Disentangling User Interest and Conformity for Recommendation with Causal Embedding

Yu Zheng[1,2], Chen Gao[1,2], Xiang Li[3], Xiangnan He[4], Depeng Jin[1,2], Yong Li[1,2†]
[1]Beijing National Research Center for Information Science and Technology
[2]Department of Electronic Engineering, Tsinghua University
[3]University of Hong Kong
[4]University of Science and Technology of China
†Corresponding Author: liyong07@tsinghua.edu.cn
China

ABSTRACT

Recommendation models are usually trained on observational interaction data. However, observational interaction data could result from users' conformity towards popular items, which entangles users' real interest. Existing methods tracks this problem as eliminating popularity bias, e.g., by re-weighting training samples or leveraging a small fraction of unbiased data. However, the variety of user conformity is ignored by these approaches, and different causes of an interaction are bundled together as unified representations, hence robustness and interpretability are not guaranteed when underlying causes are changing. In this paper, we present DICE, a general framework that learns representations where interest and conformity are structurally disentangled, and various backbone recommendation models could be smoothly integrated. We assign users and items with separate embeddings for interest and conformity, and make each embedding capture only one cause by training with cause-specific data which is obtained according to the colliding effect of causal inference. Our proposed methodology outperforms state-of-the-art baselines with remarkable improvements on two real-world datasets on top of various backbone models. We further demonstrate that the learned embeddings successfully capture the desired causes, and show that DICE guarantees the robustness and interpretability of recommendation.

CCS CONCEPTS

• **Information systems** → **Collaborative filtering**.

KEYWORDS

Recommender systems, popularity bias, causal embedding

ACM Reference Format:
Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Depeng Jin, Yong Li. 2021. Disentangling User Interest and Conformity for Recommendation with Causal Embedding. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442381.3449788

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '21, April 19–23, 2021, Ljubljana, Slovenia
© 2021 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3449788

1 INTRODUCTION

Recent years have witnessed great success in recommender systems, which provide users with personalized contents by mining user preference from the observational interaction data [36]. However, observational interaction data exhibits strong popularity bias [12], which entangles users' real interest. A user might click an item simply because many other users have clicked it, e.g. in e-commerce platforms items are often displayed with their sales values. In fact, those interactions are mainly driven by users' *conformity*, rather than real *interest*. As a crucial factor for decision making, conformity describes how users tend to follow other people. Meanwhile, conformity towards an item varies according to different users. In order to capture users' pure interest that is independent with conformity, existing approaches track this problem as eliminating popularity bias, a static and global term from the perspective of items, while ignoring the variety of users' conformity. For example, a sport lover purchases a bicycle with high sales value due to his unique tastes on specific characteristics (e.g. tire size or speed capacity), while an office worker might purchase the same bicycle only because of its high sales. Using uniform popularity bias fails to distinguish these two users' different conformity, since popularity score of an item will be the same for all users. Therefore, disentangling user interest and conformity is crucial enhance recommendation quality.

In this work, we take a different approach from user's perspective. Instead of eliminating popularity bias from the perspective of **items**, we propose to decompose the observed interactions into two factors in the **user** side, interest and conformity, and learn disentangled representations for them. Disentangling these two factors is challenging and has not been well explored. Specifically, we face three key challenges. First, conformity depends on both user and item. One user's conformity varies on different items, as well as conformity towards one item from different users. Thus, a scalar bias term for user or item is insufficient, as adopted by existing algorithms [4]. Second, learning disentangled representations is intrinsically difficult, especially when only observational interaction data is available. In other words, we only have access to the *effect*, but not the *causes*, since there is no labeled ground-truth value for interest and conformity. Third, a click interaction can come from one or both causes of interest and conformity. Therefore, it requires careful design to aggregate and balance the two causes.

Although it is a challenging task, learning disentangled representations of interest and conformity has two main advantages over approaches that only learn a unified embedding for a user or item:

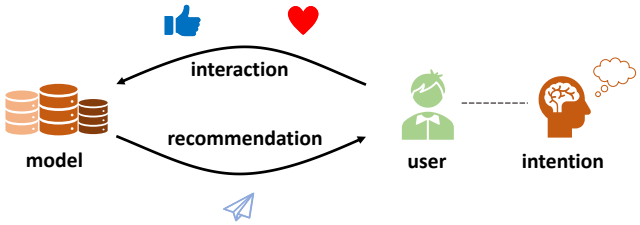


Figure 1: Feedback loop of recommender system. Users interact with the model according to user intention, and the model is trained with users' interaction data.

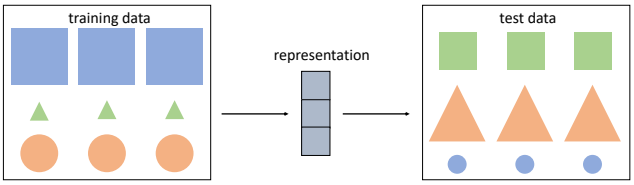


Figure 2: Shape recognition under non-IID circumstance. A model can predict shapes from size or color in training data, since rectangles are blue and large, triangles are green and small, circles are orange and medium size. However, under non-IID circumstances as in test data, color and size are different from training data, hence only models that disentangle underlying factors can survive in this example.

(1) **Robustness**. Real-world recommender systems are often trained and updated continuously using real-time user interactions, which forms a feedback loop as illustrated in Figure 1, with training data and test data NOT independent and identically distributed (IID) [13]. Causal modeling on the effect (click) and cause (interest and conformity) can lead to more robust models, with stronger generalization ability, especially in non-IID situations where underlying causes are changing [41].
(2) **Interpretability**. Interpretable recommendation can benefit both users and platforms of recommender systems, since it improves user-friendliness and facilitates algorithm developing. By disentangling underlying causes, each recommendation score is decomposed as an aggregation of interest score and conformity score. Therefore, explanations towards the two causes can be easily made according to corresponding scores.

In this paper, we present a general framework for **D**isentangling **I**nterest and **C**onformity with **C**ausal **E**mbedding (DICE). To capture the variety of conformity, we propose to learn comprehensive embeddings separately for conformity, which is independent with user interest. Instead of using simple scalar popularity value as existing approach, we develop particular methodologies in order to learn disentangled representations for interest and conformity. Specifically, we describe the causal model of how each interaction data is generated. Based on the causal model, we propose particular negative sampling strategies for specific causes based on the *colliding effect* of causal inference [37, 38], and learn separate embeddings for interest and conformity with cause-specific data. Meanwhile, we add direct supervision on the disentanglement between the two parts of embeddings. To generate final recommendation considering both user interest and conformity, we exploit multi-task and curriculum learning, which successfully balances the two causes.

We evaluate the proposed method on two large-scale benchmark datasets collected from real-world applications. Experimental results show that DICE outperforms state-of-the-art baselines with over 15% improvements in terms of Recall and NDCG. To investigate the robustness of DICE, we extract test data that is non-IID with training data by conducting *intervention* on conformity. We demonstrate that DICE consistently beats baseline methods under non-IID situations. Furthermore, we provide analytical results on the quality of learned embeddings, which illustrate superior interpretability of the proposed method.

In summary, the main contributions of this paper are as follows:

- To the best of our knowledge, this is the first work to formulate the problem of disentangling user interest and conformity for recommender systems. We tackle the causal recommendation problem from the perspective of users and show that disentangling these two factors is essential for recommender systems, with respect to robustness and interpretability.
- We propose a general framework to disentangle interest and conformity. Separate embeddings are adopted to capture the two causes, and different embeddings are trained with cause-specific data, forced to capture only one desired cause. Moreover, we exploit multi-task learning and curriculum learning to balance the two causes.
- Extensive experiments are conducted on two large-scale datasets of real-world recommender systems. Results show that DICE achieves significant improvements over state-of-the-art baseline models. Further analysis demonstrates that DICE shows great robustness under non-IID circumstances, and high interpretability of the learned embeddings is guaranteed in DICE as well.

The remainder of this paper is as follows. We first introduce the motivation and formulate the problem in Section 2. We then elaborate the proposed DICE framework in Section 3. We conduct experiments in Section 4, after that we discuss related works in Section 5. Finally, we conclude this paper in Section 6.

2 MOTIVATION AND PROBLEM OVERVIEW

Motivation. Algorithms that disentangle underlying semantics have superior generalization ability over *entangled* approaches. Here we focus on a specific form of generalization ability that is not from one data point to another data point in the same distribution, but from one distribution to another distribution. Figure 2 shows an example of shape recognition, which follows the non-IID condition on training data and test data. Suppose we are developing a shape recognition model, where we learn representations from original pictures and predict their shapes based on the learned representations. It seems like a normal task, however there are traps in it. In fact, models are easily misled by training data, since rectangles are blue and large, triangles are green and small, circles are orange and medium size. As a consequence, a model can predict shapes from color or size, rather than outline. Moreover, if test data is generated from the same distribution (*i.e.* IID with training data), bad models that pay attention to color or size would perform well on test set and we might not even notice what was going wrong. Fortunately, we force training data and test data NOT to be IID as shown in Figure 2 where color and size are totally different with training

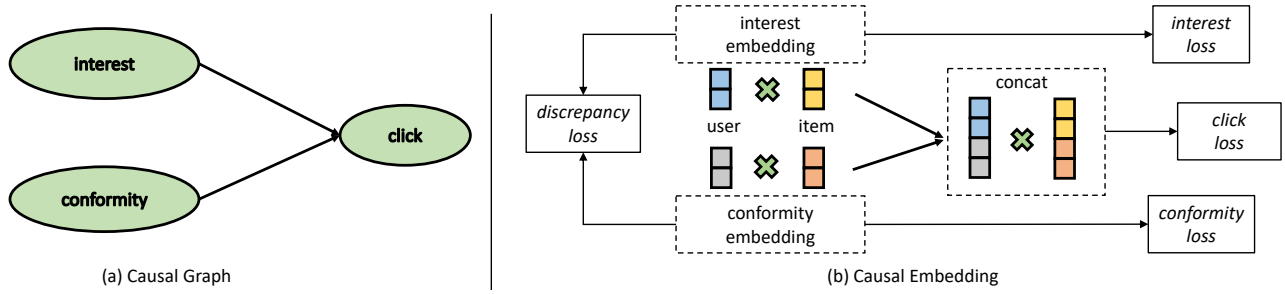


Figure 3: Causal graph and causal embeddings. (a) We make concise causal modeling on each click that it results from two independent causes, interest and conformity. (b) We adopt separate embeddings for interest and conformity, thus each user or item has two embeddings. We force each embedding to capture only one cause by training different embeddings with cause-specific data and adding direct disentanglement supervision, under the framework of multi-task curriculum learning.

data, and evaluate whether our models are robust under this intervened environment. Then only those models that disentangle the underlying semantics (shape, color and size) can survive in our test.

With respect to recommender systems, a click interaction can be triggered by users' real interest, or their conformity towards popular items. In IID situations, it is not necessary for models to distinguish between users' interest and conformity, thus models tend to recommend items according to their popularity values, due to their larger amount of training instances. However, users' conformity at training time and serving time are distinct, since recommender system is a live interactive system shown in Figure 1. Therefore, it is essential for recommendation algorithms to be robust in such non-IID situation, especially when underlying causes are different. In this work, we extend conventional causal recommendation algorithms that perform unbiased learning from biased data [26], and propose to disentangle user interest and conformity. Based on recent advancements in causal recommendation [9, 30], we construct datasets with training data and test data NOT IID. We evaluate the proposed methodology compared with state-of-the-art baseline approaches, and particularly investigate their robustness under non-IID circumstances by interventions.

Problem Formulation. Here we formulate the problem of *disentangling user interest and conformity*. Suppose the dataset O is composed of N instances of (u, i, p), where p is the popularity of item i, *i.e.* the number of interactions on item i. The distribution of p serves as a proxy for conformity distribution. We first construct intervened test set O_{test} and normal training set O_{train}, where $D_p^{O_{test}}$, the distribution of item popularity p in test set, is distinct from that in training set, $D_p^{O_{train}}$. Our goal is to maximize recommendation performance $\mathcal{R}$, like recall and NDCG, on O_{test}, with models trained on O_{train} that is NOT IID with O_{test}:

Input: Observational interaction data O, which is splitted into O_{train} and O_{test}, with non-IID conditions on popularity distribution $D_p^{O_{train}}$ and $D_p^{O_{test}}$.

Output: A predictive model estimating whether a user will click an item considering both interest and conformity.

3 DICE: THE PROPOSED APPROACH

We propose a general framework, named DICE, to learn disentangled representations for interest and conformity. Figure 3 illustrates the holistic design of DICE. To tackle the previously introduced three challenges, our proposed framework is composed of the following three stages:

- **Causal Embedding**:We propose to utilize separate embeddings instead of scalar values for interest and conformity, to solve the problem of varying conformity.
- **Disentangled Representation Learning**: In order to learn disentangled representations for interest and conformity, we divide training data into cause-specific parts, and train different embeddings with cause-specific data. Direct supervision on embedding distribution is added to reinforce disentanglement.
- **Multi-task Curriculum Learning**: Finally, we develop an easy-to-hard training strategy and exploit curriculum learning to aggregate and balance interest and conformity.

3.1 Causal Embedding

In this section, we first describe the causal model of how each interaction data is generated from interest and conformity. Then we provide the structural causal model (SCM) and causal graph for click, interest and conformity, based on which we propose to utilize separate embeddings for interest and conformity, solving the first challenge of varying conformity.

How each interaction data is generated? A click record of a user on an item mainly reflects two aspects: (1) the user's interest in the item's characteristics, (2) the user's conformity towards the item's popularity. A click can come from one or both of the two aspects. We propose an additive model to describe how each click record is generated from interest and conformity. Formally, the matching score of a given user u and item i is attained as follow:

$$S_{\text{ui}} = S_{ui}^{\text{interest}} + S_{ui}^{\text{conformity}}, \tag{1}$$

where S_{ui} represents the overall matching score, while S_{ui}^{interest} and $S_{ui}^{\text{conformity}}$ stand for a specific cause. This additive model is justified because users tend to have both particularity and conformity when interacting with recommender systems [31]. Meanwhile, additive models are widely adopted in causal inference and have been shown effective in a bunch of applications [38]. In addition, multiplicative model is also adopted in related literature [49], which decomposes the click probability as production of exposure probability and the conditional click probability given exposure. However, such multiplicative model entangles interest and conformity from the perspective of user, since users' conformity still takes effect given the

exposed items. It is worthwhile to notice that there may be causes other than interest or conformity that lead to a click interaction, but we propose to grasp these two principle factors. Meanwhile, the proposed methodology is a general framework which can be extended to scenarios with multiple causes.

SCM and causal graph for click, interest and conformity. Based on our proposed causal model in (1), we now provide the SCM of our proposed DICE framework, ζ_{DICE}, along with causal graph in Figure 3(a):

$$\begin{aligned} X_{ui}^{int} &:= f_1(u, i, N^{int}), \\ X_{ui}^{con} &:= f_2(u, i, N^{con}), \\ Y_{ui}^{click} &:= f_3(X_{ui}^{int}, X_{ui}^{con}, N^{click}), \end{aligned} \tag{2}$$

where N^{int}, N^{con} and N^{click} are independent noises. SCM ζ_{DICE} expresses causal dependence of interest, conformity and click, where f_1, f_2 and f_3 are the underlying causal mechanisms for interest X^{int}, conformity X^{con}, and click Y^{click} respectively. Practically, those causal mechanisms are decided by optimizing within a given family of functions [38], such as deep neural networks. When we consider interventions on user conformity, we simply replace X_{ui}^{con} with pre-assigned values.

SCM ζ_{DICE} in (2) explains the logic on how effect (click) is generated from causes (interest and conformity). However, particular forms of function families in f_1, f_2 and f_3 are still to be determined. As introduced previously, conformity of different users towards the same item are diverse, and so does conformity of the same user towards different items. In other words, conformity depends on both users and items, as well as interest. Therefore, function families for f_1 and f_2 should better support such flexibility in interest and conformity. We now introduce our proposed design using separate embeddings.

Separate embeddings for interest and conformity In the proposed DICE framework, we adopt two sets of embeddings to separately capture interest and conformity, instead of using scalar popularity values as in existing approach [4], since scalar values are insufficient to capture the diversity of user conformity. As shown in Figure 3(b), each user has an interest embedding $\boldsymbol{u}^{(\text{int})}$ and a conformity embedding $\boldsymbol{u}^{(\text{con})}$, and each item also has $\boldsymbol{i}^{(\text{int})}$ and $\boldsymbol{i}^{(\text{con})}$ for the two causes[1]. We use inner product to compute matching score for both causes. Based on the additive causal model in Equation (1), we sum up the two matching scores from corresponding causes, to estimate the overall score on whether a user will click an item. Therefore, the recommendation score for user u and item i is formulated as:

$$\begin{aligned} s_{ui}^{int} &= \langle \boldsymbol{u}^{(\text{int})}, \boldsymbol{i}^{(\text{int})} \rangle, \quad s_{ui}^{con} = \langle \boldsymbol{u}^{(\text{con})}, \boldsymbol{i}^{(\text{con})} \rangle, \\ s_{ui}^{click} &= s_{ui}^{int} + s_{ui}^{con}, \end{aligned} \tag{3}$$

where $\langle \cdot, \cdot \rangle$ means inner product of two embeddings. Figure 3(b) demonstrates the disentanglement design of interest embeddings and conformity embeddings. From the perspective of SCM, we restrict the family of functions for f_1 and f_2 as inner product between two sets of learnable embeddings, and set f_3 as a concise additive model which is commonly used in practice [38]. By optimizing in two high-dimensional spaces rather than finding the optimal scalar value in 1-D space like existing solutions, diversity of user conformity can be automatically captured in the proposed DICE framework, hence we solve the first challenge.

[1]Formally, users should have interest and conformity embeddings, and items should have characteristic and popularity embeddings. We overload the usage of interest and conformity on items to simplify expressions and avoid confusions.

3.2 Disentangled Representation Learning

In this section, we elaborate our designs on disentangling the two causal embeddings for interest and conformity. We propose to train different embeddings with cause-specific data, and decompose the problem into four tasks of conformity modeling, interest modeling, estimating clicks and an extra discrepancy task.

Mining Cause-specific Data. Disentanglement between interest embedding and conformity embedding means that each embedding captures only one factor, and squeezes out the second factor to the other embedding. To achieve such target, a common and reasonable idea is to train different embeddings with cause-specific data. However, we only have access to the *effect*, which is the observational click data, but we are nearly ignorant on whether the click is caused by interest or conformity. In other words, an equality in (1) is insufficient to recover interest and conformity, because there are infinite solutions to it when there are no ground-truth for two addends and only the summation is available. Therefore, we explore from observational interactions and discover cause-specific data which means those interactions are sourced from individual cause with high probability. The cause-specific data paves the way for disentangling the two underlying causes, interest and conformity.

We first introduce several notations. We use M^I to denote the matrix of interest matching score for all users and items, and M^C for conformity matching score. M^I and M^C are in $\mathbb{R}^{M \times N}$, where M and N are the number of users and items. In the causal graph in Figure 3(a), the three nodes form a *immorality*, and click is the *collider* of interest and conformity [37, 38]. In fact, the two causes of a collider are independent variables. However, if we condition on the collider, the two causes become correlated with each other, and we call it *colliding effect*. For example, whether a man is popular depends on both his appearance and temper. Appearance and temper are usually independent, and popularity is the collider of appearance and temper (appearance → popularity ← temper). Given a popular man who is not good-looking, then he is good-tempered with high probability. Similarly, an unpopular but good-looking man is most likely bad-tempered. Therefore, in our task of disentangling interest and conformity, the colliding effect can be utilized to obtain samples that are mostly resulted from one cause. Specifically, we emphasize on two particular cases in M^I and M^C which are cause-specific:

Case 1: The negative item is less popular than the positive item. If a user u clicks a popular item a, while does not click an unpopular item b, then we are not sure whether the user's interest on a is stronger than b, because users have conformity towards popular items. In other words, the click can come from the second cause (conformity). Meanwhile, we can also safely conclude that the overall score of the two causes of a is larger than b. Hence we have two inequalities in this case:

$$\begin{aligned} &M_{ua}^C > M_{ub}^C, \\ &M_{ua}^I + M_{ua}^C > M_{ub}^I + M_{ub}^C. \end{aligned} \tag{4}$$

Case 2: The negative item is more popular than the positive item. However, if a user clicks an unpopular item c, while does not click a popular item d, then the colliding effect can bring more information. Since c is less popular than d which serves as a reasonable proxy for conformity, the click on c is largely due to the user's interest. Therefore, we have three inequalities in this case, with one extra inequality on interest than the previous case:

$$\begin{aligned} &M^I_{uc} > M^I_{ud}, M^C_{uc} < M^C_{ud}, \\ &M^I_{uc} + M^C_{uc} > M^I_{ud} + M^C_{ud}. \end{aligned} \tag{5}$$

We use O to denote all training instances, which is divided into O_1 and O_2. Specifically, O_1 denotes those instances where negative samples are less popular than positive samples, and O_2 denotes the opposite cases. Correspondingly, O_1 contains the data that inequalities of (4) in *Case 1* hold true, thus O_1 can be utilized to learn conformity and click. O_2 contains the data that fits *Case 2*, hence it can be exploited to learn interest, conformity and click.

By extending one equality to multiple inequalities, we transform the problem from learning absolute values to learning relative relations, which makes the task of disentangling interest and conformity solvable. Specifically, based on these derived inequalities, we obtain user-item interactions that mainly result from one specific cause, and leverage these interactions to optimize corresponding embeddings. Take the famous matrix factorization algorithm for recommendation as an example, usually we optimize a user embedding matrix and an item embedding matrix to best regress the original interaction matrix, say M^{click}. This classical approach unifies all possible causes into one bundled representation for a user or item, thus different causes are entangled, leading to inferior robustness and interpretability under non-IID circumstances, which is quite common in recommender systems. Moreover, debias algorithms such as IPS can not fully solve this problem, since they still adopt unified representations. In contrast to existing approaches, we first decompose the original click matrix M^{click} into two cause-specific matrices, M^I and M^C, for interest and conformity respectively. Then two sets of embeddings are adopted, in order to capture interest and conformity separately, and they are further combined to regress click. Different causes are thus disentangled, which achieves better robustness under interventions. We now introduce our causal learning methodology.

With cause-specific data O_1 and O_2, it is possible to model interest and conformity separately. Meanwhile, we propose to estimate click behaviors by combining the two causes, which is the main task for recommendation. Moreover, we further add a discrepancy task in order to make the two sets of embeddings independent with each other, which enhances disentanglement. Therefore, we decompose the problem of disentangling interest and conformity into four tasks, which are conformity modeling, interest modeling, estimating clicks and discrepancy task. We utilize BPR [39] to model the pairwise quantitative relations in (4) and (5). Each positive sample is paired with certain number of negative samples, and each training instance is a triplet (u, i, j) containing user ID, positive item ID and negative item ID. We now introduce the four tasks in sequence.

Conformity Modeling For instances in both O_1 and O_2, we have inequalities for conformity modeling, which are the inequalities for M^C. Notice that the directions of inequality are different in the two cases. We use these conformity-specific data to optimize conformity embeddings. BPR loss function is exploited to regress M^C with conformity embeddings. Therefore, the loss function for conformity modeling is formulated as:

$$\begin{aligned} L^{O_1}_{\text{conformity}} &= \sum_{(u,i,j)\in O_1} \text{BPR}(\langle \boldsymbol{u}^{(\text{con})}, \boldsymbol{i}^{(\text{con})} \rangle, \langle \boldsymbol{u}^{(\text{con})}, \boldsymbol{j}^{(\text{con})} \rangle), \\ L^{O_2}_{\text{conformity}} &= \sum_{(u,i,j)\in O_2} -\text{BPR}(\langle \boldsymbol{u}^{(\text{con})}, \boldsymbol{i}^{(\text{con})} \rangle, \langle \boldsymbol{u}^{(\text{con})}, \boldsymbol{j}^{(\text{con})} \rangle), \\ L^{O_1+O_2}_{\text{conformity}} &= L^{O_1}_{\text{conformity}} + L^{O_2}_{\text{conformity}}. \end{aligned} \tag{6}$$

Interest Modeling In O_2, negative items are more popular than positive items, and those interactions are largely due to users' interest. These data is interest-specific, and we have inequalities for interest modeling. We also use BPR to optimize interest embeddings to learn such pairwise preference, in order to regress M^I. The loss function only takes effect for instances in O_2:

$$L^{O_2}_{\text{interest}} = \sum_{(u,i,j)\in O_2} \text{BPR}(\langle \boldsymbol{u}^{(\text{int})}, \boldsymbol{i}^{(\text{int})} \rangle, \langle \boldsymbol{u}^{(\text{int})}, \boldsymbol{j}^{(\text{int})} \rangle). \tag{7}$$

Estimating Clicks This is the main target for recommender systems, and we combine the two causes to estimate clicks as introduced in (3), with a concise additive model. For each instance in training set O, which is the union of O_1 and O_2, we use BPR to maximize the margin between scores of positive items and negative items, so as to regress M^{click}. The loss function for click estimation is thus formulated as follow:

$$L^{O_1+O_2}_{\text{click}} = \sum_{(u,i,j)\in O} \text{BPR}(\langle \boldsymbol{u}^t, \boldsymbol{i}^t \rangle, \langle \boldsymbol{u}^t, \boldsymbol{j}^t \rangle). \tag{8}$$

$\boldsymbol{u}^t$, $\boldsymbol{i}^t$ and $\boldsymbol{j}^t$ are concatenation of interest embedding and conformity embedding for user and item:

$$\boldsymbol{u}^t = \boldsymbol{u}^{(\text{int})} \| \boldsymbol{u}^{(\text{con})}, \boldsymbol{i}^t = \boldsymbol{i}^{(\text{int})} \| \boldsymbol{i}^{(\text{con})}, \boldsymbol{j}^t = \boldsymbol{j}^{(\text{int})} \| \boldsymbol{j}^{(\text{con})}, \tag{9}$$

where $\|$ means concatenation of two embeddings. We use the concatenation form here for simplicity, which is equivalent to the summation form in (3). The BPR loss pushes the recommendation score for the positive item i to be higher than the negative item j.

Interest modeling and conformity modeling disentangle the two causes by training different embeddings with different cause-specific data. Meanwhile, the main task on estimating clicks also strengthens this disentanglement as a constraint. For example, in terms of a training instance (u, i, j) where negative item j is more popular than positive item i, interest modeling task forces the two sets of embeddings to learn that user u's interest in i is larger than j, and conformity modeling task forces them to learn that user u's conformity towards item i' is less than j. Meanwhile, estimating clicks forces them to learn that the overall strength on i is larger than j. Therefore, what the model really learns is that the advantage of i over j with respect to interest dominates the disadvantage in conformity, which can be best learned by capturing only one cause with one embedding.

Discrepancy Task Besides the three tasks above that disentangle interest and conformity by optimizing different embeddings with cause-specific data, we impose direct supervision on the embedding distribution to reinforce this disentanglement. Suppose

$\mathbf{E}^{(\text{int})}$ and $\mathbf{E}^{(\text{con})}$ represent two sets of embeddings of all users and items. We examine three candidate discrepancy loss functions, which are L1-inv, L2-inv and distance correlation (*dCor*). L1-inv and L2-inv *maximize* L1 and L2 distances between $\mathbf{E}^{(\text{int})}$ and $\mathbf{E}^{(\text{con})}$ respectively. We refer to [45, 46] for details on *dCor*. From high level, *dCor* is a more reasonable choice, since it focuses on the correlations of pairwise distances between interest embeddings and conformity embeddings. The three options for discrepancy loss function are $-L1(\mathbf{E}^{(\text{int})}, \mathbf{E}^{(\text{con})})$, $-L2(\mathbf{E}^{(\text{int})}, \mathbf{E}^{(\text{con})})$ and $dCor(\mathbf{E}^{(\text{int})}, \mathbf{E}^{(\text{con})})$. We will compare them in experiments.

Figure 3(b) illustrates the four decomposed tasks using disentangled embeddings for interest and conformity. By training different embeddings with cause-specific data and imposing direct supervision on the embedding distribution, we solve the second challenge of learning disentangled representations.

3.3 Multi-task Curriculum Learning

In the proposed framework, we overcome the last challenge of aggregating interest and conformity by multi-task curriculum learning. To be specific, we train causal embeddings with the four above tasks simultaneously, and combine these loss functions together:

$$L = L_{\text{click}}^{O_1+O_2} + \alpha(L_{\text{interest}}^{O_2} + L_{\text{conformity}}^{O_1+O_2}) + \beta L_{\text{discrepancy}}. \quad (10)$$

Since estimating clicks is the main task for recommendation, α and β should be less than 1 from intuition. Meanwhile, discrepancy task directly influences the distribution of embeddings, thus too large β would negatively impact interest and conformity modeling.

As introduced previously, we obtain two or three inequalities when the negative sample is less or more popular than the positive sample, respectively. Notice that those inequalities will hold true with high probability when the popularity gap is sufficiently large. Therefore, we develop **P**opularity based **N**egative **S**ampling with **M**argin (PNSM) to guarantee those quantitative relations. Specifically, if the popularity of the positive sample is p, then we will sample negative instances from items with popularity larger than $p + m_{up}$, or lower than $p - m_{down}$, where m_{up} and m_{down} are positive margin values. By sampling negative items with popularity margin, we gain high confidence in our causal models. Later experiments show that popularity based negative sampling with margin is of crucial importance for learning disentangled and robust representations.

Inspired by curriculum learning [7], we adopt an easy-to-hard strategy on training DICE by adding decay on margin values and loss weights. Specifically, when margin values m_{up} and m_{down} are large, we have high confidence on those inequalities for interest and conformity modeling, which means the tasks are *easier* and we set high loss weights α for L_{interest} and $L_{\text{conformity}}$. As we train the model, we increase the difficulty by decaying margin values, as well as loss weights α, by a factor of 0.9 after each epoch. With curriculum learning, the proposed approach learns stronger disentanglement for high-confidence samples. Furthermore, this adaptive design also makes the proposed method not sensitive to initial values of hyper-parameters. We will compare the performance of curriculum learning with normal learning in experiments. Interest and conformity are elegantly aggregated by multi-task curriculum learning, hence the last challenge is addressed.

Table 1: Statistics of datasets. (Ent. stands for entropy value of the number of interactions for all items. Larger entropy value of test data shows the non-IID condition.

Dataset	User	Item	Interaction	Ent. Train	Ent. Test
Movielens-10M	37962	4819	1371473	6.22	7.97
Netflix	32450	8432	2212690	6.85	8.54

In summary, we propose an additive causal model on user interest and conformity. Based on SCM ζ_{DICE}, we develop separate causal embeddings for individual cause, which capture the diversity of conformity and interest. A bunch of inequalities are derived from our causal model, decomposing the causal learning task into conformity modeling, interest modeling, estimating clicks and discrepancy task. Disentangled representations for underlying causes are obtained by training different embeddings with cause-specific data. To attain robust recommendation, multi-task curriculum learning is adopted to aggregate the two causes. Meanwhile, our causal framework is based on how data is generated and hence they are model-independent. Therefore, the proposed DICE methodology provides a highly general framework for disentangling user interest and conformity, which can be smoothly integrated into existing recommendation models. In our experiments, we successfully develop DICE on top of state-of-the-art recommender systems based on Graph Convolutional Networks.

4 EXPERIMENTS

In this section, we conduct experiments to show the effectiveness of the proposed framework. Specifically, we aim to answer the following research questions:

- **RQ1**: How does our proposed DICE framework perform compared with state-of-the-art causal recommendation methods under non-IID circumstances? Particularly, is it necessary to replace scalar bias term with embedding?
- **RQ2**: Can the proposed DICE framework guarantee interpretability and robustness?
- **RQ3**: What is the role of each component in the proposed methodology, including negative sampling, conformity modeling, curriculum learning, and discrepancy loss?
- **RQ4**: What is the effect of intervened data inserted into training set? How does DICE perform when no intervened training data is available?

4.1 Experimental Settings

Datasets We conduct experiments on two million-scale datasets collected from real-world applications, Movielens-10M dataset [20] and Netflix Prize dataset [8], and Table 1 lists the statistics of two datasets.

Data Preprocessing In order to measure the performance of causal learning under non-IID circumstances, intervened test sets are needed, and thus all datasets are transformed following the standard protocol introduced in related literatures [9, 30]. We binarize the datasets by keeping ratings of five stars as one, and others as zero. To conduct intervention on conformity, we randomly sample 40% of the records with equal probability in terms of items, and leave the other 60% as training data. In other words, items are sampled with probability as **inverse** popularity, which means

Table 2: Overall performance on Movielens-10M dataset and Netflix dataset.

Dataset		Movielens-10M						Netflix					
		TopK = 20			TopK = 50			TopK = 20			TopK = 50		
Model	Method	Recall	HR	NDCG	Recall	HR	NDCG	Recall	HR	NDCG	Recall	HR	NDCG
MF	None	0.1286	0.4429	0.0846	0.2346	0.6295	0.1170	0.1122	0.5194	0.0943	0.1928	0.6749	0.1185
	IPS	0.1335	0.4434	0.0852	0.2376	0.6288	0.1174	0.1058	0.4882	0.0864	0.1855	0.6562	0.1112
	IPS-C	0.1367	0.4564	0.0875	0.2429	0.6383	0.1203	0.1119	0.5046	0.0919	0.1938	0.6700	0.1174
	IPS-CN	0.1412	0.4700	0.0925	0.2509	0.6477	0.1264	0.1080	0.5042	0.0935	0.1912	0.6621	0.1185
	IPS-CNSR	0.1365	0.4588	0.0895	0.2419	0.6366	0.1219	0.1110	0.5159	0.0948	0.1937	0.6713	0.1192
	CausE	0.1157	0.4066	0.0744	0.2121	0.5924	0.1037	0.0935	0.4641	0.0782	0.1651	0.6272	0.0994
	DICE	**0.1634**	**0.5197**	**0.1084**	**0.2872**	**0.6975**	**0.1468**	**0.1258**	**0.5545**	**0.1070**	**0.2164**	**0.7090**	**0.1345**
GCN	None	0.1378	0.4625	0.0898	0.2513	0.6505	0.1247	0.1026	0.4908	0.0870	0.1842	0.6609	0.1112
	IPS	0.1394	0.4645	0.0919	0.2538	0.6473	0.1275	0.1101	0.5091	0.0950	0.1941	0.6657	0.1203
	IPS-C	0.1478	0.4829	0.0971	0.2654	0.6632	0.1339	0.1157	0.5219	0.1004	0.2037	0.6816	0.1270
	IPS-CN	0.1119	0.3997	0.0701	0.2281	0.6112	0.1057	0.0726	0.3991	0.0643	0.1472	0.5841	0.0866
	IPS-CNSR	0.1300	0.4427	0.0852	0.2336	0.6282	0.1171	0.0826	0.4337	0.0715	0.1589	0.6124	0.0940
	CausE	0.1027	0.3729	0.0632	0.2044	0.5811	0.0941	0.0838	0.4289	0.0677	0.1569	0.6119	0.0902
	DICE	**0.1812**	**0.5563**	**0.1228**	**0.3100**	**0.7216**	**0.1629**	**0.1420**	**0.5910**	**0.1217**	**0.2367**	**0.7340**	**0.1499**

popular items are less selected. Moreover, we cap the probability at 0.9 to limit the number of items that do not show in training set [9]. Finally, we obtain a 70/10/20 split for training set (60% normal and 20% intervened), validation set (10% intervened) and test set (20% intervened). Test data can be regarded as recommendation result under a fully random policy. As a consequence, conformity in test data is distinct from that in training data, since users have access to all items with equal probability in test data, rather than seeing more popular items in training data. We refer to [9, 30] for details on extracting an intervened test set from original interaction data. To show that training data and test data are non-IID, we count the number of interactions for each item and calculate the entropy, hence larger entropy value indicates that different items are of more equal probability to be exposed to users. As illustrated in Table 1, entropy on test data is much larger than that on training data for both datasets. In other words, models are trained on normal data, while evaluated on intervened data.

Recommendation Models Causal approaches usually serve as additional methods upon backbone recommendation models. We use the most adopted recommendation model, Matrix Factorization (MF) [29] to compare different approaches. Meanwhile, we also incorporate the state-of-the-art collaborate filtering model, Graph Convolutional Networks (GCN) [19, 21, 50], to investigate whether algorithms generalize across different recommendation models. Specifically, we use BPR-MF [39] and LightGCN [21], which are both state-of-the-art recommendation models.

Experiment Setups For IPS based models, we fix the embedding size as 128. While for CausE and DICE, the embedding size is fixed as 64, since they contain two sets of embeddings. Therefore, the number of parameters are the same for all methods to guarantee fair comparison. We set α as 0.1 and β as 0.01, which shows great performance and agnostic to both datasets and backbone models in experiments. We use BPR [39] as the loss function for all baselines. We use Adam [27] for optimization. Other hyper-parameters for our method and baselines are tuned by grid search. The code and data are available at https://github.com/tsinghua-fib-lab/DICE.

4.2 Performance Comparison (RQ1)

4.2.1 ***Overall Performance.*** We compare our approach with the following state-of-the-art causal recommendation methods:

- **IPS** [26, 40]: IPS eliminates popularity bias by re-weighting each instance according to item popularity. Specifically, weight for an instance is set as the inverse of corresponding item popularity value, hence popular items are imposed lower weights, while the importance for long-tail items are boosted.
- **IPS-C** [10]: This method adds max-capping on IPS value to reduce the variance of IPS.
- **IPS-CN** [18]: This method further adds normalization which also achieved lower variance than plain IPS, at the expense of introducing a small amount of bias.
- **IPS-CNSR** [18]: Smoothing and re-normalization are added to attain more stable output of IPS.
- **CausE** [9]: This method requires a large biased dataset and a small unbiased dataset. Each user or item has two embeddings to perform matrix factorization (MF) on the two datasets respectively, and L1 or L2 regularization is exploited to force the two sets of embeddings similar with each other.

We also include simple MF and GCN without using any causal methods for comparison. We evaluate top-k recommendation performance for implicit feedback [39], which is the most common setting for recommendation. We adopt three frequently used metrics, which are Recall, Hit Ratio and NDCG.

Results on two datasets are listed in Table 2. We have the following observations:

- **Our proposed DICE framework outperforms baselines with significant improvements with respect to all metrics on both datasets.** For example, DICE makes over 15% improvements with respect to NDCG@50 using MF as backbone on Moveilens-10M dataset, and over 20% improvements with respect to Recall@20 using GCN as backbone on Netflix dataset. Results show that the disentanglement design of interest embeddings and conformity embeddings successfully distinguish the

two causes of user interactions. It allows the framework to capture invariant interest from training data, and adapt to intervened conformity in test cases.

- **DICE is highly general framework which can be combined with various recommendation models.** Besides attaining the best performance on both datasets, the proposed DICE framework also outperforms all other baselines with both recommendation models, MF and GCN. The proposed concise causal model is sourced from how the data is generated, thus the proposed framework is independent with backbone recommendation models. Results based on MF and GCN illustrate that DICE is a general framework, which can be smoothly integrated into various embedding based recommendation algorithms.
- **Entangled causal models are not stable on different datasets and metrics.** From the results in Table 2, entangled causal models like IPS and CausE can not make improvements consistently on different datasets and metrics. For example, IPS-CN achieves the second best performance on Movielens-10M dataset, but fails to make improvements on Netflix dataset with MF as recommendation model. In addition, IPS-CNSR attains decent performance with respect to NDCG on Netflix dataset with MF as recommendation model, but it is even worse than None (no causal model) in terms of another metric, HR. Without disentangling interest and conformity, those causal models are not stable on different datasets and metrics. In contrast, the disentangled DICE framework attains consistent improvements by disentangling the underlying causes.

4.2.2 ***Comparison between Embedding and Scalar.*** Using a scalar bias term for each item and user is frequently adopted to capture the influence of popularity [4]. However, it is insufficient to express the diversity of user conformity. For example, user a has stronger conformity towards item s than user b, thus bias term for user a should be higher than user b. However, user a would have weaker conformity towards item t than user b, which requires bias term for user a to be lower than user b. It is common in practice since users tend to have different conformity in their familiar and unfamiliar fields, such as categories of items or genres of movies. The above contradiction demonstrates the limited power of using scalar values to capture user conformity. In our work, we propose to exploit embeddings instead of simple scalars. By raising dimensions of solution space, the diversity of user conformity is guaranteed. For example, the above contradiction can be easily resolved by using 2-D vectors for user conformity and item popularity, rather than 1-D scalars.

We compare the proposed DICE framework using embeddings with existing algorithms using scalar values. We include bias terms for both users and items. Specifically, we compare DICE with BIAS-U (adding scalar bias term for each user), BIAS-I (adding scalar bias term for each item) and BIAS-UI (adding scalar bias term for each user and item) on both MF and GCN. Figure 4 shows the results on the two datasets. DICE outperforms all other models with scalar bias terms with significant margin, proving that simple scalar values are insufficient to capture the diversity of user conformity. Experiments on both MF and GCN show that it is necessary to use embeddings rather than scalar values for conformity modeling.

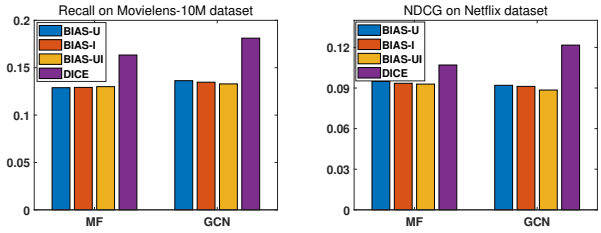


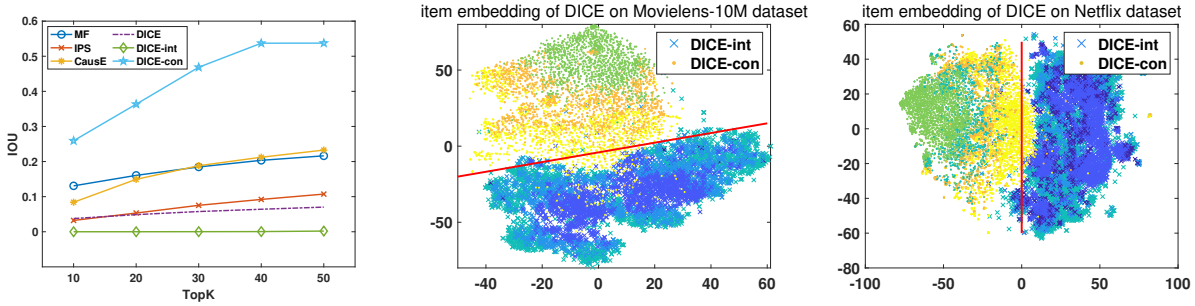
Figure 4: Comparison between using embeddings and using scalars on two datasets.

4.3 Interpretability and Robustness (RQ2)

As introduced previously, disentangled algorithms are generally more interpretable and robust than entangled competitors. In this section, we investigate whether the proposed DICE framework has such advantages.

4.3.1 ***Interpretability based on Disentangled Embedding.*** We investigate the quality of embedding disentanglement in DICE. As there is ground-truth for popularity, which serves as a pseudo proxy for conformity. We first study whether conformity embeddings capture the desired cause. Here we introduce another two versions of the framework, DICE-int and DICE-con. They only use interest or conformity embeddings for recommendation, respectively. Note that in DICE we concatenate the two embeddings. We compare the overlapped recommended items of all methods with ItemPop, which recommends the top popular items. Intersection Over Union (IOU) is used as the metric. Figure 5(a) illustrates the results on Movielens-10M dataset. We observe that using conformity embeddings greatly simulates the ItemPop algorithm, and the overlapped items even surpass 50% when TopK is above 40. Compared with other baselines like IPS and CausE with IOU less than 20%, DICE-con is much more similar to ItemPop, which validates that conformity embeddings indeed capture the desired cause. The IOU value of DICE-con around 0.5 demonstrates that users tend to confirm with popular items, but different users have their own variance in conformity. If all the users are of the same conformity towards popular items, the IOU value would be close to 1. On the other hand, there is almost no overlapped items between DICE-int and ItemPop, proving that conformity information is almost fully squeezed out from interest embeddings. Therefore, interpretations for interest and conformity can be made based on corresponding embeddings.

Besides calculating the similarity with ItemPop, we visualize the learned item embeddings in DICE using t-SNE [34]. Figure 5(b) shows the learned item embeddings on two datasets, where *crosses* represent interest embeddings and *dots* represent conformity embeddings. With special causal learning design and direct supervision on disentanglement, the two sets of embeddings are far from each other, separated by a linear classifier (red line in the figure). Moreover, we divide all the items to three groups based on their popularity, which are popular, normal and unpopular. In Figure 5(b), items of different groups are painted in different colors. We observe that conformity embeddings are layered according to item popularity, where items of similar popularity are near in the embedding space. Notice that if we use scalar values, items of the three groups will form three segments in a straight line, which is insufficient to capture the diversity of conformity. On the other



(a) (b)

Figure 5: (a) Overlapped items with ItemPop. Larger IOU means recommendation result is more similar to ItemPop which recommends top popular items. (b) Visualization of the learned item embeddings of DICE on Movielens-10M and Netflix dataset. Interest embeddings are represented by crosses, and conformity embeddings are represented by dots.

hand, with respect to interest embeddings, items of different popularity are mixed with each other. Visualizations of the learned item embeddings illustrate the high quality of disentanglement in the proposed framework. Based on disentangled embeddings, reasonable interpretations can be made, which is crucial for recommendation. We also compare the item embedding quality of DICE with baseline methods, including MF, CausE and IPS. Our proposed methodology learns highly interpretable representations for user conformity, and successfully captures the diversity of conformity by using embeddings instead of scalars.

4.3.2 ***Robustness under Intervention.*** Algorithms that disentangle underlying causes are generally more robust than entangled approaches under intervention [41]. In our experiments, we conduct intervention by constructing a different test set which is non-IID with training set, in terms of conformity. Users have access to all items with *equal* probability, rather than seeing more popular items in training set. Specifically, the probability of an instance to be included into test set is the inverse of its corresponding item popularity value. Notice that we cap the probability at 0.9 to avoid too many cold-start items in test set, which controls the strength of intervention. With lower capping value, we impose weaker intervention, hence users are more likely to be exposed with popular items. On the contrary, larger capping value attains stronger intervention and different items receive more equal opportunity to be recommended. Therefore, it provides an elegant way to evaluate the robustness of recommender systems under distinct levels of intervention, by simply changing the capping value. In our experiments, we investigate how the proposed framework performs at different strength of intervention, as well as state-of-the-art methods. Figure 6 shows the results of DICE and IPS-CNSR. We compare the performance of the two approaches with capping value as 0.5, 0.7, and 0.9. The three cases represent quite different interventions on user conformity, since users are more likely to conform towards popular items when capping value is 0.5 due to larger exposure probability for popular items, while they tend to interact according to their real interest when capping value is 0.9 since items are exposed in an almost random manner. Results in Figure 6 illustrates that the proposed DICE framework outperforms IPS-CNSR consistently under all degrees of intervention, which proves the robustness of disentangling user interest and conformity.

Table 3: Comparison between the proposed negative sampling strategy PNSM with traditional random strategy on Movielens-10M dataset.

	Top-K=20			Top-K=50		
Name	Recall	HR	NDCG	Recall	HR	NDCG
PNSM	**0.1634**	**0.5197**	**0.1084**	**0.2872**	**0.6975**	**0.1468**
RANDOM	0.1274	0.4394	0.0843	0.2306	0.6255	0.1160

4.4 Study on DICE (RQ3)

Ablation studies on DICE are also conducted to investigate the effectiveness of several components, including negative sampling, conformity modeling, curriculum learning and discrepancy loss.

4.4.1 ***Impact of Negative Sampling.*** As introduced previously, we adopt **P**opularity based **N**egative **S**ampling with **M**argin (PNSM) to gain high confidence of our causal models. Specifically, when the popularity gap between negative item and positive item is sufficiently large, those derived inequalities on interest and conformity would hold true with high probability. Therefore, we sample items that are significantly more or less popular than the positive item. We require that the popularity gap is larger than a margin value.

In this section, we compare PNSM with the commonly used fully random negative sampling strategy. Table 3 shows the results on Movielens-10M dataset. We observe that popularity based negative sampling with margin significantly outperforms random negative sampling. Specifically, Recall and NDCG of PNSM are better than RANDOM with over 20% improvements. PNSM also improves Hit Ratio@20 and Hit Ratio@50 by over 10%. Results of PNSM and RANDOM verify that sampling negative items with large popularity margin is crucial in the proposed framework. It is reasonable since the proposed causal learning methodology depends on those derived inequalities from causal model 1, which will hold true with high probability when the negative items are significantly more or less popular than the positive one.

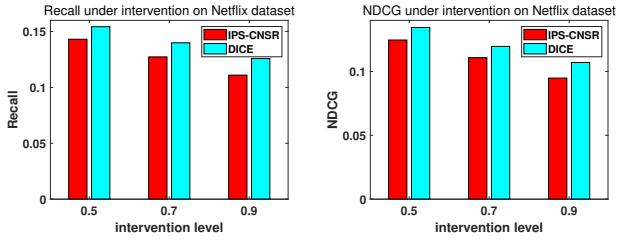


Figure 6: Performance comparison between DICE and IPS-CNSR under different levels of intervention.

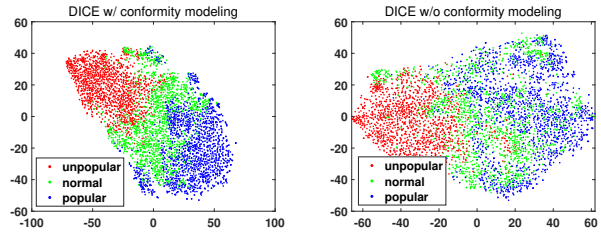
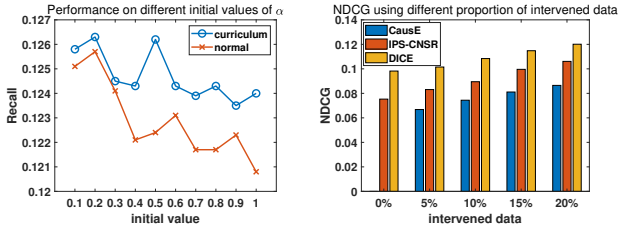


Figure 7: Visualization of the learned item embeddings in DICE on Movielens-10M dataset with and without conformity modeling.



(a) (b)

Figure 8: (a) Comparison between curriculum learning and normal learning. (b) Performance of different proportion of intervened training data.

4.4.2 ***Impact of Conformity Modeling.*** We also investigate the effect of conformity modeling in DICE. Specifically, we remove the conformity modeling task in DICE and compare it with the full version of DICE. We found that recommendation performance does not decrease much, however, removing conformity modeling task indeed influences the learned embeddings. Figure 7 illustrates the learned conformity embeddings in DICE with and without conformity modeling task. We observe that in DICE with conformity modeling task, embeddings are layered according to item popularity, and items of similar popularity are near in the embedding space. However, when we remove the conformity modeling task, the distribution of conformity embeddings becomes messy, and there are more *outliers* in all groups. To be specific, popular items and normal items tend to overlap with each other in the embedding space. Meanwhile, there are also a fraction of normal items lying in the layer of unpopular items.

Conformity modeling task leverages the popularity gap between positive item and negative item to learn pairwise relationships using separate embeddings. From the embedding visualization in Figure 7, we can confirm the effect of conformity modeling task in DICE on learning high-quality interpretable representations.

4.4.3 ***Impact of Curriculum Learning.*** In the proposed framework, we adopt multi-task curriculum learning to aggregate different causes. Specifically, we make several hyper-parameters adaptive to form a easy-to-hard curriculum for causal learning. These hyper-parameters include loss weight α and negative sampling margin values m_{up} and m_{down}. As we train the causal embeddings, we decay these hyper-parameters by a factor of 0.9 to increase the difficulty. In experiments, we initialize these hyper-parameters with different values, and investigate the effect of curriculum learning. Figure 8(a) shows the results of curriculum learning and normal learning on different initial values of loss weight α. We can observe that curriculum learning is consistently better than normal cases. Meanwhile, curriculum learning is not sensitive to initial value due to the easy-to-hard decaying strategy, while normal training without adaptive hyper-parameters is not as stable as curriculum learning and performance drops at large α values.

4.4.4 ***Impact of Discrepancy Loss.*** We provide three options for discrepancy loss, L1-inv, L2-inv and *dCor*. We examine the three candidates on two datasets with two backbones. Overall, *dCor* attains better performance than L1-inv and L2-inv with over 2% improvements. However, *dCor* relies on heavy matrix computations which is much more time-consuming than L1-inv and L2-inv. Specifically, training with *dCor* (about 100s per epoch) as discrepancy loss is much slower than L1-inv and L2-inv (about 44s per epoch), which means L1-inv and L2-inv might be more appropriate for large scale applications.

4.5 Study on Intervened Training Data (RQ4)

In previous experiments, all the algorithms are trained with a large fraction of normal data (60%) and a small fraction of intervened data (10%). Adding extra intervened data is not only a hard requirement of certain baseline method (CausE), but also reduces the difficulty of causal learning. However, intervened data is often too expensive to obtain in real-world recommender systems, *e.g.* random recommendation policy will greatly damage user experience. Therefore, in this section, we investigate how different algorithms perform when we change the proportion of intervened training data, and we also include the most challenging task of not using any intervened data for training. Figure 8(b) demonstrates the performance of DICE, CausE and IPS-CNSR using different proportion of intervened data. Without surprise all the methods got improved performance when we add more intervened data into training set, since it allows models to get access to intervention information which is more similar to test cases. Meanwhile, the proposed DICE framework achieves remarkable improvements against baselines in all cases from 0% to 20%. The proposed DICE framework can still disentangle interest and conformity with even no intervened data, and outperforms other baselines significantly. Notice that there is no result for CausE at 0% since CausE requires intervened training data.

To summarize, we conduct extensive experiments to evaluate the performance of DICE. We compare it with state-of-the-art baseline methods under non-IID circumstances, and DICE outperforms other methods with significant improvements. We emphasize that it is crucial to use embeddings instead of scalars to fully capture the variety of user conformity, which is also proved by experiments against biased MF and biased GCN. Since the main advantages

of disentangled algorithms over entangled algorithms are interpretability and robustness, we further conduct experiments to show DICE indeed provides interpretable results and guarantees robustness under intervention. Moreover, we conduct ablation studies to investigate the role of negative sampling, conformity modeling and curriculum learning. At last, we also study the impact of the proportion of intervened training data and different options for discrepancy loss.

5 RELATED WORK

Causal Recommendation. Existing causal solutions for recommender systems formulate the problem as eliminating popularity bias, from the perspective of items [1, 3, 5, 11, 12, 24, 35, 43]. A bunch of algorithms for unbiased recommendation are proposed in recent literatures, aiming to reduce popularity bias as much as possible [2, 10, 17, 18, 25, 26, 30, 40, 48, 49]. Among them, Inverse Propensity Scoring (IPS) based methods are mostly adopted and achieve state-of-the-art performance. IPS re-weights each instance as the inverse of corresponding item popularity value, thus popular items are imposed lower weights, while the long-tail items are boosted. IPS guarantees zero bias, however, it is with high variance. A series of variants have been proposed to attain more stable results based on IPS. Bottou *et al.* [10] add max-capping on IPS value, Gruson *et al.* [18] further add normalization, and smoothing and re-normalization are also added to reduce the variance of IPS[18]. IPS and its variants attain unbiased or low-biased recommendation, only from the perspective of items, while ignoring the variety of users' conformity. Imposing different weights is insufficient to comprehensively capture user conformity, since it inherently depends on both user and item.

Besides IPS, Bonner *et al.* [9] proposed CausE that performs two MF on a large biased dataset and a small unbiased dataset respectively. L1 or L2 regularization are exploited to force the two factorized embeddings similar with each other. However, conformity is still not taken into consideration in CausE. In recommendation with explicit feedback (*e.g.* rating prediction), Sinha *et al.* [42] decomposed observed ratings to the union of real ratings and recommender influence. With several strong assumptions, they attained a closed-form solution to recover real ratings from observational ratings based on SVD. However, these assumptions turn out to be invalid in the more prevalent implicit feedback setting.

Unlike aforementioned approaches that ignore user conformity and bundle different causes into unified representations, our approach achieves causal recommendation with disentangled embeddings for user interest and conformity. To our knowledge, our proposed methodology is the first attempt to tackle the causal recommendation problem from the perspective of users, attaining superior robustness and interpretability by disentangling user interest and conformity.

Disentangled Representation Learning. Learning representations in which different semantics are disentangled is crucial for robust use of neural models [6, 32, 41, 44]. Existing approaches mainly focus on computer vision [15, 16, 22, 23, 28]. For example, β-VAE [22] learns interpretable representations from raw images in an unsupervised manner. Disentangled representation learning in recommender systems was not explored until recently [14, 33, 47]. Ma *et al.* [33] proposed to use Variational Auto-Encoder to disentangle macro-level concept such as intention on different items, and disentangle micro-level factors like color or size of an item. Wang *et al.* [47] utilized Graph Convolutional Networks to learn disentangled representations for different latent user intentions. These methods decompose user intent into finer granularity, such as the brand or color of an item, while ignoring user conformity, which is essential for recommendation.

6 CONCLUSION AND FUTURE WORK

In this paper, we propose a general framework for disentangling user interest and conformity for recommendation with causal embedding. We develop a concise additive causal model and formulate this model with both causal graph and SCM. Separate embeddings are adopted for interest and conformity according to the proposed SCM. We extract cause-specific data from observational interactions and train different embeddings with different cause-specific data to achieve disentanglement between interest and conformity. The two causes are aggregated and balanced by multi-task curriculum learning. Based on concise and reasonable causal models, DICE consistently outperforms state-of-the-art algorithms with remarkable improvements. Experiments show that DICE is more robust under non-IID circumstances, compared with other baselines. Analysis on disentanglement demonstrates that user interest and conformity are largely independent in the two sets of embeddings. The learned embeddings are of high quality and interpretability, which is promising to explore novel applications using the learned disentangled representations.

DICE decomposes each click interaction into two causes, interest and conformity. A particular meaningful direction for future work is extending DICE to include finer level of causes. For example, the macro-level cause interest could be further divided into micro-level causes such as intentions towards the brand, price or color of items. Overall, we believe disentangling interest and conformity opens new doors for understanding user-item interactions of recommender systems.

ACKNOWLEDGMENTS

This work was supported in part by The National Key Research and Development Program of China under grant 2020AAA0106000, the National Natural Science Foundation of China under U1936217, 61971267, 61972223, 61941117, 61861136003, U19A2079.

REFERENCES

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 42–46.
[2] Aman Agarwal, Kenta Takatsu, Ivan Zaitsev, and Thorsten Joachims. 2019. A General Framework for Counterfactual Learning-to-Rank. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 5–14.
[3] Punam Bedi, Anjali Gautam, Chhavi Sharma, et al. 2014. Using novelty score of unseen items to handle popularity bias in recommender systems. In *2014 International Conference on Contemporary Computing and Informatics (IC3I)*. IEEE, 934–939.
[4] Robert M Bell, Yehuda Koren, and Chris Volinsky. 2008. The bellkor 2008 solution to the netflix prize. *Statistics Research Department at AT&T Research* 1 (2008).
[5] Alejandro Bellogín, Pablo Castells, and Iván Cantador. 2017. Statistical biases in Information Retrieval metrics for recommender systems. *Information Retrieval Journal* 20, 6 (2017), 606–634.

[6] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. 2019. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912* (2019).

[7] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. 41–48.

[8] James Bennett, Stan Lanning, et al. 2007. The netflix prize. In *Proceedings of KDD cup and workshop*, Vol. 2007. Citeseer, 35.

[9] Stephen Bonner and Flavian Vasile. 2018. Causal embeddings for recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 104–112.

[10] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research* 14, 1 (2013), 3207–3260.

[11] Rocío Cañamares and Pablo Castells. 2017. A Probabilistic Reformulation of Memory-Based Collaborative Filtering: Implications on Popularity Biases. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 215–224.

[12] Rocío Cañamares and Pablo Castells. 2018. 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*. 415–424.

[13] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *RecSys*. 224–232.

[14] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2020. Bias and Debias in Recommender System: A Survey and Future Directions. *arXiv preprint arXiv:2010.03240* (2020).

[15] Emilien Dupont. 2018. Learning disentangled joint continuous and discrete representations. In *Advances in Neural Information Processing Systems*. 710–720.

[16] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728* (2018).

[17] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. 2018. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 198–206.

[18] Alois Gruson, Praveen Chandar, Christophe Charbuillet, James McInerney, Samantha Hansen, Damien Tardieu, and Ben Carterette. 2019. Offline Evaluation to Make Decisions About PlaylistRecommendation Algorithms. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 420–428.

[19] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in neural information processing systems*. 1024–1034.

[20] Joseph A. Harper, F. Maxwell anUntitled.texd Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (Dec. 2015), 19 pages. https://doi.org/10.1145/2827872

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

[22] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *Iclr* 2, 5 (2017), 6.

[23] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. 2018. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems*. 517–526.

[24] Dietmar Jannach, Lukas Lerche, Iman Kamehkhosh, and Michael Jugovac. 2015. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction* 25, 5 (2015), 427–491.

[25] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. 2019. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 383–390.

[26] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 781–789.

[27] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[28] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[29] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[30] Dawen Liang, Laurent Charlin, and David M Blei. 2016. Causal inference for recommendation. In *Causation: Foundation to Application, Workshop at UAI*. AUAI.

[31] Yiming Liu, Xuezhi Cao, and Yong Yu. 2016. Are You Influenced by Others When Rating? Improve Rating Prediction by Conformity Modeling. In *Proceedings of the 10th ACM Conference on Recommender Systems*. 269–272.

[32] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *International Conference on Machine Learning*. 4114–4124.

[33] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning disentangled representations for recommendation. In *NeurIPS*. 5712–5723.

[34] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[35] Benjamin Marlin, Richard S Zemel, Sam Roweis, and Malcolm Slaney. 2012. Collaborative filtering and the missing at random assumption. *arXiv preprint arXiv:1206.5267* (2012).

[36] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G. Azzolini, Dmytro Dzhulgakov, Andrey Mallevich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Ansha Yu, Volodymyr Kondratenko, Stephanie Pereira, Xianjie Chen, Wenlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. 2019. Deep Learning Recommendation Model for Personalization and Recommendation Systems. *CoRR* abs/1906.00091 (2019). https://arxiv.org/abs/1906.00091

[37] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic Books.

[38] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference*. The MIT Press.

[39] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[40] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*. 1670–1679.

[41] Bernhard Schölkopf. 2019. Causality for Machine Learning. *arXiv preprint arXiv:1911.10500* (2019).

[42] Ayan Sinha, David F Gleich, and Karthik Ramani. 2016. Deconvolving feedback loops in recommender systems. In *Advances in neural information processing systems*. 3243–3251.

[43] Harald Steck. 2011. Item popularity and recommendation accuracy. In *Proceedings of the fifth ACM conference on Recommender systems*. 125–132.

[44] Raphael Suter, Đorđe Miladinović, Bernhard Schölkopf, and Stefan Bauer. 2018. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. *arXiv preprint arXiv:1811.00007* (2018).

[45] Gábor J Székely and Maria L Rizzo. 2009. Brownian distance covariance. *The annals of applied statistics* (2009), 1236–1265.

[46] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. 2007. Measuring and testing dependence by correlation of distances. *The annals of statistics* 35, 6 (2007), 2769–2794.

[47] Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tonog Xu, and Tat-Seng Chua. 2020. Disentagnled Graph Collaborative Filtering. *SIGIR* (2020).

[48] Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. 2018. The deconfounded recommender: A causal inference approach to recommendation. *arXiv preprint arXiv:1808.06581* (2018).

[49] Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. 2018. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 279–287.

[50] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. 2018. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 974–983.