completely eliminating popularity bias. However, we argue that not all popularity bias is evil. Popularity bias not only results from conformity but also *item quality*, which is usually ignored by existing methods. Some items exhibit higher popularity as they have intrinsic better property. Blindly removing the popularity bias would lose such important signal, and further deteriorate model performance. To sufficiently exploit such important information for recommendation, it is essential to disentangle the benign popularity bias caused by item quality from the harmful popularity bias caused by conformity.

Although important, it is quite challenging as we lack an explicit signal to differentiate the two factors of popularity bias. In this paper, we propose to leverage temporal information as the two factors exhibit quite different patterns along the time: item quality revealing item inherent property is stable and static while conformity that depends on items' recent clicks is highly time-sensitive.

Correspondingly, we further propose a novel **Ti**me-aware **D**is**E**ntangled framework (**TIDE**), where a click is generated from three components namely the static item quality, the dynamic conformity effect, as well as the user-item matching score returned by any recommendation model. Lastly, we conduct interventional inference so that the recommendation can benefit from the benign popularity bias while circumvent the harmful one. Extensive experiments on four real-world datasets demonstrated the effectiveness of TIDE.

Index Terms—Recommendation, Popularity Bias, Conformity, Item Quality

1 Introduction

Recent years have witnessed flourishing publications on recommendation, most of which aim at inventing machine learning models to fit users' historical behavior data [1]. However, the observation data usually exhibits severe *popularity bias*, *i.e.,* the distribution over items is quite imbalanced and even long-tailed. Such skewed distribution may be caused by the users' *conformity*, deviating from reflecting users' true preference. As a crucial factor for users' decision-making, conformity describes the tendency that user behaves following the group. In a typical recommender system, a user may click an item simply because he finds the item clicked by many other users, rather than based on his own judgement. As a result, recommendation model trained on such biased data would yield unexpected results, e.g., capturing skewed user preference and amplifying the long-tail effect. Given the wide existence of popularity bias and its negative impact on recommendation, we cannot emphasize too much the importance of tackling popularity bias.

Existing efforts mainly focus on entirely eliminating popularity bias to recover true user preference. However, we argue that *not all popularity bias is harmful*. Besides conformity effect, the uneven item distribution can also be attributed to the diversity of item quality. For example, some items exhibit higher popularity as they have intrinsic better properties, e.g., attractive story, harmonious music and professional actors for a typical movie. Blindly removing the popularity bias would lose such important signal, making the model fail to differentiate superb items that deserve more opportunities to be recommended. Therefore, we arrive at a dilemma: eliminating popularity bias would lose important quality signal, while maintaining popularity bias would suffer undesirable conformity effect. Now a question is raised: *is there a solution that enjoys the merit of the popularity bias while circumvents its bad effect?* To achieve this goal, it is essential to disentangle the harmful popularity bias caused by the conformity from the benign one caused by the item quality.

Although important, this problem has been under explored in the literature. The main challenge is the lack of explicit signals for disentanglement. Since we only have access to item popularity scores, which do not tell what factor causes this result. To deal with this problem, we propose to leverage the temporal information in differentiating the benign and harmful factors, as they exhibit quite different patterns along time: item quality which reveals item intrinsic property is stable and static, while conformity that depends on the number of recent clicks is highly time-

- *Zihao Zhao and Xiangnan He are with the School of Information Science and Technology, University of Science and Technology of China, Hefei, China. E-mail: zzh1998@mail.ustc.edu.cn, cjwustc@ustc.edu.cn and hexn@ustc.edu.cn.*
- *Jiawei Chen is with the College of Computer Science, Zhejiang University, Hangzhou, China. Most of his work was done when he was in the University of Science and Technology of China. E-mail: sleepyhunt@zju.edu.cn.*
- *Sheng Zhou is with the College of Computer Science, Zhejiang University, Hangzhou, China. E-mail: sleepyhunt@zju.edu.cn and zhousheng_zju@zju.edu.cn.*
- *Xuezhi Cao, Fuzheng Zhang, Wei Wu are with Meituan Dianping, Beijing, China. E-mail: caoxuezhi@meituan.com, zhangfuzheng@meituan.com, wuwei30@meituan.com.*
- *Jiawei Chen and Xiangnan He are the corresponding authors.*

sensitive. We also conduct empirical analyses on real-world datasets to validate this point, with making the following two interesting observations: (1) *The more popular an item is, the larger average rating value the item tends to acquire.* This observation reveals the existence of benign popularity bias — items with higher popularity usually suggest better quality and would receive more praise. (2) *From the temporal view, for a large proportion of items, the rating value exhibits negative correlation with the item popularity at that time.* This observation reveals temporal dynamic of harmful popularity bias — conformity exerts varying negative impact on users' behaviors with time going by.

Based on the above insights, we propose a **Ti**me-aware **D**is**E**ntangled framework (**TIDE**) for tackling popularity bias. We resort to the causal graph and assume click data is generated from three different components: (1) a time-invariant module that captures the quality of the item; (2) a temporal dynamic module that encodes the conformity effect by scrutinizing the number and time of recent clicks on the item; (3) a normal recommendation model that estimates user interest matching on the item. Such disentangled model provides opportunity to make better recommendation — inheriting the benign components while circumventing the harmful ones. Towards this end, during the inference stage, we conduct causal intervention on the conformity module to make the prediction beneficial from the item quality and interest matching score while immune to the harmful conformity effect.

Lastly, in terms of leveraging popularity bias in recommendation, the most relevant work is the recently proposed PDA [2]. However, we argue that directly injecting (predicted) item popularity score into prediction is insufficient for satisfactory recommendation as the harmful conformity effect is also injected. Distinct from PDA, our TIDE distills the benign popularity bias in prediction and yields significant empirical improvement.

In a nutshell, this work makes the following main contributions:

- To the best of our knowledge, this is the first work to study the problem of disentangling the benign popularity bias caused by item popularity from the harmful popularity bias caused by conformity in recommendation.
- We propose a novel time-aware disentangled framework TIDE for tackling popularity bias in recommendation. TIDE performs disentangled training by leveraging temporal information while resorts to intervention to block the harmful conformity effect during inference stage.
- Extensive experiments on four well-known benchmark datasets demonstrate the superiority of the proposed method over a range of state-of-the-arts. We will release our source code to facilitate future research.

The rest of this paper is organized as follows. We formulate the task and empirically explore popularity bias in section 2. We further present our proposed TIDE in section 3. The experimental results are presented in section 4. We briefly review related works in section 5. Finally, we conclude the paper and present some directions for future work in section 6.

2 PRELIMINARIES

In this section, we formulate the task and explore popularity bias on real-world datasets.

2.1 Problem Definition

We use uppercase character (e.g., U) to denote a random variable and lowercase character (e.g., u) to denote its specific value. We use characters in calligraphic font (e.g., $\mathcal{U}$) to represent the sample space of the corresponding random variable.

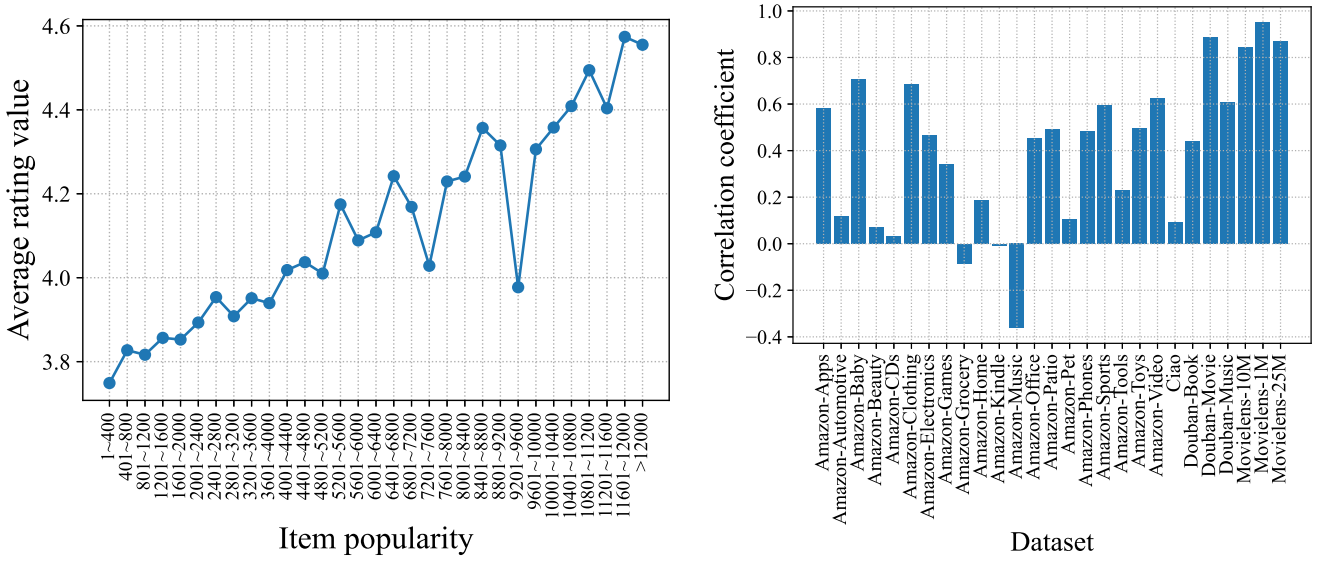
Suppose we have a recommender system with a user set $\mathcal{U}$ and an item set $\mathcal{I}$. Let u (or i) denote a user (or an item) in $\mathcal{U}$ (or $\mathcal{I}$). Let $\mathcal{D}$ denote the historical user behavior data, which was sequentially collected before the time T and notated as a set of triples, *i.e.*, $\mathcal{D} = \{(u_k, i_k, t_k)\}_{1 \le k \le |\mathcal{D}|}$, where the triple (u_k, i_k, t_k) denotes the user u_k has clicked the item i_k at the time t_k. For convenience, we collect users' feedback on the specific item i before time t as $\mathcal{D}_i^t = \{(u_l, i_l, t_l) \in \mathcal{D} | i_l = i, t_l < t\}$. Also, we define the popularity p_i of the item i as the number of observed interactions on i, *i.e.*, $p_i = |\mathcal{D}_i^T|$. The task of a recommendation system can be stated as follows: learning a recommendation model from $\mathcal{D}$ so that it can capture user preference and make a high-quality recommendation.

Popularity Bias, which denotes the uneven (usually long-tailed) distribution over the interaction frequency of items, is common in recommender systems. There are two factors resulting in popularity bias: (1) item quality, revealing the inherent excellence of items, which is benign; (2) conformity effect, describing a user tends to behave towards group norms while deviating from her own preference, which is harmful. This paper aims at disentangling the two factors such that the recommendation can benefit from the benign factor while circumvent the harmful one.

2.2 Empirical Analyses of Popularity Bias

In this subsection, to reveal the existence of the two factors and their properties, we conducted empirical analyses on real-world recommendation datasets including Amazon[1], Ciao[2], Douban[3] and Movielens[4]. Besides click information, these datasets also contain users' ratings on their clicked items, which provide ground truth label of their preference. A larger rating value suggests the user is more satisfied with the item. Two statistical analyses have been conducted: (1) We first explore the correlation between item popularity and their average ratings. We divide items into 30 groups according to their popularity p_i (where we segment popularity interval uniformly). We then calculate the average ratings of items in each group. The result on a typical dataset Douban-Movie is presented in Figure 1(a). We also report

1. https://jmcauley.ucsd.edu/data/amazon/
2. https://www.cse.msu.edu/~tangjili/datasetcode/truststudy.htm
3. https://github.com/DeepGraphLearning/RecommenderSystems/blob/master/socialRec/README.md#douban-data
4. http://files.grouplens.org/datasets/movielens/



(a) Average ratings with popularity on Douban-Movie

(b) Correlation Coefficient on various datasets

Fig. 1. We divide items into 30 groups according to their popularity and then calculate average rating values of items in each group. The left subplot (a) shows the relation of average rating values with their popularity on Douban-Movie. The right subplot (b) presents the Correlation Coefficient between average ratings and popularity on various datasets.

the Pearson Correlation Coefficient [3] between the average rating and popularity in terms of groups on various datasets in Figure 1(b). (2) We then explore the temporal dynamic of popularity bias. For each item, we calculate the Pearson Correlation Coefficient between the rating value and the time-aware *instant popularity* at that time, where *instant popularity* of item i at time t is defined as the number of clicks on the item during the past half year (*i.e.*, $|\mathcal{D}_i^t| - |\mathcal{D}_i^{t-t_o}|$, in which t_o denotes a period of half year[5]). The distribution of the calculated coefficients over items on two typical datasets is presented in Figure 2(a), 2(b). Here we filter out items with less than 20 interactions and exclude not significant results with $p > 0.2$. We also visualize the temporal evolution of the instant popularity for five randomly-selected items (Figure 2(c)), as well as an example of the relation between the rating value and the instant popularity (Figure 2(d)).

Two important observations are concluded from these results.

Observation 1. *The more popular an item is, the larger average rating value the item tends to have.*

Figure 1(b) demonstrates item average rating values exhibit positive correlation with item popularity in a large portion of datasets. This result suggests that popularity bias is not always harmful. The higher popularity of some items can be attributed to their better intrinsic quality, consequently, these items are more likely to be favored by users. Item popularity provides an important signal regarding to item quality, which is profitable to boost recommendation performance. Nevertheless, item popularity can not be directly leveraged in recommendation. Popularity would also be affected by the conformity effect, deviating from the quality. It can be seen from the severe fluctuation of the curve in Figure 1(a). Also, popularity exhibits weakly-positive or even negative correlation with average ratings in a considerable portion of datasets as the effect of the item quality is approached or even overrode by the conformity effect. Thus, we need to disentangle the effects from the two factors so that the recommendation can benefit from such benign knowledge while circumvent the impact of the harmful one.

Observation 2. *From the temporal view, for a large proportion of items, the rating value exhibits negative correlation with the item temporal popularity at that time.*

Figure 2(c), 2(d) demonstrates the dynamics of item instant popularity that conformity effect depends on. Besides, we observe that, when the instant popularity becomes larger, when the conformity exerts larger impact on user behavior, user's behavior deviates from his own preference to a large extent. Thus we can see the negative correlation between average ratings and instant popularity (Figure 2(a), 2(b)). This observation reveals the temporal dynamics of harmful popularity bias and motivates us to leverage temporal information in disentanglement to remove the harmful effect.

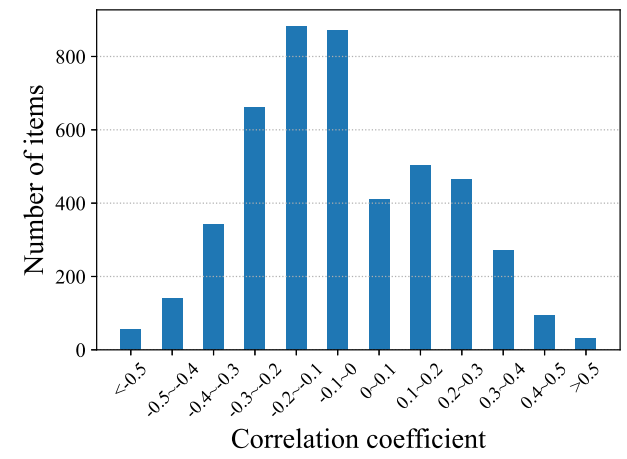
Based on above analyses, we make the following hypothesis, which lays foundation for our proposed method:

Hypothesis 1. *Popularity bias is mainly caused by both conformity effect and diverse item quality. Item quality that reveals item intrinsic property is stable and static, while conformity that depends on recent clicks is highly time-sensitive.*

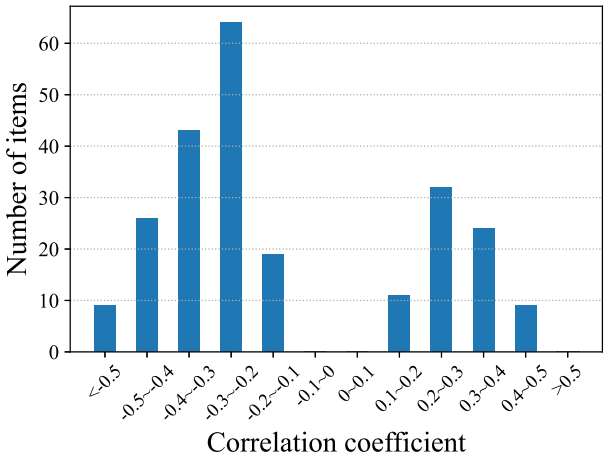
3 Time-aware Disentangled Framework

In this section, we present our time-aware disentangled framework (TIDE) for tackling popularity bias.

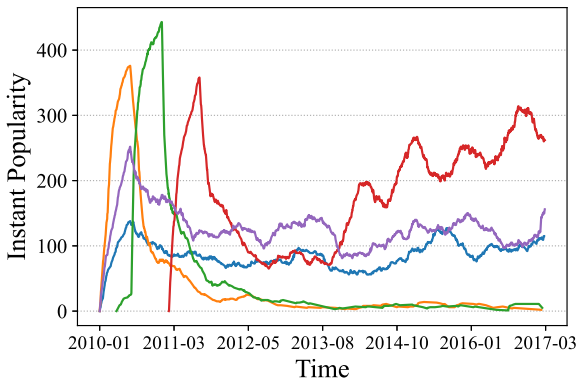
5. Here we simply choose half year for analyses, while the results in terms of other t_o (e.g., 1 month, 3 months, 1 year, 3 years) are presented in Appendix.



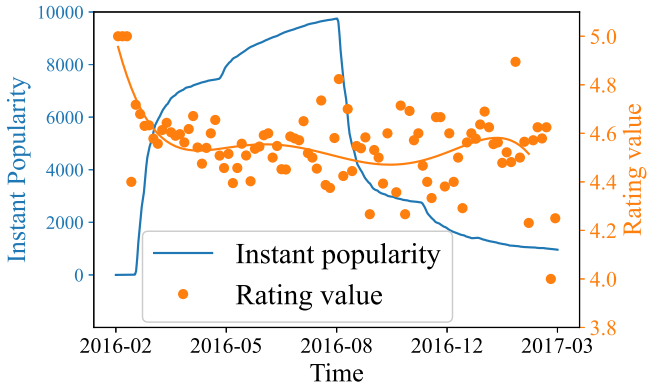
(a) Coefficient distribution on the dataset Douban-Movie.



(b) Coefficient distribution on the dataset Amazon-Music.

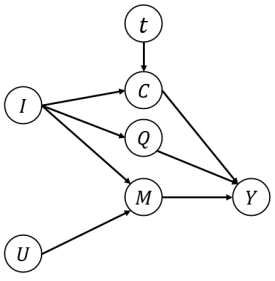


(c) Temporal evolving of the instant popularity for five randomly-selected items.

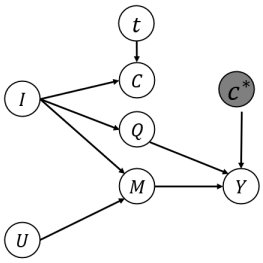


(d) The rating value with the instant popularity for an exemplified item. For better visualization, here we scatter the average ratings occurred within a week. Also, we plot a fitting curve for the rating value.

Fig. 2. We calculate the correlation coefficient between the rating value and the instant popularity at that time for each item, where instant popularity denotes the number of clicks on the item during the past half year. The subplots (a) and (b) illustrate the distribution of the calculated coefficient over items on two typical datasets; The subplot (c) illustrates the temporal evolving of the instant popularity for five randomly-selected items on Douban-Movie; The subplot (d) visualizes the relation of the rating value with the instant popularity for an exemplified item.



(a) Causal graph of our TIDE.



(b) Performing intervention on C during inference stage

Fig. 3. The subplot (a) illustrates causal graph of TIDE while the subplot (b) illustrates how we conduct interventional inference on TIDE.

3.1 Disentangled Learning

TIDE resorts to a causal graph as shown in Figure 3(a), consisting of seven types of nodes: (1) U: user; (2) I: item; (3) t: time; (4) C: conformity effect; (5) Q: item quality; (6) M: matching scores; (7) Y: prediction on user behavior.

TIDE assumes an observed click is generated from the following three disentangled components:

(1) $I \rightarrow Q \rightarrow Y$: This link denotes the effect of item quality on user behavior. An item with higher quality is more likely to be favored by a user. Here we simply use a time-irrelevant item-specific variable q_i for each item i to capture its inherent quality.

(2) $(I, t) \rightarrow C \rightarrow Y$: These links represent the time-aware conformity effect on user behavior. As suggested in Hypothesis 1, the impact of conformity not only depends on the time point t of this interaction, but also on the time and the number of past interactions on the item i. As such, we formulate the following parameterized function $g_\beta(\cdot)$ to estimate the strength of conformity effect of item i at time t:

$$c_i^t = g_\beta(t, \mathcal{D}_i^t) = \beta_i \sum_{(u_l, i_l, t_l) \in \mathcal{D}_i^t} \exp(-\frac{|t - t_l|}{\tau}), \quad (1)$$

where a parameter β_i is introduced for each item i to re-scale the effect, as conformity usually exhibits more severe

on some items (*e.g.,* soap opera) than others (*e.g.,* science documentary). Here we simply accumulate the stimulations from past interactions while discount their contribution according to the time interval, so that every interaction of item i before current time t would contribute to its conformity bias value c_i^t with different degree according to the time interval from now$(t-t_l)$. This setting is coincident with our intuition — the currently popular items would have larger impact on us than the ones that were popular in the far past. We also introduce a coefficient τ to control the sensitivity of c_i^t to the time. A smaller τ would make the model focus more on recent interactions and immunize the interactions occurred long time ago.

(3) $(U, I) \rightarrow M \rightarrow Y$: these links project user and item features (*e.g.,* IDs) into their matching scores $m_{ui} = f_\theta(u, i)$. $f_\theta(u, i)$ can be implemented by various recommendation models, such as MF [4], LightGCN [5], DIN [6], etc.

Finally these three components are aggregated into a final prediction score for recovering the observed historical interactions:

$$\hat{y}_{ui}^t = \text{Tanh}(q_i + c_i^t) \times \text{Softplus}(m_{ui}), \tag{2}$$

where a parameter q_i is introduced to capture the quality of each item i. Tanh$(\cdot)$ is an activation function that project the combined value (always positive) into interval [0,1] to make the model more stable; and Softplus$(\cdot)$ is an activation function to ensure the positivity of the matching score. Tanh$(q_i + c_i^t)$ can be understood as popularity bias which combines the benign effect from the item quality q_i $(Q \rightarrow Y)$ and the harmful effect from the conformity c_i^t $(C \rightarrow Y)$.

We can still apply the commonly-used BPR [4] recommendation loss over the final prediction score to train the model. Formally, the training loss is given as follows:

$$L = \sum_{(u,i,t)\in\mathcal{D}, j\sim P_n} -\log(\sigma(\hat{y}_{ui}^t - \hat{y}_{uj}^t)), \tag{3}$$

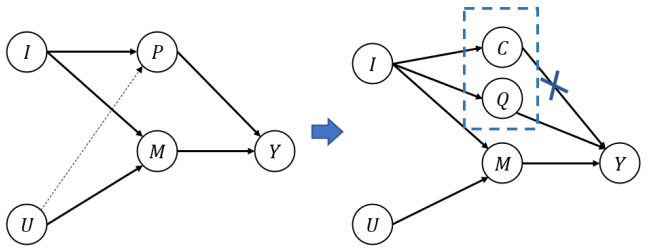
where $\sigma(\cdot)$ represents the sigmoid function. We conduct negative sampling to draw 4 negative samples (j) for each positive instance (i) from distribution P_n for training our model. As recent work [2], here we simply use a uniform negative sampling strategy for fair comparison. Note that we have omitted the L_2 regularization terms for clarity.

3.2 Intervention-based Inference

As shown in Figure 3(a), I influences Y through three paths: $I \rightarrow Q \rightarrow Y$ through item quality, $I \rightarrow C \rightarrow Y$ through conformity effect and $I \rightarrow M \rightarrow Y$ through user-item matching score. In order to make the recommendation benefit from the useful factors while circumvent the harmful one, we perform the causal intervention to cut off the path $I \rightarrow C \rightarrow Y$ as shown in Figure 3(b) where improper effect from the conformity has been removed. Formally, we directly intervene c_i^t with a fixed value c^* and make the prediction as:

$$\hat{y}_{ui}^* = \tanh(q_i + c^*) \times \text{Softplus}(m_{ui}). \tag{4}$$

We simply set the c^* as 0 in our experiments.



(a) Causal graph of recent work (b) Causal graph of TIDE

Fig. 4. Causal graph comparison of recent work with our TIDE.

3.3 Links to Recent Work

Recent years have witnessed various debiasing strategies for popularity bias. Among which, causal inference is the most successful and representative strategy [2], [7], [8]. We argue that the inherent nature of this kind of methods is disentanglement — undo the effect of the popularity bias to recover user preference on items. The causal graph of these methods can be simply summarized as Figure 4(a). Although this graph may be different from the causal graph claimed in the original papers, Figure 4(a) is indeed coincident with their models. For example, PDA [2] assumes a click is generated with combining item popularity score and user-item matching score, *i.e.,* $\hat{y}_{ui} = p_i^\gamma \times \text{Elu}'(m_{ui})$; DICE [7] makes a similar assumption except that they model the sensitivity of users to item popularity (as marked by the dash line in Figure 4(a)).

This work lies on this scheme but we further conduct disentanglement of popularity bias. As Figure 4(b) shows, we split the path regarding to popularity bias $(I \rightarrow P \rightarrow Y)$ into two paths: $I \rightarrow Q \rightarrow Y$ the benign effect from item quality and $I \rightarrow C \rightarrow Y$ the harmful effect from conformity. Besides, during the inference stage, instead of blindly removing popularity bias as [7], [8] (cutting $I \rightarrow P \rightarrow Y$) or leveraging complete popularity bias in prediction as [2], we utilize partial popularity bias — leveraging benign part (maintain path $I \rightarrow Q \rightarrow Y$) while removing harmful part (cut path $I \rightarrow C \rightarrow Y$). In this way, our TIDE can distill useful information from item popularity and thus yield empirical improvement over them.

4 Experiments

In this section, we conduct experiments to evaluate the performance of our proposed TIDE. Our experiments are intended to address the following research questions:

- **RQ1:** Does TIDE outperform SOTA methods for popularity bias?
- **RQ2:** Is it beneficial to model both static item quality and dynamic conformity effect? Is it beneficial to remove the effect of conformity during the inference stage?
- **RQ3:** Do the learned parameters q_i capture item quality?

4.1 Experimental Setup

Datasets. We choose four well-known datasets Douban-Movie, Amazon-CDs, Amazon-Music and Ciao for our experiments, in which Douban-Movie has a strong positive

TABLE 1
Statistics of the datasets.

Dataset	User #	Item #	Interaction #	Date
Douban-Movie	48,799	26,813	7,409,868	2010.1-2017.3
Amazon-CDs	75,258	64,443	1,097,592	1997.11-2014.7
Ciao	5,868	10,724	143,217	2000.5-2011.4
Amazon-Music	5,541	3,568	64,706	1998.4-2014.7

correlation coefficient while Amazon-Music is the most negative dataset as shown in 1(b), and the other two datasets have a relatively small correlation coefficient. We select diverse datasets in experiments to demonstrate the effectiveness and robustness of our model. These datasets contain users' rating records in a chronological order, where each interaction is rated ranging from 1 to 5 points indicating users' satisfaction from low to high. Since it is unreliable to include users and items with few interactions for evaluation, we conduct 5-core filtering for the datasets Ciao, Amazon-CDs and Amazon-Music, and 10-core filtering for Douban-Movie. The statistics of the datasets are described in Table 1. We follow the setting of PDA [2] and split the datasets chronologically. Specifically, we split the datasets into 10 parts according to the interaction time, and each part has the same time interval. The first nine parts are used for training, while the last part is left for validation and testing, in which the interactions of half of the users are organized as the validation set while others are organized as the test set. We also transform the data into binary implicit feedback for experiments as [2], [9]. That is, as long as there exists a rating, the corresponding implicit feedback is assigned a value of 1, suggesting the item has been interacted (*i.e.,* clicked) by the user.

Evaluation Methodology. We train a model with binary training data and evaluate its performance on the following two tasks:

- *Click prediction task:* We evaluate how accurate a model forecasts users' future clicks. Specifically, we apply the model to sort the items that have not been interacted, and test whether the top-K items would be clicked by the user in the future (*i.e.,* in test data). For the matrics, we employ Recall@K (called CP-Rec@K in this task), Precision@K (CP-Pre@K) and Normalized Discounted Cumulative Gain@K (CP-NDCG@K) for evaluating model performance in this task.
- *Preference prediction task:* Note that click is not always coincident with user preference. We further evaluate how a model retrieves relevant items that users are indeed fond of. We resort to the ground truth rating value, and consider the item with a high rating value (*e.g.,* 5) as positive. As we do not know user's true preference on unrated items, in this task, we just rank the rated items in the test data and evaluate whether the positive items are retrieved within Top-K positions. Specifically, precision@K (marked as PP-Pre@K) and recall@K (PP-Rec@K) are adopted in this task. Also, considering the number of rated items is usually small, we set a relatively small K (*e.g.,* $K = 3$).

Comparison methods. Five types of methods are tested in our experiments:

- MF [4]: the basic matrix factorization model with BPR loss.
- MF-IPS [10], [11]: a classic strategy for eliminating popularity bias by re-weighting each instance according to item popularity. We refer to [12] and apply a max-capping trick on IPS value to reduce variance.
- DICE [7]: a framework that leverages cause-specific data to disentangle user preference and popularity bias into two sets of embeddings.
- PD and PDA [2]: a state-of-the-art method that performs deconfounded training while intervenes the popularity bias during model inference. We report two versions of this work: PD that directly uses matching score for recommendation; PDA that leverages predicted item popularity score in recommendation. As PDA demonstrates superior performance over ranking-based methods [13], [14], we do not include these methods as baselines.
- TIDE: the method proposed in this work. We mainly test two versions of TIDE: TIDE-full, combining all the effect from three components for predicting user future click, *i.e.,* we use $\hat{y}_{ui}^{T}$ for ranking; TIDE-int, which performs intervention to cut off the effect from the conformity, *i.e.,* $\hat{y}_{ui}^{*}$ is utilized.

Implementation details. Matrix Factorization (MF) has been selected as the main backbone recommendation model for experiments, and it would be straightforward to replace it with more sophisticated models such as Factorization Machine [15], or Neural Network [5], [16]. We also utilize reparametrization trick to ensure the positivity of the learned q_i and β_i, *i.e.,* $q_i \leftarrow Softplus(q_i), \beta_i \leftarrow Softplus(\beta_i)$. We optimize our TIDE with Adam optimizer. Grid search is used to find the best hyper-parameters based on the performance on the validation set. The search space of learning rate and weight decay of the parameters in MF is {1e-4, 1e-3, 1e-2}; also, we set the decay of q_i and β_i as 0, and search their initialization in [-5,-1] with step 1 and learning rate in {1e-4, 1e-3, 1e-2, 1e-1}; τ is set as 1e7, batch-size is set as 2,048.

For the experiments on LightGCN-based models, we set the search space and parameter setting as same as the MF-based model except the batch size of Douban-movie and Amazon-CDs is increased to 8,192 for speed and the number of convolutional layers is searched in {2, 3, 4} as advised by [5]. We adopt the early stopping strategy that stops training if performance on the validation data does not increase for 10 epochs. The setting of compared methods is either determined by grid search in our experiments or suggested by their original papers.

All experiments are conducted on a server with 2 Intel E5-2620 CPUs, 4 NVIDIA GTX2080 GPUs and 256G RAM. The source code will be available at Github [6].

4.2 Performance Comparison (RQ1)

Performance on click prediction task. Table 2 presents the performance of the compared methods on the click

6. https://github.com/zzhUSTC2016/TIDE

TABLE 2
Performance comparison on the click prediction task with MF [4] as backbone. The boldface font denotes the winner in that column. $K = 20$.

Datasets	Douban-Movie			Amazon-CDs			Amazon-Music			Ciao		
Metrics	CP-Rec@K	CP-Pre@K	CP-Ndcg@K	CP-Rec@K	CP-Pre@K	CP-Ndcg@K	CP-Rec@K	CP-Pre@K	CP-Ndcg@K	CP-Rec@K	CP-Pre@K	CP-Ndcg@K
MF	0.0223	0.0342	0.0370	0.0119	0.0030	0.0035	0.0362	0.0068	0.0080	0.0107	0.0076	0.0086
MF-IPS	0.0220	0.0337	0.0366	0.0118	0.0030	0.0035	0.0378	0.0063	0.0071	0.0109	0.0068	0.0078
DICE	0.0202	0.0323	0.0343	0.0073	0.0019	0.0021	0.0357	0.0058	0.0060	0.0145	0.0103	0.0110
PD	0.0355	0.0465	0.0520	0.0140	0.0032	0.0036	0.0418	0.0071	0.0083	0.0177	0.0110	0.0118
PDA	0.0408	0.0534	0.0596	0.0194	0.0044	0.0052	0.0656	0.0111	0.0125	0.0189	0.0144	**0.0159**
TIDE-full	**0.0483**	**0.0590**	**0.0671**	**0.0243**	**0.0058**	**0.0068**	**0.0837**	**0.0152**	**0.0175**	**0.0244**	**0.0148**	0.0154
Impv	18.59%	10.53%	12.63%	25.27%	30.93%	31.98%	27.69%	36.71%	39.79%	28.99%	2.50%	-2.77%

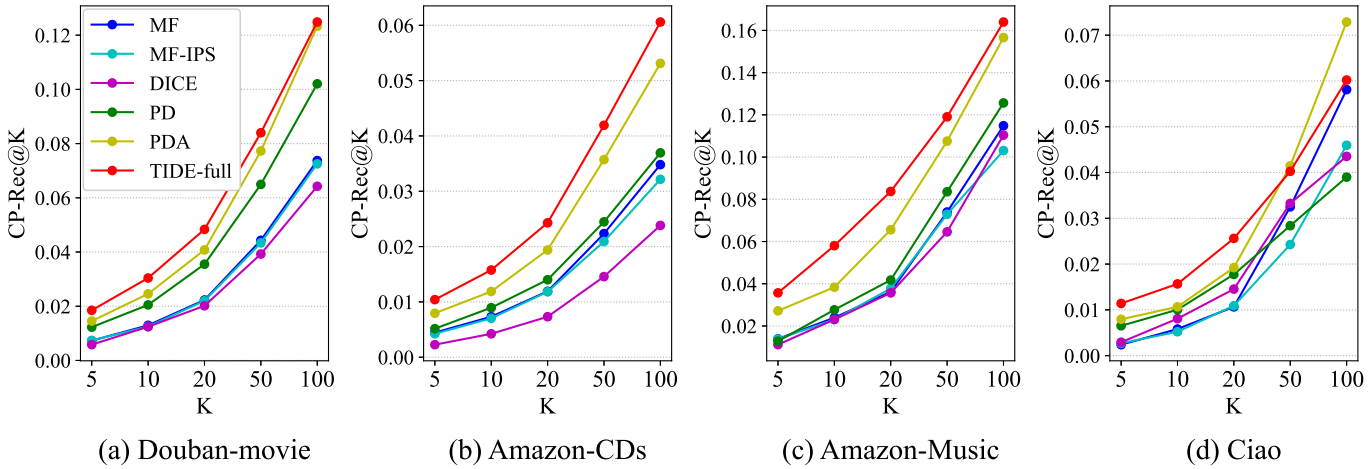


Fig. 5. Performance comparison of CP-Rec@K where K is set as different value when MF is the backbone model.

TABLE 3
Performance comparison on the preference prediction task. The boldface font denotes the winner in that column.

Datasets	Douban-Movie		Amazon-CDs		Amazon-Music		Ciao	
Metrics	PP-Rec@3	PP-Pre@3	PP-Rec@3	PP-Pre@3	PP-Rec@3	PP-Pre@3	PP-Rec@3	PP-Pre@3
MF	0.1690	0.4357	0.4234	0.6970	0.4692	0.7031	0.2609	0.5564
MF-IPS	0.1676	0.4317	0.4226	0.6983	0.4628	0.6976	0.2658	0.5641
DICE	0.1735	0.4509	0.4195	0.6928	0.4528	0.6794	0.2591	0.5256
PD	0.1621	0.4133	0.4222	0.6956	0.4687	0.7031	0.2664	0.5744
PDA	0.1659	0.4109	0.4277	0.7031	0.4617	0.6922	0.2368	0.5205
TIDE-full	0.1570	0.3873	0.4302	0.7074	0.4678	0.6976	0.2593	0.5538
TIDE-int	**0.1780**	**0.4693**	**0.4362**	**0.7178**	**0.4855**	**0.7250**	**0.2670**	**0.5795**

prediction task in terms of three evaluation metrics. The boldface font denotes the winner in that column. For the sake of clarity, the row 'Impv' shows the relative improvement achieved by TIDE-full over all the baselines. Overall, with few exceptions, our TIDE-full outperforms all compared baselines. Especially in the dataset Amazon-Music, the improvements are quite impressive — 27.69%, 36.71% and 39.79% in terms of Precision, Recall and NDCG respectively. To further validate the performance of our model, we report the metric CP-Rec at different K value. As shown in Figure 5, our model TIDE-full outperforms other methods consistently in all datasets with few exceptions. These results validate that, by utilizing both the item quality information and the user conformity effect, TIDE-full can capture more precise popularity bias and thus make a more accurate prediction of users' future behavior.

Performance on preference prediction task. Table 3 presents the performance of the compared methods on preference prediction task. We have the following observations: (1) PDA, which consistently outperforms PD in the click prediction task, performs worse in this task. This interesting phenomenon reveals the negative impact of popularity bias. Blindly injecting popularity bias without filtering out its harmful ingredient would deteriorate the model's capability to capture user interests. Similar results can be seen from the worse performance of TIDE-full than TIDE-int. (2) Overall, with few exceptions, our TIDE-int outperforms all compared methods in this task. This result validates the effectiveness of disentangling benign and harmful factors of popularity bias. Without disentanglement, existing methods sink into a dilemma — they either fail to utilize the important signal of the item quality (*e.g.*, TIDE-int outperforms PD, DICE, MF-IPS), or are disturbed by the harmful conformity effect (*e.g.*, TIDE-int outperforms PDA and MF). By disentangling the two factor and intervening the harmful factor during the inference, our TIDE-int method could

TABLE 4
Performance comparison on the click prediction task with LightGCN [5] as backbone. The boldface font denotes the winner in that column. $K = 20$.

Datasets	Douban-Movie			Amazon-CDs			Amazon-Music			Ciao		
Metrics	CP-Rec@K	CP-Pre@K	CP-Ndcg@K	CP-Rec@K	CP-Pre@K	CP-Ndcg@K	CP-Rec@K	CP-Pre@K	CP-Ndcg@K	CP-Rec@K	CP-Pre@K	CP-Ndcg@K
LightGCN	0.0227	0.0363	0.0390	0.0137	0.0036	0.0040	0.0517	0.0087	0.0095	0.0171	0.0094	0.0105
PD_{GCN}	0.0332	0.0481	0.0534	0.0143	0.0036	0.0040	0.0425	0.0080	0.0084	0.0146	0.0088	0.0190
PDA_{GCN}	0.0409	0.0559	0.0624	0.0195	0.0047	0.0052	0.0662	0.0117	0.0123	0.0182	**0.0142**	**0.0150**
$TIDE\text{-}full_{GCN}$	**0.0480**	**0.0651**	**0.0734**	**0.0234**	**0.0056**	**0.0062**	**0.0765**	**0.0139**	**0.0154**	**0.0234**	0.0133	0.0138
Impv	17.29%	16.49%	17.62%	19.93%	19.06%	19.54%	15.45%	19.04%	25.94%	28.60%	-6.33%	-7.95 %

enjoy the merit of the popularity bias while circumvent its bad effect.

Performance with GCN-based backbone model. To further validate the effectiveness and the generalization of TIDE, we make an experiment on a typical GCN-based backbone model, *i.e.*, LightGCN. The results are presented in Table 4. Here we simply choose the most SOTA and relevant baselines PD and PDA for comparison. As we can see, with few exceptions our TIDE still outperforms the compared methods in this setting.

4.3 Ablation Study (RQ2)

We conduct ablation study to explore whether it is essential to model both factors and whether it is essential to perform interventional inference. We compare our TIDE-full and TIDE-int with the following special cases: (1) TIDE-noq and TIDE-noc: where item quality (Q) or conformity effect (C) is removed in both training and inference stage; (2) TIDE-e: which is trained as same as TIDE-int but only uses matching score for recommendation. The characteristics and performance on the preference prediction task are presented in Table 5.

Effectiveness of modeling both factors. We observe that the method modeling two factors (TIDE-int) consistently outperforms the cases just considering one aspect (TIDE-noq and TIDE-noc). This result is coincident with our intuition — modeling both factors is beneficial for capturing popularity bias as well as for distilling useful knowledge about item quality from it.

Effectiveness of interventional inference. From Table 5, we observe TIDE-int is consistently superior over TIDE-e and TIDE-full. This result demonstrates the mix nature of popularity bias — containing both benign and harmful signals. The model that roughly maintains (TIDE-full) or removes (TIDE-e) both of them would result in undesirable performance.

4.4 Exploratory Analysis (RQ3)

To answer the question RQ3, we now explore the learned q_i from two perspectives to provide insights into how TIDE captures item quality.

Distribution of learned q_i. Figure 6 visualizes the distribution of the learned q_i with their average rating value (simply marked as AR_i) on a typical dataset Douban-Movie. We can observe the strong positive correlation between them, suggesting our learned parameters q_i capture the item quality successfully. Also, comparing with Figure 1(a), the curve in Figure 6 is more stable and exhibits less fluctuation.

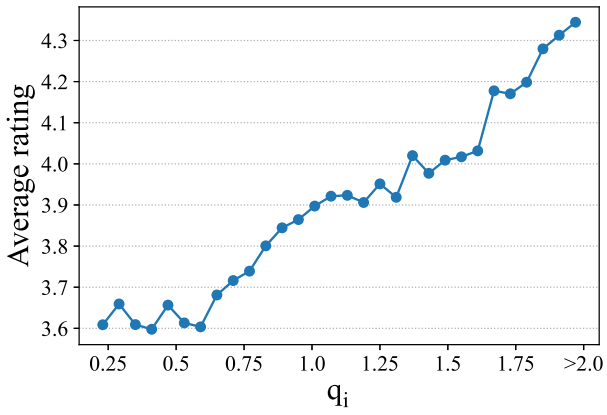


Fig. 6. We divide items into 30 groups according to their learned q_i and then calculate average rating values of items in each group. This figure visualizes the relation of the average rating value with the learned q_i on Douban-Movie.

To further demonstrate the ability of q_i in capturing item quality, we also report the results on the dataset Amazon-Music where item popularity has a negative correlation with the average ratings. In Figure 7, we plot the relation (red line) between the learned q_i and the average rating AR_i, as well as the relation (blue line) between item popularity and AR_i for comparison. The result shows that although in such a hard dataset, q_i can still capture information about item quality and filter out the distraction of the severe conformity effect.

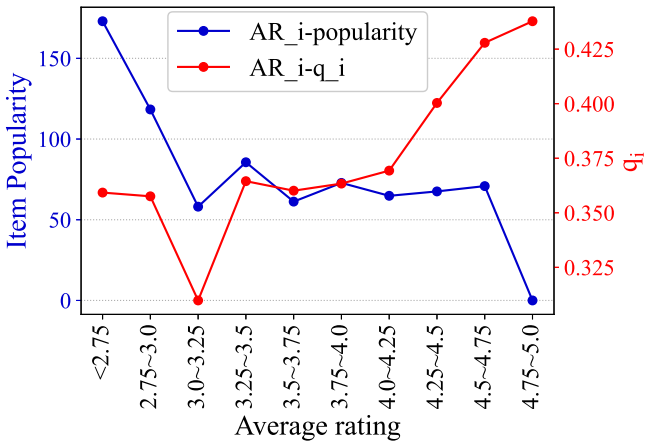


Fig. 7. We divide items in Amazon-Music into 10 groups according to their average rating, and visualize the average item popularity (Blue line) and the average q_i (Red line) in each group.

TABLE 5
Characteristics of TIDE and its variants. We also report their performance on the preference prediction task. M: Matching Scores, Q: Quality, C: Conformity

Methods	Training with?			Inference with?			Performance							
							Douban-Movie		Amazon-CDs		Amazon-Music		Ciao	
	M?	Q?	C?	M?	Q?	C?	PP-Rec@3	PP-Pre@3	PP-Rec@3	PP-Pre@3	PP-Rec@3	PP-Pre@3	PP-Rec@3	PP-Pre@3
MF	√	×	×	√	×	×	0.1690	0.4357	0.4234	0.6970	0.4692	0.7031	0.2609	0.5564
TIDE-noc	√	√	×	√	√	×	0.1706	0.4394	0.4319	0.7112	0.4657	0.7031	0.2494	0.5333
TIDE-noq	√	×	√	√	×	√	0.1566	0.3871	0.4255	0.6988	0.4562	0.6831	0.2547	0.5410
TIDE-e	√	√	√	√	×	×	0.1527	0.3750	0.4234	0.6977	0.4843	**0.7250**	0.2651	0.5564
TIDE-full	√	√	√	√	√	√	0.1570	0.3873	0.4302	0.7074	0.4678	0.6976	0.2593	0.5538
TIDE-int	√	√	√	√	√	×	**0.1780**	**0.4693**	**0.4362**	**0.7178**	**0.4855**	**0.7250**	**0.2670**	**0.5795**

Ranking correlation comparison. We further validate the stronger correlation of the average rating value with q_i than with popularity p_i. We calculate the Kendall Tau Ranking Correlation Coefficient (RCC) [17] between the item lists ranked by them. RCC essentially measures the probability of two random items being in the same order in the two ranked lists, and would be more robust and rational than Pearson Correlation Coefficient (PCC) especially for the recommendation task. The result is presented in Table 6. We observe RCC between q_i and AR_i is consistently larger than RCC between p_i and AR_i in all four datasets. Besides, to our surprise, we observe the absolute values of both metrics are relatively small. More seriously, RCC between p_i and AR_i is negative on the datasets Amazon-CDs, Amazon-Music and Ciao. This result validates the challenging of tackling popularity bias. There exists a gap between the value and ranking — positive correlation in terms of value may not result in positive correlation in ranking. Although popularity exhibits positive correlation with AR_i in PCC, its ranking result is easily distorted by other factors in popularity and deviates from reflecting positive correlation. TIDE filters out conformity effect from popularity bias and relatively captures more stable and precise knowledge of item quality.

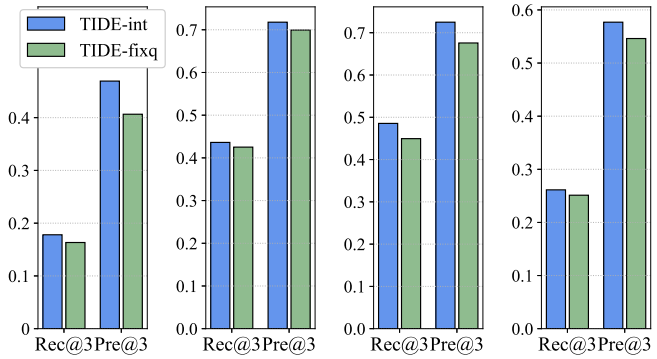
TABLE 6
Ranking Correlation Coefficient (RCC) between the lists ranked by average rating AR_i and by the learned parameter q_i, as well as between the lists ranked by average rating AR_i and by popularity p_i.

	Douban-Movie	Amazon-CDs	Amazon-Music	Ciao
AR_i with q_i	0.0947	0.0883	0.0815	0.0777
AR_i with p_i	0.0384	-0.0646	-0.03284	-0.0318

Effectiveness of learning diverse q_i. To validate the necessary of learning diverse q_i, we compare TIDE-int with its special case TIDE-fixq, where q_i for all items are fixed as a constant value. The results are presented in Figure 8. In all datasets, TIDE-int consistently outperforms TIDE-fixq with a certain margin. This result demonstrates that by training personalized q_i for each item, our model indeed learns some useful information, which is beneficial for capturing item quality and promoting recommendation performance.

5 Related Work

In this section, we review the most related works from the following two perspectives.



(a)Douban-movie (b)Amazon-CDs (c)Amazon-Music (d)Ciao

Fig. 8. Comparison of TIDE-int with its special case TIDE-fixq, where all q_i are constrained to a constant value.

Popularity Bias in recommendation. Popularity bias depicting uneven (usually long-tailed) distribution over the interaction frequency of items, is common in a recommender system. The negative impact of popularity bias has been studied in a large number of recent literature. For example, some works [7], [18] argue that such skewed distribution may be caused by user conformity, deviating from reflecting users' true preference. As such, recommendation models trained on such biased data would give skewed prediction. Worse still, recommendation model not only inherits the bias, but also amplifies bias, making the popular items dominate the top recommendations [13], [19], [20], [21], [22], [23]. This phenomenon has been empirically verified by Abdollahpouri [24]. They find popular items are recommended to a much greater degree than even what their initial popularity warrants. It would decrease serendipity [25], [26], [27] and fairness [13], [28], [29], [30] of recommendation results, hurting user experience and causing customer churn.

Recent works on tackling popularity bias can be mainly classified into four types: (1) Inverse propensity scoring (IPS) [11], [31] is a classic strategy that directly adjust the data distribution with re-weighting each instance according to item popularity. (2) Ranking adjustment is another type of method [13], [14] that directly re-rank the recommendation list to improve the recommendation opportunity of unpopular items. Although simple and straightforward, this type of methods relies on heuristic artificial design and usually sacrifices recommendation accuracy. (3) Regularization has been introduced by some researchers to push the model

towards balanced recommendation [32], [33], [34], [35]. For example, Chen *et al.* [32] leverage regularization to transfer the knowledge from these well-trained popular items to the long-tail items; Bonner *et al.* [33] leverage regularization to distill knowledge from the uniform data for addressing popularity bias. (4) Causal inference has been leveraged for addressing popularity bias. These methods mainly assume the generative process of the data with causal graphs and then disentangle the popularity bias from the user preference accordingly [2], [7], [8].

However, most of existing methods focus on eliminating popularity bias. In fact, popularity bias is not always evil. It may not only result from the users' conformity to the group, but also from item quality. It would be valuable to leverage such important signal in boosting recommendation performance. To the best of our knowledge, only one work [2] considers to leverage popularity bias into recommendation. However, they directly injecting (predicted) item popularity score into prediction, which is insufficient for satisfactory recommendation as the harmful conformity effect is also injected. Different from these works, we consider the double-edged nature of popularity bias. We aim at disentangling the benign popularity bias from the harmful one, so that the recommendation can benefit from the merit while circumvent the harmful.

Biases in recommendation. Besides popularity bias, recent works have studied other types of biases in recommendation including: Selection bias, which happens as users are free to choose which items to rate, so that the observed ratings are not a representative sample of all ratings [36], [37], [38], [39]; Exposure bias, which happens in implicit feedback data as users are only exposed to a part of specific items [24], [36], [40], [41]; Position bias, which happens as users tend to interact with items in higher position of the recommendation list [36], [42]; Unfairness [43], [44], which denotes the system systematically and unfairly discriminates against certain individuals or groups of individuals in favor others. Generally, there are substantial works on addressing these biases issues. We encourage the readers to refer to the survey [36] for more details.

Disentanglement in recommendation. In terms of disentanglement, existing efforts can be classified into two lines. The first type of methods is designed for debiasing. As discussed above, this type of methods aim at disentangling user true preference from the various data biases [7], [8]. Another type of methods lie in disentangled representation learning. This kind of methods aims at learning a finer-granularity representation of users and items, which is beneficial for robust and explainable recommendation. For example, Ma *et al.* [45] leverage Variational Auto-Encoder [46] to disentangle high-level concepts associated with user intentions as well as low-level factors (*e.g.*, size or color of a shirt). Similarly, Wang *et al.* [47] learn disentangled user representation with the merits of the interaction graph.

6 Conclusion

This paper studies an important but unexplored problem — how to disentangle the benign popularity bias caused by item quality from the harmful popularity bias caused by conformity. We first conduct empirical analyses on real-world datasets and observe quite different patterns of these two factors along time: item quality revealing item inherent property is stable and static while conformity that depends on item recent clicks is highly time-sensitive. We then propose a novel time-aware disentangled framework (TIDE), where a click is generated from three components namely the static item quality, the dynamic conformity effect, as well as the user-item matching score. We further provide an interventional inference strategy such that the recommendation can benefit from the benign popularity bias while circumvent the harmful one. Extensive experiments on four real-world datasets demonstrated the effectiveness of the proposed disentangled model as well as its interventional inference strategy.

One interesting direction for future work is to explore a more sophisticated conformity model $g_\beta(\cdot)$, which could capture more complex patterns and potentially achieve better performance than simple sum-exponential structure. Besides, this work demonstrates popularity bias is double-edged. We believe other biases may also have this nature. It will be valuable to transfer the experience of this work to tackle other biases and to explore their benign and harmful effect on recommendation.

Appendix

We provide more data analyses for better understanding of our observations. In Figure 2, we define the instant popularity as the number of clicks on the item during the past half year. To show that our observations are not sensitive to the lengths of the time slot, we report the corresponding results of Figure 2(a) and 2(c) with different lengths of the time slot ranging from 1 month to 3 years as shown in Figure 9 and Figure 10. These results validate that our observation 2 is stable with the length of time slot.

Figure 11 gives more examples demonstrating the negative correlation between the average ratings and the instant popularity.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (2021ZD0111802), the National Natural Science Foundation of China (62102382, 61972372, U21B2026), the Meituan Inc. through Research Cooperation Project, and the CCCD Key Lab of Ministry of Culture and Tourism.

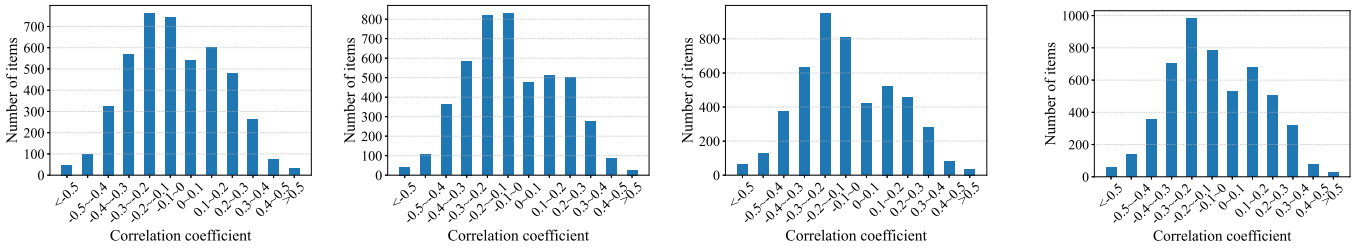
References

[1] L. Wu, X. He, X. Wang, K. Zhang, and M. Wang, "A survey on neural recommendation: From collaborative filtering to content and context enriched recommendation," *CoRR*, vol. abs/2104.13030, 2021.

[2] Y. Zhang, F. Feng, X. He, T. Wei, C. Song, G. Ling, and Y. Zhang, "Causal intervention for leveraging popularity bias in recommendation," *arXiv preprint arXiv:2105.06067*, 2021.

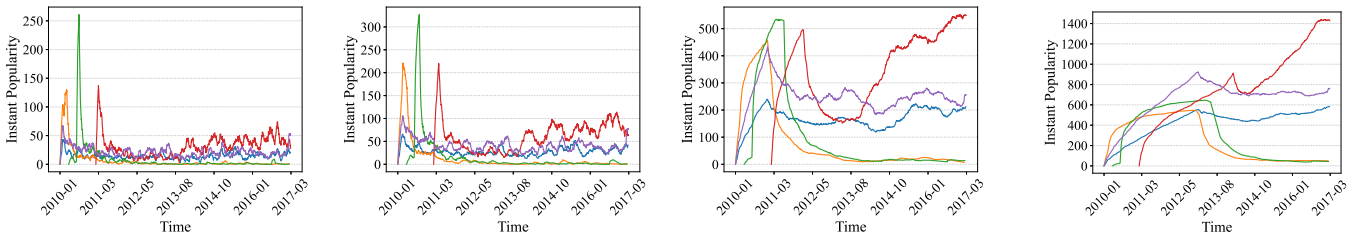
[3] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[4] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: bayesian personalized ranking from implicit feedback," *CoRR*, vol. abs/1205.2618, 2012.



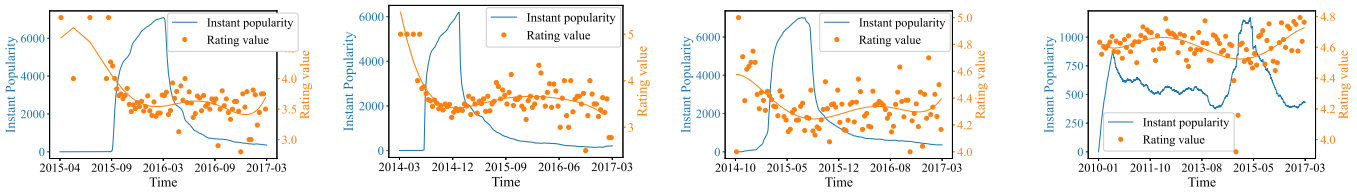
(a) 1 month (b) 3 months (c) 1 year (d) 3 years

Fig. 9. The distribution of the correlation coefficient between the rating value and the instant popularity on Douban-Movie, where instant popularity denotes the number of clicks on the item in different lengths of time slot.



(a) 1 month (b) 3 months (c) 1 year (d) 3 years

Fig. 10. This figure illustrates the temporal evolving of the instant popularity for five randomly-selected items on Douban-Movie with instant popularity calculated in different lengths of time slot.



(a) Item a (b) Item b (c) Item c (d) Item d

Fig. 11. More examples on Douban-movie which visualize the relation of the rating value with the instant popularity.

[5] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "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*, 2020, pp. 639–648.

[6] G. Zhou, X. Zhu, C. Song, Y. Fan, H. Zhu, X. Ma, Y. Yan, J. Jin, H. Li, and K. Gai, "Deep interest network for click-through rate prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1059–1068.

[7] Y. Zheng, C. Gao, X. Li, X. He, Y. Li, and D. Jin, "Disentangling user interest and conformity for recommendation with causal embedding," in *Proceedings of the Web Conference 2021*, 2021, pp. 2980–2991.

[8] T. Wei, F. Feng, J. Chen, Z. Wu, J. Yi, and X. He, "Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system," in *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, F. Zhu, B. C. Ooi, and C. Miao, Eds. ACM, 2021, pp. 1791–1800.

[9] J. Chen, C. Wang, S. Zhou, Q. Shi, Y. Feng, and C. Chen, "Samwalker: Social recommendation with informative sampling strategy," in *The World Wide Web Conference*. ACM, 2019, pp. 228–239.

[10] T. Joachims, A. Swaminathan, and T. Schnabel, "Unbiased learning-to-rank with biased feedback," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2017, pp. 781–789.

[11] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims, "Recommendations as treatments: Debiasing learning and evaluation," in *international conference on machine learning*. PMLR, 2016, pp. 1670–1679.

[12] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson, "Counterfactual reasoning and learning systems: The example of computational advertising." *Journal of Machine Learning Research*, vol. 14, no. 11, 2013.

[13] H. Abdollahpouri, R. Burke, and B. Mobasher, "Managing popularity bias in recommender systems with personalized re-ranking," in *The thirty-second international flairs conference*, 2019.

[14] Z. Zhu, Y. He, X. Zhao, Y. Zhang, J. Wang, and J. Caverlee, "Popularity-opportunity bias in collaborative filtering," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021, pp. 85–93.

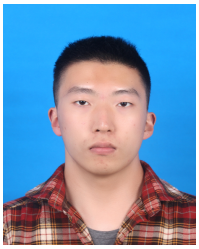
[15] S. Rendle, "Factorization machines with libfm," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, pp. 1–22, 2012.

[16] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th International Conference on World Wide Web*. ACM, 2017, pp. 173–182.

[17] M. B. Abdullah, "On a robust correlation coefficient," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 39, no. 4, pp. 455–460, 1990.

[18] R. Cañamares and P. Castells, "Should i follow the crowd? a probabilistic analysis of the effectiveness of popularity in recommender systems," in *The 41st International ACM SIGIR Conference*

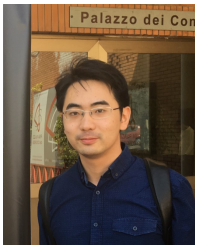
on Research & Development in Information Retrieval, 2018, pp. 415–424.
[19] C. Anderson, *The long tail: Why the future of business is selling less of more*. Hachette Books, 2006.
[20] M. Mansoury, H. Abdollahpouri, M. Pechenizkiy, B. Mobasher, and R. Burke, "Feedback loop and bias amplification in recommender systems," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 2145–2148.
[21] E. Brynjolfsson, Y. J. Hu, and M. D. Smith, "From niches to riches: Anatomy of the long tail," *Sloan management review*, vol. 47, no. 4, pp. 67–71, 2006.
[22] Ò. Celma and P. Cano, "From hits to niches? or how popular artists can bias music recommendation and discovery," in *Proceedings of the 2nd KDD Workshop on Large-Scale Recommender Systems and the Netflix Prize Competition*, 2008, pp. 1–8.
[23] Y.-J. Park and A. Tuzhilin, "The long tail of recommender systems and how to leverage it," in *Proceedings of the 2008 ACM conference on Recommender systems*, 2008, pp. 11–18.
[24] H. Abdollahpouri and M. Mansoury, "Multi-sided exposure bias in recommendation," *arXiv preprint arXiv:2006.15772*, 2020.
[25] Y. Ge, S. Zhao, H. Zhou, C. Pei, F. Sun, W. Ou, and Y. Zhang, "Understanding echo chambers in e-commerce recommender systems," in *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, 2020, pp. 2261–2270.
[26] Q. Lu, T. Chen, W. Zhang, D. Yang, and Y. Yu, "Serendipitous personalized ranking for top-n recommendation," in *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, vol. 1. IEEE, 2012, pp. 258–265.
[27] A. J. Chaney, B. M. Stewart, and B. E. Engelhardt, "How algorithmic confounding in recommendation systems increases homogeneity and decreases utility," in *Proceedings of the 12th ACM Conference on Recommender Systems*, 2018, pp. 224–232.
[28] H. Abdollahpouri, M. Mansoury, R. Burke, and B. Mobasher, "The connection between popularity bias, calibration, and fairness in recommendation," in *Fourteenth ACM Conference on Recommender Systems*, 2020, pp. 726–731.
[29] ——, "The impact of popularity bias on fairness and calibration in recommendation," *CoRR*, vol. abs/1910.05755, 2019.
[30] ——, "The unfairness of popularity bias in recommendation," in *Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019), Copenhagen, Denmark, September 20, 2019*, ser. CEUR Workshop Proceedings, R. Burke, H. Abdollahpouri, E. C. Malthouse, K. P. Thai, and Y. Zhang, Eds., vol. 2440. CEUR-WS.org, 2019.
[31] A. Gruson, P. Chandar, C. Charbuillet, J. McInerney, S. Hansen, D. Tardieu, and B. Carterette, "Offline evaluation to make decisions about playlistrecommendation algorithms," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 420–428.
[32] Z. Chen, R. Xiao, C. Li, G. Ye, H. Sun, and H. Deng, "Esam: Discriminative domain adaptation with non-displayed items to improve long-tail performance," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 579–588.
[33] S. Bonner and F. Vasile, "Causal embeddings for recommendation," in *Proceedings of the 12th ACM conference on recommender systems*, 2018, pp. 104–112.
[34] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Correcting popularity bias by enhancing recommendation neutrality." in *RecSys Posters*, 2014.
[35] J. Oh, S. Park, H. Yu, M. Song, and S.-T. Park, "Novel recommendation based on personal popularity tendency," in *2011 IEEE 11th International Conference on Data Mining*. IEEE, 2011, pp. 507–516.
[36] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, and X. He, "Bias and debias in recommender system: A survey and future directions," *arXiv preprint arXiv:2010.03240*, 2020.
[37] B. M. Marlin and R. S. Zemel, "Collaborative prediction and ranking with non-random missing data," in *Proceedings of the third ACM conference on Recommender systems*. ACM, 2009, pp. 5–12.
[38] J. Chen, C. Wang, M. Ester, Q. Shi, Y. Feng, and C. Chen, "Social recommendation with missing not at random data," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 29–38.
[39] J. M. Hernández-Lobato, N. Houlsby, and Z. Ghahramani, "Probabilistic matrix factorization with non-random missing data." in *ICML*, 2014, pp. 1512–1520.
[40] D. Liang, L. Charlin, J. McInerney, and D. M. Blei, "Modeling user exposure in recommendation," in *Proceedings of the 25th International Conference on World Wide Web*. ACM, 2016, pp. 951–961.
[41] J. Chen, Y. Feng, M. Ester, S. Zhou, C. Chen, and C. Wang, "Modeling users' exposure with social knowledge influence and consumption influence for recommendation," in *Proceedings of the 27th ACM International on Conference on Information and Knowledge Management*. ACM, 2018, pp. 953–962.
[42] Z. Ovaisi, R. Ahsan, Y. Zhang, K. Vasilaky, and E. Zheleva, "Correcting for selection bias in learning-to-rank systems," in *Proceedings of The Web Conference 2020*, 2020, pp. 1863–1873.
[43] M. D. Ekstrand and D. Kluver, "Exploring author gender in book rating and recommendation," *User modeling and user-adapted interaction*, pp. 1–44, 2021.
[44] A.-A. Stoica, C. Riederer, and A. Chaintreau, "Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 923–932.
[45] J. Ma, C. Zhou, P. Cui, H. Yang, and W. Zhu, "Learning disentangled representations for recommendation," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 5711–5722.
[46] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014.
[47] X. Wang, H. Jin, A. Zhang, X. He, T. Xu, and T.-S. Chua, "Disentangled graph collaborative filtering," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1001–1010.



Zihao Zhao is currently a Master Degree student with the School of Information Science and Technology, University of Science and Technology of China. He received his bachelor degree from University of Science and Technology of China in 2020. He has been awarded 2019 National Encouragement scholarship. His research interest includes recommendation system and graph representation learning.

Jiawei Chen is a Research Fellow in School of Computer Science, Zhejiang University. He received Ph.D. in Computer Science from Zhejiang University in 2020. His research interests include information retrieval, data mining, and causal inference. He has published over ten academic papers on international conferences and journals such as WWW, AAAI, SIGIR, CIKM, ICDE and TOIS. Moreover, he has served as the PC/SPC member for top-tier conferences including SIGIR, WWW, WSDM, ACMMM, AAAI, IJCAI and the invited reviewer for prestigious journals such as TNNLS, TKDE, TOIS.

Sheng Zhou is currently working as an assisant professor in College of Software and Engineering, Zhejiang University. He received the Ph.D degree with the College of Computer Science and Technology, Zhejiang University, No. 38 Zheda Road, Hangzhou, Zhejiang, China. He is working with Zhejiang Provincial Key Laboratory of Service Robot, College of Computer Science, Zhejiang University. His current research interests include Data mining, Graph Neural Networks and Knowledge distillation.



Xiangnan He is a professor at the University of Science and Technology of China (USTC). He received his Ph.D. in Computer Science from the National University of Singapore (NUS). His research interests span information retrieval, data mining, and multi-media analytics. He has over 100 publications that appeared in top conferences such as SIGIR, WWW, and KDD, and journals including TKDE, TOIS, and TNNLS. His work has received the Best Paper Award Honorable Mention in WWW 2018 and ACM SIGIR 2016. He serves as the associate editor for ACM Transactions on Information Systems (TOIS), Frontiers in Big Data, AI Open etc. Moreover, he has served as the PC chair of CCIS 2019 and SPC/PC member for several top conferences including SIGIR, WWW, KDD, MM, WSDM, ICML etc., and the regular reviewer for journals including TKDE, TOIS, etc.

Xuezhi Cao is a senior researcher in Meituan's NLP team. He obtained his Ph.D. degree from Shanghai Jiao Tong University in 2018. His research interests include knowledge graph, recommender system, and social network. He has over 15 publications in top conferences including SIGIR, WWW, AAAI, etc.

Fuzheng Zhang is now a technical leader in Meituan. He is leading a team focus on AI technologies such as knowledge graph, NLP, and information retrieval. Before this, Dr. Zhang was a researcher in Microsoft Research Asia. He obtained his Ph.D. degree in computer science, and is supervised jointly by University of Science and Technology of China and Microsoft Research Asia. He has published over 50 top-tier international conference papers and journal articles including KDD, WWW, AAAI, and IJCAI. He has received the best paper award in ICDM2013 and CIKM2020. He has long served as the reviewers on top-tier international conferences and journals, such as KDD, WWW, and TKDE.

Wei Wu is now a technical leader in Meituan. He is leading a team focus on AI technologies such as knowledge graph, NLP, and information retrieval. Before this, Dr. Zhang was a researcher in Microsoft Research Asia. He obtained his Ph.D. degree in computer science, and is supervised jointly by University of Science and Technology of China and Microsoft Research Asia. He has published over 50 top-tier international conference papers and journal articles including KDD, WWW, AAAI, and IJCAI. He has received the best paper award in ICDM2013 and CIKM2020. He has long served as the reviewers on top-tier international conferences and journals, such as KDD, WWW, and TKDE.