# Debiased Explainable Pairwise Ranking from Implicit Feedback

Khalil Damak
khalil.damak@louisville.edu
Knowledge Discovery & Web Mining Lab, Dept. of Computer Science & Engineering, University of Louisville
Louisville, Kentucky, USA

Sami Khenissi
sami.khenissi@louisville.edu
Knowledge Discovery & Web Mining Lab, Dept. of Computer Science & Engineering, University of Louisville
Louisville, Kentucky, USA

Olfa Nasraoui
olfa.nasraoui@louisville.edu
Knowledge Discovery & Web Mining Lab, Dept. of Computer Science & Engineering, University of Louisville
Louisville, Kentucky, USA



## ABSTRACT

Recent work in recommender systems has emphasized the importance of fairness, with a particular interest in bias and transparency, in addition to predictive accuracy. In this paper, we focus on the state of the art pairwise ranking model, Bayesian Personalized Ranking (BPR), which has previously been found to outperform pointwise models in predictive accuracy, while also being able to handle implicit feedback. Specifically, we address two limitations of BPR: (1) BPR is a black box model that does not explain its outputs, thus limiting the user's trust in the recommendations, and the analyst's ability to scrutinize a model's outputs; and (2) BPR is vulnerable to exposure bias due to the data being Missing Not At Random (MNAR). This exposure bias usually translates into an unfairness against the least popular items because they risk being under-exposed by the recommender system. In this work, we first propose a novel explainable loss function and a corresponding Matrix Factorization-based model called Explainable Bayesian Personalized Ranking (EBPR) that generates recommendations along with item-based explanations. Then, we theoretically quantify additional exposure bias resulting from the explainability, and use it as a basis to propose an unbiased estimator for the ideal EBPR loss. The result is a ranking model that aptly captures both debiased and explainable user preferences. Finally, we perform an empirical study on three real-world datasets that demonstrate the advantages of our proposed models.

## CCS CONCEPTS

• **Information systems** → **Collaborative filtering**; **Recommender systems**; **Information retrieval**; • **Computing methodologies** → **Machine learning**.

## KEYWORDS

Fairness in AI, Debiased Machine Learning, Pairwise Ranking, Explainability, Exposure Bias

**ACM Reference Format:**
Khalil Damak, Sami Khenissi, and Olfa Nasraoui. 2021. Debiased Explainable Pairwise Ranking from Implicit Feedback. In *Fifteenth ACM Conference on Recommender Systems (RecSys '21), September 27-October 1, 2021, Amsterdam, Netherlands.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3460231.3474274


*RecSys '21, September 27-October 1, 2021, Amsterdam, Netherlands*

ACM ISBN 978-1-4503-8458-2/21/09…$15.00
https://doi.org/10.1145/3460231.3474274

## 1 INTRODUCTION

Bayesian Personalized Ranking (BPR) is a state of the art pairwise ranking approach [36] that has recently received significant praise in the recommender systems community because of its capacity to rank implicit feedback data with high accuracy compared to pointwise models [18]. Aiming to rank relevant items higher than irrelevant items, pairwise ranking recommender systems often assume that all non-interacted items as irrelevant. Hence, these systems rely on the assumption that implicit feedback data is Missing Completely At Random (MCAR), meaning that the items are equally likely to be observed by the users [40], consequently any missing interaction is missing because the user chose not to interact with it. However, given the abundance of items on most e-commerce, entertainment, and other online platforms, it is safe to assume the impossibility of any user being exposed to *all* the items. Thus, missing interactions should be considered Missing Not At Random (MNAR). This means that the user may have been exposed to part of the items, but chose not to interact with them, which can be a sign of irrelevance; and was not exposed to the rest of the items. This MNAR property is translated into an exposure bias. This type of bias is usually characterized by a bias against less popular items that have a lower propensity of being observed [6].

Moreover, most accurate recommender systems tend to be black boxes that do not justify why or how an item was recommended to a user. This might engender unfairness issues if, for example, particularly inappropriate or offensive content gets recommended to a user. This kind of unfairness can be better diagnosed and mitigated with an explanation. In fact, it could be important for the user to know why or how the inappropriate item was recommended. For example, an Italian user might think that the movie recommendation "The Godfather" is offensive because of the way it depicts, in an unfair stereotypical way, a certain Italian community in the US. However, the explanation "Because you liked the movie "Scarface"" can be important in this case, because it clarifies that the movie recommendation was not tied to a community, but rather resulted from the user also liking another similar "mafia" sub-genre movie. Furthermore, explanations have been shown to help users make more accurate decisions, which translates into an increased user satisfaction [2, 4]. Bayesian Personalized Ranking [36] treats comparisons between any positive and negative items the same, regardless of which ones can or cannot be explained. Thus, while BPR aptly captures and models ranking based preference, it does not yet capture an *explainable* preference. It is this *explainable preference*, in

addition to an unbiased preference ranking, that we seek to achieve in this work. We thus propose models that address explainability *and* exposure bias in pairwise ranking from implicit feedback and achieve the following contributions:

- Proposing an explainable loss function based on the state of the art Bayesian Personalized Ranking (BPR) loss [36] along with a corresponding Matrix Factorization (MF)-based model called Explainable Bayesian Personalized Ranking (EBPR). To the extent of our knowledge, no work has introduced neighborhood-based explainability to pairwise ranking.
- Conducting a theoretical study of the additional exposure bias coming from the item-based explanations.
- Proposing an unbiased estimator for the ideal EBPR loss, called UEBPR, based on the Inverse Propensity Scoring (IPS) estimator [37]. To our knowledge, no prior work has tried to address the additional exposure bias that could result from neighborhood-based explainability.
- Performing an empirical study on three real-world datasets to compare the effectiveness of the proposed models, in terms of ranking, explainability, and both exposure and popularity debiasing.
- Investigating the properties of the proposed neighborhood based explainable models, revealing and explaining a desirable inherent popularity debiasing that is built into these models. This opens the path to a new family of future debiasing strategies, where the debiasing is rooted in an explainable neighborhood-based rationale.

In addition, we make our implementations of all the models presented in this paper available for reproducibility[1].

## 2 BACKGROUND

In this section, we start by reviewing previous work on explainability and counteracting exposure bias in recommendation. While it is impossible to do justice to every past contribution with an exhaustive review, we try to focus on the most representative or related work. Then we review Bayesian Personalized Ranking (BPR).

### 2.1 Explainability in Recommendation

The types of explanations in recommendation have varied with the type of data used [4, 43] Some explanations are content-based, meaning that they usually come from features. These were used in works that employed sentiment analysis on user reviews along with learned latent features to generate explanations in the form of user or item features [53], textual sentences [53] or word clusters [52]. Other research efforts used attention mechanisms to explain recommendations [9, 10, 27, 41]. The generated explanations are important regions in the textual [41] or image [9, 10, 27] inputs. Other methods relied on post-hoc approaches that try to extract explanations for the recommendations after they occur. For instance, [35] and [11] use influence functions to determine the effect of each input interaction on the recommendation; while [13] proposed an approach that forward-propagates song segments through the trained recurrent neural network model to determine the most explanatory segment in a song recommendation. In contrast to the above methods, some explainable recommender systems rely solely on feedback data such as ratings or interactions. Hence, they have the advantage of (1) accommodating collaborative filtering (CF) models and (2) not requiring any additional content or metadata to generate explanations for CF. These explanations tend to depend only on the rating data and they are mainly neighborhood-based, and can be either user-based or item-based [2, 21]. Explanations can be obtained by using classical, inherently interpretable, user-based or item-based collaborative filtering techniques [21, 39] or by using model-based approaches. The latter are most related to our work. Among model-based approaches, Explainable Matrix Factorization (EMF) [2] pre-computes a user or item-based neighborhood style explainability matrix from the ratings, and then uses this matrix in a regularization term that is added to obtain an explainable recommendation reconstruction loss to guide the learning and yield explainable recommendations. This approach provides a simple and flexible way to add explainability to latent factor loss-based models to obtain a single integrated explainable model. It also has the advantage of not being a post-hoc approach, and hence not incurring the cost of learning two separate models, nor risking lack of fidelity from deviations between the explaining model and the predictive model. For all these reasons, EMF was later adopted in several works, such as [12] which extended it and tried to improve the novelty of the recommendations; and in [45] which modified the calculation of the explainability matrix by integrating the neighbors' weights to improve performance. Other works used influence functions to generate neighborhood-based explanations. For instance, [30] proposed a probabilistic factorization model, which employs an influence mechanism to evaluate the importance of the users' historical data and present the most related users and items as explanations for the predicted rating.

### 2.2 Exposure Bias in Recommendation

Bias in recommendation can be categorized into seven types [6] that occur within the various stages of the recommendation feedback loop [23, 24, 32, 42] between the user, the data, and the model. Among these categories, in the user-to-data phase, we find *exposure bias*, which is the focus of our work in this paper. Exposure bias happens when users are only exposed to a portion of the items, and hence, unobserved interactions do not always represent negative preferences [6]. The techniques that have been introduced to mitigate exposure bias, vary in whether they treat bias during the training or evaluation [6]. The common approach that is used in the evaluation phase incorporates an Inverse Propensity Scoring (IPS) modification of the ranking evaluation metrics, where more popular items are down-weighted and less popular items are up-weighted [49]. Exposure debiasing in training is usually achieved by considering the unobserved interactions as negatives with a certain confidence [6]. These methods differ in the way they define or approximate the confidence weight. One group of methods [14, 22] considers a uniform weight for all negative items that is lower than one; while a second group [33, 34] utilizes the user activity, for instance the number of interacted items, to weight the negative interactions; and a third group uses item popularity [20, 50] and user-item similarity [28] to achieve a similar goal. Recent work, [38] and [37], proposed IPS-based unbiased estimators for the ideal

[1] https://github.com/KhalilDMK/EBPR

pointwise and pairwise losses, respectively. In their experiments, they estimated the propensity of an interaction using the relative item popularity. On the other hand, [25] proposed a regularization term that penalizes non-uniform exposure. Departing from the previously mentioned methods, other work proposed negative sampling processes in order to mitigate exposure bias. This negative sampling is usually done by exploiting side information such as social network information [8] or item-based knowledge graphs [46]. Another approach is to integrate the capacity to learn the exposure probability within the model [7, 8, 29], which in turn requires assumptions on the probability distribution of exposure. Finally, [3, 31, 47, 51] consider users' sequential behavior to address exposure bias with multi-task learning.

### 2.3 Bayesian Personalized Ranking for Pairwise Ranking

The Bayesian Personalized Ranking (BPR) loss was introduced in [36] as the first loss that is "optimized for ranking" in the implicit feedback pairwise ranking setting. In other words, it learns the users' preference of a positive item over a negative item. In this case, positive and negative items are those that the user has, respectively, interacted with and not interacted with. This is opposed to pointwise prediction, which can be seen as a predictive classification problem of the relevance of an item to a user. Pairwise ranking has received increasing attention and praise over the years from the recommender system community due to its high performance in top-N recommendation compared to pointwise ranking [18]. The BPR objective function is defined as follows:

$$L_{BPR} = \frac{1}{|D|} \sum_{(u,i_+,i_-)\in D} -log\sigma(f_\Omega(u,i_+,i_-)), \tag{1}$$

where $D = \{(u,i_+,i_-)|u\epsilon U, i_+\epsilon I_u^+, i_-\epsilon I_u^-\}$ is the training data. $I_u^+$ is the set of positive (interacted) items by user $u$ and $I_u^-$ is the set of negative (non-interacted) items by user $u$ such that $I_u^- = I \setminus I_u^+$. $f_\Omega$ is a hypothesis with parameters $\Omega$ that quantifies how much user $u$ prefers (following the order relation $>_u$ defined in [36]) item $i_+$ over item $i_-$, and $\sigma$ is the Sigmoid function. When the BPR loss is applied to Matrix Factorization (MF) with the parameters $\Omega$ consisting of the respective user and item latent matrices $P\epsilon\mathbb{R}^{n\times K}$ and $Q\epsilon\mathbb{R}^{m\times K}$, the preference model is given by

$$f_\Omega(u,i_+,i_-) = P_u \cdot Q_{i_+}^T - P_u \cdot Q_{i_-}^T. \tag{2}$$

Applying the Sigmoid function to the output of the preference model yields the preference probability, which is the probability of user $u$ preferring item $i_+$ over item $i_-$: $P_\Omega(i_+ >_u i_-) = P(i_+ >_u i_-|\Omega) = \sigma(f_\Omega(u,i_+,i_-))$. Note that in equation 1, as in the remainder of this paper, we omitted any regularization terms from the equations for simplicity, although we use L2 regularization in our implementation.

## 3 EXPLAINABLE BAYESIAN PERSONALIZED RANKING

To the extent of our knowledge, no work has introduced neighborhood based explainability to pairwise ranking. More importantly, although neighborhood-based explainability can be expected to be vulnerable to exposure bias, there is a need to mitigate any additional exposure bias coming from the explainability. The BPR model learns to rank positive (interacted) items by a user higher than any negative (non-interacted) item. This objective treats comparisons between any positive and negative items the same, regardless of which ones can or cannot be explained based on any given style of explanation, for instance based on neighborhoods. Thus, while BPR aptly captures and models a ranking based preference, it does not yet capture an *explainable* preference. In fact, as demonstrated in [2], it is important to consider the interpretability of the items to the users, often referred to as *explainability*, when learning a recommendation objective, and this can be computed based on readily available rating data, for instance from similar items. Hence, given a definition for a measure of explainability $E_{ui}$, of an item $i$ to a user $u$, our aim is to condition the BPR objective function to capture what we call *explainable preference*. This means giving more importance to the explainable items that it is learning to rank higher, and less importance to the explainable items that it is learning to rank lower. In other words, if the objective function is learning to rank, for a user $u$, an item $i_+$ higher than an item $i_-$, then we would additionally want to give an even higher importance to this preference if it is also accompanied by a higher explainability $E_{ui_+}$ of item $i_+$ to user $u$ and a lower explainability $E_{ui_-}$ of item $i_-$ to user $u$. We formulate this *explainable preference* desiderata into a modified objective to obtain Explainable Bayesian Personalized Ranking (EBPR) as follows:

Definition 1 (Explainable Bayesian Personalized Ranking (EBPR) Objective Function). *Given an explainability matrix $E = (E_{ui})_{u=1..|U|,i=1..|I|} \in [0,1]^{|U|\times|I|}$, where $E_{ui}$ is a measure of explainability of item $i$ to user $u$, the EBPR objective function is defined as*

$$L_{EBPR} = \frac{1}{|D|} \sum_{(u,i_+,i_-)\in D} -E_{ui_+}(1-E_{ui_-})log\sigma(f_\Omega(u,i_+,i_-)). \tag{3}$$

The intuition is to weight the contribution of an instance $(u,i_+,i_-)$ into the loss by $E_{ui_+}(1-E_{ui_-})$, in proportion to the degree that the positive item is considered to be more explainable and the negative item is considered less explainable. Hence, the higher the explainability $E_{ui_+}$ and the lower the explainability $E_{ui_-}$, the more the instance $(u,i_+,i_-)$ will contribute to the learning. This also means that, when generating a recommendation list to a user $u$, the items ranked at the top of the list would be expected to have higher explainability than the items ranked lower in the list. Thus the multiplicative explainability term can be seen as one way to formulate an *explainable preference* function, that is furthermore flexible, since any explainability score can be incorporated.

The latter objective function may seem counter-intuitive due to the fact that the loss increases when the explainability weighting term $E_{ui_+}(1-E_{ui_-})$ increases. However, the model learns with the update equations regardless of the value of the loss. Hence, instead of trying to reduce the loss further when the explainability weighting term $E_{ui_+}(1-E_{ui_-})$ increases, we aim to increase the *contribution* of the instance $(u,i_+,i_-)$ to the learning objective. To gain a better insight, we derive the gradient used in the update equations of EBPR, with respect to the model parameters $\Omega$:

$$\frac{\partial L_{EBPR}}{\partial \Omega} = \frac{-1}{|D|} \sum_{(u,i_+,i_-) \in D} E_{ui_+}(1-E_{ui_-}) \frac{e^{-f_\Omega(u,i_+,i_-)}}{1+e^{-f_\Omega(u,i_+,i_-)}} \frac{\partial f_\Omega(u,i_+,i_-)}{\partial \Omega}. \tag{4}$$

From (2), we have

$$\frac{\partial f_\Omega(u,i_+,i_-)}{\partial \Omega} = \begin{cases} Q_{i_+k} - Q_{i_-k} & if\ \Omega = P_{uk}, \\ P_{uk} & if\ \Omega = Q_{i_+k}, \\ -P_{uk} & if\ \Omega = Q_{i_-k}, \\ 0 & otherwise. \end{cases}$$

The amplitude of the gradient with respect to parameter $\Omega$ is thus an increasing function of the explainability weighting factor $E_{ui_+}(1 - E_{ui_-})$ in a way that confirms the desired explainable preference aim. For instance, in the extreme case where either the positive item is not explainable at all or the negative item is completely explainable, the update equation is zeroed out. Hence, no contribution will come from the corresponding instance to the learning. This is reasonable and desirable since the aforementioned case depicts a *non* explainable preference, where either the positive item is not explainable or the negative item is explainable. Either case undermines the explainability of the preference.

## 3.1 Explainability Matrix

Various measures of explainability can be defined given the characterized order relation of an item $i$ being "more explainable" than an item $j$ to a user $u$. The notion of explainability may depend on user or item metadata if using a content-based or hybrid approach. But in a purely collaborative filtering approach, such as in our case, it should be neighborhood-based as proposed in [2], which further categorized the explanations as user-based or item-based. User-based explanations are based on user similarities and generate explanations in the form of "this item was recommended because certain similar users liked it". Item-based explanations use item-similarities and generate explanations in the form "the item was recommended because you liked similar items". We extend the idea of neighborhood-based explainability from [2] because it has shown success as an intuitive method for modifying loss-based recommendation models [12, 45]. Both item-based and user-based measures of explainability can be defined by relying solely on the interaction matrix (or rating matrix, depending on the type of feedback). However, in this work, we focus only on item-based explanations which are expected to be more intuitive and informative to the user than user-based explanations. This is because the user knows the items that they interacted with but does not necessarily know their neighbors who have similar interactions with items. That said, a user-based explainability matrix can be similarly defined by applying the same strategy, described below, on the transpose of the interaction matrix. We define the measure of *explainability* $E_{ui}$ as the probability of user $u$ interacting with item $i$'s neighbors, as shown below.

Definition 2 (Item-based explainability for Implicit Feedback).

$$E_{ui} = P(Y_{uj} = 1 | j \in N_i^\eta), \tag{5}$$

*where $N_i^\eta$ is the neighborhood of item $i$ which is a set of item $i$'s $\eta$ most similar items given a similarity measure. $Y_{ui}$ is a Bernoulli random variable that takes value 1 if user $u$ interacted with item $i$ and 0 otherwise:*

$$Y_{ui} = \begin{cases} 1 & if\ i \in I_u^+, \\ 0 & otherwise. \end{cases}$$

The explainability $E_{ui}$ can also be reformulated as $E_{ui} = \frac{|N_i^\eta \cap I_u^+|}{\eta}$. This means that for a specific item, the more neighboring items a given user has interacted with, the higher the explainability of that item will be to this user. In our experiments, we use the Cosine similarity between items to generate the neighborhoods.

*3.1.1 Justifications for the Choice of Explainability.* In contrast to post-hoc explainability approaches, which generate explanations after the predictions have been made, our approach pre-computes explanation scores, then uses them to learn an explainable model. This leads to two advantages: (1) better transparency since there is no post-hoc model and (2) avoiding the heavy cost of post-hoc model training and explanation generation at prediction time.

Aiming toward *transparency* is also why we chose to use *neighborhood-based explainability*. More specifically, our aim is to explain recommendations using *only* the input data used by the recommendation algorithm, and not any additional data that is not used to generate predictions. Consequently and because BPR uses no metadata, the explanations must be sourced from only the interaction data.

## 3.2 Training Complexity of EBPR

The complexity of learning the BPR model is $O(|D|K)$, where $|D|$ is the size of the training data, and $K$ is the number of latent factors. This is because the complexity of forward and backward propagating an instance stems from computing two dot products, which is $O(K)$. Considering that generating the explainability matrix can be done offline in the data pre-processing phase, no additional time complexity needs to be added to the training process of EBPR compared to BPR. That said, the explainability matrix is computed only once, and the most significant part of the computation is computing the similarity values initially, which can be done very efficiently, owing to the sparsity of the interactions and the power law in the data distribution, allowing the use of sparse structures and locality sensitive hashing [15].

# 4 EXPOSURE BIAS IN EBPR

As proved in [37], the estimator optimized in BPR is biased against the ideal pairwise loss. This is because the choice of the positive and negative items depends on the interaction random variable $Y_{ui}$ instead of the relevance. In fact, there is a discrepancy between interaction and relevance. Assuming that a relevant item is interacted implies that all non-interacted items are irrelevant, even if they were not exposed. This biases the BPR loss. Explainability too relies on the interaction random variable, hence amplifying this bias. To model exposure and relevance, we consider two Bernoulli random variables: $O_{u,i} \sim Ber(\theta_{ui})$, where $\theta_{ui} = P(O_{ui} = 1)$, represents the exposure propensity of item $i$ relative to user $u$; and $R_{u,i} \sim Ber(\gamma_{ui})$, where $\gamma_{ui} = P(R_{ui} = 1)$, represents the probability of item $i$ being relevant to user $u$. $O_{u,i}$ and $R_{u,i}$ represent, respectively, whether item $i$ is exposed or relevant to user $u$. We only know if user $u$

interacted with item $i$ when the item is both observed and relevant. In other words, $Y_{ui} = O_{ui}R_{ui}$ [37]. However, there could be relevant unobserved items that the user did not get a chance to observe in order to interact with. To handle this issue, [37] proposed an Inverse Propensity Scoring (IPS) based estimator, as was done earlier for explicit feedback ratings in [40], that is unbiased with respect to the ideal pairwise estimator. The latter is defined as follows.

Definition 3 (Unbiased estimator for the ideal BPR loss).

$$L_{BPR}^{unbiased} = \frac{1}{|U||I|^2} \sum_{(u,i_+,i_-)\in U\times I\times I} -\frac{Y_{ui_+}}{\theta_{ui_+}}(1-\frac{Y_{ui_-}}{\theta_{ui_-}}) log\sigma(f_\Omega(u,i_+,i_-)). \tag{6}$$

Given that the explainability scores $E_{ui}$ also rely on the interaction random variable $Y_{ui}$, it is reasonable to suspect that the explainability weighting of the loss could introduce some additional exposure bias. In fact, it would be ideal to use the relevance to define a more ideal explainability matrix as follows.

Definition 4 (Ideal explainability matrix).

$$E_{ui}^{ideal} = P(R_{uj} = 1 | j \in N_i^\eta). \tag{7}$$

This being done, we use the ideal explainability matrix to define the ideal EBPR loss as follows.

Definition 5 (Ideal EBPR loss).

$$L_{EBPR}^{ideal} = \frac{1}{|U||I|^2} \sum_{(u,i_+,i_-)\in U\times I\times I} -\gamma_{ui_+}(1-\gamma_{ui_-})E_{ui_+}^{ideal}(1-E_{ui_-}^{ideal}) \times log\sigma(f_\Omega(u,i_+,i_-)). \tag{8}$$

To quantify the additional bias, we compare the ideal EBPR loss to an IPS-based estimator similar to the one defined in Definition 3, but with explainability weighting. We call the latter estimator pUEBPR loss, where the "pU" stands for *partially unbiased*, and formulate it as follows.

Definition 6 (Partially Unbiased Explainable BPR (pUEBPR) loss).

$$L_{pUEBPR} = \frac{1}{|U||I|^2} \sum_{(u,i_+,i_-)\in U\times I\times I} -\frac{Y_{ui_+}}{\theta_{ui_+}}(1-\frac{Y_{ui_-}}{\theta_{ui_-}})E_{ui_+}(1-E_{ui_-}) \times log\sigma(f_\Omega(u,i_+,i_-)). \tag{9}$$

The pUEBPR loss eliminates the initial exposure bias of BPR without taking into account the impact of adding explainability. Thus it is not a complete debiasing. However, as we will show below, this *partial* debiasing loss will allow us to quantify the additional bias coming from adding the explainability weighting to BPR. Next, we prove that the *explainability* weighting in the EBPR loss introduces *additional* exposure bias. Then we proceed to eliminate this additional bias in the next section.

Proposition 1 (Additional exposure bias from explainability weighting in EBPR). *(proof is omitted) The explainability weighting in the EBPR loss introduces additional non-zero exposure bias, given by*

$$Add\_Bias\_EBPR = \mathbb{E}[L_{pUEBPR}] - L_{EBPR}^{ideal} \neq 0. \tag{10}$$

## 5 UNBIASED EBPR ESTIMATOR

We follow the same IPS-based methodology on the explainability weighting to propose an unbiased estimator for the ideal EBPR loss:

Definition 7 (Unbiased EBPR (UEBPR) estimator).

$$L_{UEBPR} = \frac{1}{|U||I|^2} \sum_{(u,i_+,i_-)\in U\times I\times I} -\frac{Y_{ui_+}}{\theta_{ui_+}}(1-\frac{Y_{ui_-}}{\theta_{ui_-}})\frac{E_{ui_+}}{\theta_{uN_{i_+}^\eta}}(1-\frac{E_{ui_-}}{\theta_{uN_{i_-}^\eta}}) \times log\sigma(f_\Omega(u,i_+,i_-)), \tag{11}$$

*where $\theta_{uN_i^\eta} = P(O_{uj} = 1 | j \in N_i^\eta)$ is the probability of user $u$ being exposed to the neighbors of item $i$. $\theta_{uN_i^\eta}$ can also be considered as the item's neighborhood propensity relative to user $u$.*

Now, we prove that this new UEPBR estimator is unbiased for the ideal EBPR loss in the following proposition.

Proposition 2. *The UEBPR estimator is unbiased for the ideal EBPR loss, meaning that*

$$\mathbb{E}[L_{UEBPR}] = L_{EBPR}^{ideal}. \tag{12}$$

Proof.

$$\begin{aligned}
\mathbb{E}[L_{UEBPR}] &= \frac{1}{|U||I|^2} \sum_{(u,i_+,i_-)\in U\times I\times I} -\frac{\mathbb{E}[Y_{ui_+}]}{\theta_{ui_+}}(1-\frac{\mathbb{E}[Y_{ui_-}]}{\theta_{ui_-}}) \\
&\times \frac{E_{ui_+}}{\theta_{uN_{i_+}^\eta}}(1-\frac{E_{ui_-}}{\theta_{uN_{i_-}^\eta}}) log\sigma(f_\Omega(u,i_+,i_-)) \\
&= \frac{1}{|U||I|^2} \sum_{(u,i_+,i_-)\in U\times I\times I} -\gamma_{ui_+}(1-\gamma_{ui_-})\frac{E_{ui_+}}{\theta_{uN_{i_+}^\eta}}(1-\frac{E_{ui_-}}{\theta_{uN_{i_-}^\eta}}) \\
&\times log\sigma(f_\Omega(u,i_+,i_-)) \\
&= \frac{1}{|U||I|^2} \sum_{(u,i_+,i_-)\in U\times I\times I} -\gamma_{ui_+}(1-\gamma_{ui_-})\frac{P(O_{uj}=1,R_{uj}=1|j\in N_{i_+}^\eta)}{\theta_{uN_{i_+}^\eta}} \\
&\times (1-\frac{P(O_{uj}=1,R_{uj}=1|j\in N_{i_-}^\eta)}{\theta_{uN_{i_-}^\eta}}) log\sigma(f_\Omega(u,i_+,i_-)) \\
&= \frac{1}{|U||I|^2} \sum_{(u,i_+,i_-)\in U\times I\times I} -\gamma_{ui_+}(1-\gamma_{ui_-})\frac{\theta_{uN_{i_+}^\eta}E_{ui_+}^{ideal}}{\theta_{uN_{i_+}^\eta}}(1-\frac{\theta_{uN_{i_-}^\eta}E_{ui_-}^{ideal}}{\theta_{uN_{i_-}^\eta}}) \\
&\times log\sigma(f_\Omega(u,i_+,i_-)) = L_{EBPR}^{ideal}
\end{aligned}$$

□

To get the last line, we assume conditional independence between exposure and relevance given the neighborhood, a much less restrictive (and thus more realistic) assumption than global independence.

## 6 EXPERIMENTAL EVALUATION

We evaluate the impact of introducing explainability and counteracting exposure bias by tuning and then comparing the models described in Sections 2.3 - 5 in terms of ranking performance, explainability, and debiasing capabilities.

### 6.1 Data Used

We use three datasets: The Movielens 100K [16] (ml-100k), The Yahoo! R3 [48] (yahoo-r3) and the Last.FM 2K [5, 26] (lastfm-2k)

**Table 1: Datasets used for evaluation.**

| Dataset | Task | Users | Items | Interactions | Sparsity |
|---|---|---|---|---|---|
| ml-100k | Movie rec. | 943 | 1,682 | 100,000 | 93.6% |
| yahoo-r3 | Song rec. | 15,400 | 1,000 | 311,704 | 97.9% |
| lastfm-2k | Artist rec. | 1,874 | 17,612 | 92,780 | 99.7% |

datasets. These datasets consist of, respectively, 100K movie interactions, over 311K song interactions, and over 92K artist interactions. The interactions consist of either ratings or play counts, which were converted into binary interactions, regardless of their values. In fact, any rating or play count over the threshold of zero is considered a positive interaction. Then we filtered out users with less than 10 interactions in the lastfm-2k dataset to ensure enough training and evaluation samples for every user and reduce the data sparsity. The other two datasets similarly have at least 10 interactions per user. The datasets' properties are summarized in Table 1.

## 6.2 Experimental Setting

We follow the standard Leave-One-Out (LOO) procedure [19, 36] that consists of considering the latest interaction of each user as a test item and comparing it to 100 randomly sampled negative items. In the training, we sample, at every epoch, one negative item for every positive user-item interaction. We implement "BPR", "UBPR", "EBPR", "pUEBPR" and "UEBPR" and tune their hyperparameters on every dataset by comparing the averages over two replicates of 15 random hyperparameter configurations. We further split the training data into training and validation sets for the hyperparameter tuning. We consider the last interaction of every user from the training data along with 100 sampled negatives (disjoint from those in the test set) as a validation set. For each random hyperparameter configuration, we choose a value for the number of latent features, batch size and L2 regularization within the respective sets {5, 10, 20, 50, 100}, {50, 100, 500} and {0, 0.00001, 0.001}. We initially fixed the neighborhood size to 20 to ensure a fair comparison in terms of explainability metrics. However we will investigate the impact of neighborhood size later in Section 7.6. This being done, we then re-train every model on the merged train and validation sets with its best hyperparameter configuration for five replicates and report the average results on the test set. We also perform Tukey tests for pairwise comparison [17] to check the significance of the results. Note that, in our implementation of the unbiased models, namely UBPR, pUEBPR, and UEBPR, we only use positive and negative interaction pairs in the training to ensure that all models are trained on the exact same datasets, and truly assess the impact of every component in the loss. Also note that the goal of the experiments is to assess the impact of the added explainability and debiasing components on BPR. For this reason, we leave for future work, the task of comparing our algorithms to additional baselines.

## 6.3 Evaluation Metrics

We use Normalized Discounted Cumulative Gain ($NDCG@\mathcal{K}$) and Hit Ratio ($HR@\mathcal{K}$) for the ranking evaluation, and Mean Explainability Precision ($MEP@\mathcal{K}$) [1] and Weighted MEP ($WMEP@\mathcal{K}$) for the explainability evaluation. $MEP@\mathcal{K}$ measures the proportion of explainable items within the list of Top K recommendations, as follows

$$MEP@\mathcal{K}(TopK) = \frac{1}{|U|}\sum_{u=1}^{|U|}\frac{|\{i \in TopK(u)\} \cap \{E_{ui} > 0\}|}{\mathcal{K}}, \quad (13)$$

where $TopK$ is the top $\mathcal{K}$ recommendation matrix in which every row represents the Top $\mathcal{K}$ recommendations of a user. We further extend $MEP@\mathcal{K}$ to be able to weight the items' contributions to the numerator by their explainability values, since $MEP@\mathcal{K}$ rewards items that are considered to be explainable (i.e., with explainability score above a given threshold) in the same way, regardless of how different their explainability values are. Hence, we propose a weighted version of MEP that weights items' contributions by their explainability values. The *Weighted MEP (WMEP)* is given by

$$WMEP@\mathcal{K}(TopK) = \frac{1}{|U|}\sum_{u=1}^{|U|}E_{ui}\frac{|\{i \in TopK(u)\} \cap \{E_{ui} > 0\}|}{\mathcal{K}}. \quad (14)$$

Note that when training a model, we hide all test interactions when generating the explainability matrix to avoid any data leakage from the test set. Then, when evaluating the model on the test set, we generate an explainability matrix that considers all interactions to ensure an evaluation of the actual explainability of the test items to users. Furthermore, we evaluate the popularity debiasing of the models in three aspects, namely Novelty, Popularity and Diversity. To evaluate the novelty of a model, we use Expected Free Discovery (EFD) [44], which is a measure of the ability of a system to recommend relevant long-tail items [44]. EFD is defined as

$$EFD@\mathcal{K}(TopK) = -\frac{1}{|U|}\sum_{u=1}^{|U|}\frac{1}{\mathcal{K}}\sum_{i \in TopK(u)} log_2\hat{\theta}_{ui}. \quad (15)$$

Note that we use an estimator of the propensity $\hat{\theta}_{ui}$ to compute the popularity as we will see later in Section 6.4. Next, to evaluate the popularity of the recommendations, we compute the average popularity at $\mathcal{K}$, using

$$Avg\_Pop@\mathcal{K}(TopK) = \frac{1}{|U|}\sum_{u=1}^{|U|}\frac{1}{\mathcal{K}}\sum_{i \in TopK(u)}\hat{\theta}_{ui}. \quad (16)$$

Finally, to evaluate recommendation diversity, we compute the Average Pairwise Similarity between the items in a top $\mathcal{K}$ recommendation list, which is given by [44]

$$Div@\mathcal{K}(TopK) = \frac{1}{|U|}\sum_{u=1}^{|U|}\frac{1}{\mathcal{K}(\mathcal{K}-1)}\sum_{\substack{i,j \in TopK(u)\\ i<j}} sim(i,j), \quad (17)$$

where $sim(i, j)$ is a measure of similarity between item $i$ and item $j$'s interaction vectors. In our experiments, we use the Cosine similarity. All ranking and explainability metrics are computed at a cutoff $\mathcal{K}$ = 10 for Top 10 recommendation.

### 6.4 Propensity Estimation

Following [37], we estimate the propensity of an item to a user by the relative item popularity of the item such that:

$$\hat{\theta}_{ui} = \sqrt{\frac{\sum_{j=1}^{|U|} Y_{ji}}{\max_{l \in I} \sum_{j=1}^{|U|} Y_{jl}}}. \tag{18}$$

The total propensity of item $i$ within its neighborhood can be defined as the average propensity of the items in the neighborhood[2]; i.e., $\hat{\theta}_{uN_i^\eta} = \frac{1}{\eta} \sum_{l \in N_i^\eta} \hat{\theta}_{ul}$.

## 7 RESULTS AND DISCUSSION

### 7.1 Overall Ranking and Explainability Results

Table 2 lists the results of all the models in terms of ranking performance and explainability. Overall, for both the ml-100k and yahoo-r3 datasets, the explainable models EBPR and pUEBPR outperformed all the other models in terms of ranking performance and explainability for almost all the metrics. Moreover, whenever EBPR was not the best performer, it was still second to best. On the lastfm-2k dataset, the non-explainable models (BPR and UBPR) reached better ranking performance than the explainable models (EBPR, pUEBPR and UEBPR). However, the explainable models were still the winners in terms of explainability (MEP and WMEP). Our interpretation of the exception in the lastfm-2k dataset, is that it is likely due to the extremely high sparsity of this dataset (99.7%), which in turn impacts the similarity based computations to determine the neighborhoods used in computing the explainability values. This in turn degrades the learning of the explainable models due to the vanishing gradient problem. We will investigate this issue further in Section 7.5, where we will investigate the effect of the data sparsity on the learning of the explainable models.

### 7.2 Advantages of using Explainability Weighting in the Learning Objective

In order to demonstrate the advantages of the proposed explainability weighting in (3), we compare EBPR to BPR and pUEBPR to UBPR because these models only differ by the explainability weighting of the loss. In both the ml-100k and yahoo-r3 datasets, going from BPR to EBPR almost always improves both the ranking and explainability performances. However, going from UBPR to pUEBPR improves the explainability but does not always improve the ranking performance. In fact, the ranking performance improves on the yahoo-r3 dataset but not on the ml-100k dataset. Nevertheless, we will see later, in Section 7.6, that pUEBPR outperforms UBPR on the ml-100k dataset when further tuning the neighborhood size. These results are somewhat surprising since while our initial aim was to improve the explainability of the recommended list, we ended up also gaining in ranking accuracy. In other words, explainability does not necessarily require sacrificing accuracy.

### 7.3 Impact of Debiasing on Performance

Contrary to what we noticed from the overall improved performance when adding explainability to any of the models, we notice a different trend in the accuracy when debiasing both models. In fact, on all three datasets, all the evaluation metrics decreased overall every time that debiasing was added: from EBPR to pUEBPR to UEBPR, and from BPR to UBPR. Hence, although the explainable models still perform better overall than the non-explainable models, debiasing explainable models seems to be degrading the ranking performance. However, as the IPS weighting aimed to mitigate the exposure bias in the training phase, the evaluation sets still suffer from exposure bias. And given that the ranking metrics are based on the interaction, rather than relevance, they cannot properly quantify the benefits of the debiasing. To truly evaluate the impact of the exposure debiasing, we evaluate the models in terms of their capacity to capture the *true relevance* which is only available in the yahoo-r3 dataset as described in the following subsection.

### 7.4 Impact of Debiasing on Relevance Modeling

The yahoo-r3 dataset provides an unbiased test set, in which a subset of 5,400 users were provided 10 random songs to rate. The fact that the songs were chosen at random ensures that the test set is free of exposure bias, because all the rated songs have the same propensity of exposure. Thus, the ratings in the unbiased test set represent the true relevance of the items to the users. Hence, evaluating a model in terms of ranking performance on this test set reflects its capacity to capture the true relevance. We re-train all the tuned models on the yahoo-r3 dataset, and evaluate it on the test set in terms of Mean Average Precision at cutoff 5 (*MAP*@5), and *NDCG*@5, where for both metrics, we assess the relevance of the top $\mathcal{K}$ predicted items for each user, given by their true rating-based ranking. We chose a cutoff of 5 because there are 10 test items per user. We summarize the results in Table 3. Almost all the unbiased models performed better than their biased versions, except for pUEBPR which performed slightly better than UEBPR. This is probably due to the nature of the neighborhood propensity estimation. However, overall, the explainable and unbiased models, pUEBPR and UEBPR, were the best performers in terms of ranking performance in an unbiased evaluation setting. This demonstrates the impact of the loss debiasing in better accounting for the true relevance.

### 7.5 Impact of Data Sparsity on Learning

In order to study the effect of the data sparsity on the performance of the explainable models, following our discussion in Section 7.1, we decided to explore the relationship between sparsity and explainability for the one data set (lastfm-2k) for which the performance trends differed. We do this by assessing the evolution of the explainability values from the explainability matrix, while gradually decreasing the sparsity of the dataset. To reduce the data sparsity, we gradually, filtered out items with fewer than a certain threshold of interactions, namely 5, 10, 15, 20, 25, 30, 35 and 40 user interactions. For each generated dataset, we compute the explainability matrix and calculate the average explainability value $E_{ui}$ in (5). We show the evolution of the average explainability with respect to the sparsity of the lastfm-2k dataset in Fig. 1. We also show the average explainability values obtained from the ml-100k and yahoo-r3 datasets for comparison purposes. The original lastfm-2k dataset

[2]In our implementation, we ended up omitting the constant denominator in the sum as this yielded better results.

**Table 2: Model comparison in terms of ranking performance and explainability on the three real interaction datasets that were described in Table 1. All evaluation metrics are computed at a cutoff $\mathcal{K}$=10 (Top 10) and reported as the averages over 5 replicates. The best results are in bold and second to best results are underlined. A value with * is significantly higher than the next best value (p-value < 0.05).**

| Dataset | ml-100k | | | | yahoo-r3 | | | | lastfm-2k | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | NDCG | HR | MEP | WMEP | NDCG | HR | MEP | WMEP | NDCG | HR | MEP | WMEP |
| BPR | 0.3807* | **0.6625** | 0.9182* | 0.3467* | 0.3315* | 0.5466 | 0.8910* | 0.1594* | **0.7260*** | **0.9086*** | 0.2142 | 0.0452 |
| UBPR | 0.3676* | 0.6401 | 0.9063* | 0.3342 | 0.3203 | 0.5422 | 0.8815 | 0.1562 | 0.6613* | 0.8340* | 0.2338 | 0.0468* |
| EBPR | **0.3821*** | 0.6568* | **0.9314** | **0.3645*** | **0.3521** | **0.5674** | **0.9461*** | **0.1808*** | 0.6309* | 0.7876* | **0.2629*** | **0.0485*** |
| pUEBPR | 0.3648* | 0.6356* | 0.9282* | 0.3595* | 0.3494* | 0.5662* | 0.9394* | 0.1778* | 0.5938* | 0.7556* | 0.2456* | 0.0471* |
| UEBPR | 0.3542 | 0.6204 | 0.8986 | 0.3332 | 0.3421* | 0.5565* | 0.9234* | 0.1710* | 0.5567 | 0.7284 | 0.2349* | 0.0461 |

**Table 3: Model comparison in terms of ranking performance on the unbiased yahoo-r3 test set: Average results over 5 replicates. The best results are in bold and second to best are underlined. A value with * is significantly higher than the next best value (p-value <0.05).**

| | BPR | UBPR | EBPR | pUEBPR | UEBPR |
|---|---|---|---|---|---|
| NDCG@5 | 0.6140 | 0.6152 | 0.6178* | **0.6187** | 0.6180 |
| MAP@5 | 0.4710 | 0.4727 | 0.4752* | **0.4764** | 0.4756 |

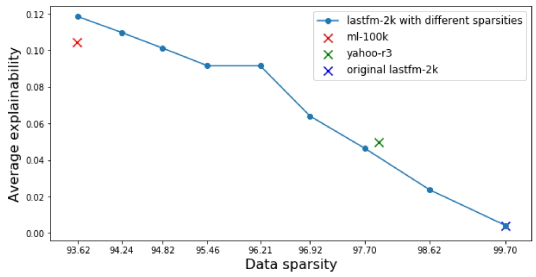


**Figure 1: Evolution of the average explainability with increasing sparsity of the lastfm-2k dataset. The average explainability values from the ml-100k and yahoo-r3 datasets are also shown for comparison.**

has an average explainability of 0.0041 which is at least one order of magnitude lower than the average explainability values of 0.1043 and 0.0497 on the ml-100k and yahoo-r3 datasets, respectively. In the explainable models (EBPR, pUEBPR and UEBPR), the explainability values are multiplication factors in the update equations (4). Hence, having explainability values that are close to 0 will cause the gradients to vanish and the learning to stall. Fig. 1 shows a decreasing linear relationship between the explainability values and the data sparsity. Moreover, when reducing the lastfm-2k data sparsity to values near the respective sparsities of the ml-100k (93.6%) and yahoo-r3 (97.9%) datasets, we obtained average explainability values near those obtained from these two datasets. Thus, the data sparsity directly affects the scale of the explainability values. Higher data sparsity leads to lower explainability values and, consequently, a higher risk of vanishing gradients. This confirms our suspicion, in Section 7.1, that the explainable models struggle with extremely sparse data due to the vanishing gradients problem.

## 7.6 Impact of Neighborhood Size on Performance

The impact of the neighborhood size is two fold: First, the neighborhood size directly impacts the explainability values of items to users, which in turn impact the values of MEP and WMEP. For that reason, we used the same neighborhood size of 20 for all models in the hyperparameter tuning. Second, the explainability values, which depend on the neighbohood size, also impact the training of the explainable models EBPR, pUEBPR and UEBPR. Thus, to compare all models fairly in terms of ranking performance, the neighborhood size must be tuned for these explainable models. In this section, we study the impact of the neighborhood size on the ranking accuracy and explainability. We vary the neighborhood size and re-train all the models in their optimal hyperparameter configurations. We show the results on the ml-100k dataset in Fig. 2. We only show the results on the ml-100k dataset to avoid clutter and because we reached similar conclusions for the other two datasets. As expected, the ranking accuracy (NDCG and HR) did not vary for the non-explainable models (BPR and UBPR) for the varying neighborhood sizes, contrarily to the explainable models (EBPR, pUEBPR and UEBPR), whose ranking prediction metrics showed different trends. EBPR and pUEBPR reached their highest ranking at a neighborhood size of 25, while UEBPR reached its maximum performance at 20. It is interesting to note that after tuning the neighborhood size, EBPR outperformed BPR and pUEBPR outperformed UBPR in both HR and NDCG which confirms our conclusions in Section 7.2, regarding the impact of the explainability weighting on the performance. The explainability metrics show opposite trends with MEP increasing and WMEP decreasing when increasing the neighborhood size. This is due to the fact that larger neighborhood sizes lead to sparser neighborhoods and thus smaller explainability values, and the latter are used as a scale in the WMEP metric. Taking aside the trends, we see that the comparative performance of the models is somewhat consistent for different neighborhood sizes: Overall, EBPR yields the best explainability performance for all neighborhood sizes, followed by pUEBPR.

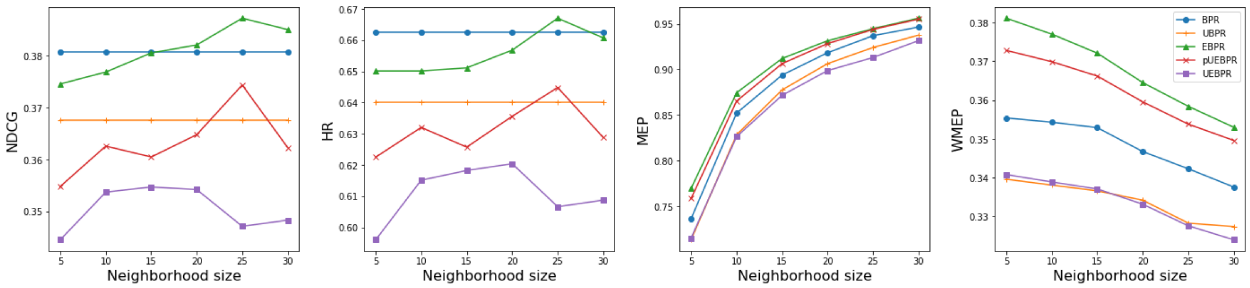


**Figure 2: Evolution of NDCG@10, HR@10, MEP@10 and WMEP@10 with increasing neighborhood size on the ml-100k dataset.**

**Table 4: Model comparison in terms of Novelty (EFD), Popularity (Avg_Pop) and Diversity (Div) on the three datasets. All evaluation metrics are computed at a cutoff $\mathcal{K}$=10 (Top 10) and reported as the averages over 5 replicates. The best results are in bold and second to best results are underlined. HB means the higher the better and LB means the lower the better. Any value with * is significantly higher than the next best value (p-value < 0.05).**

| Dataset | ml-100k | | | yahoo-r3 | | | lastfm-2k | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | EFD (HB) | Avg_Pop (LB) | Div (LB) | EFD (HB) | Avg_Pop (LB) | Div (LB) | EFD (HB) | Avg_Pop (LB) | Div (LB) |
| BPR | 1.2029 | 0.4739 | 0.2675 | 1.7681 | 0.3460 | 0.0811* | 2.7714 | 0.2000 | 0.0184 |
| UBPR | 1.3445* | 0.4397* | 0.2497* | 1.8157 | 0.3348* | **0.0789*** | 3.1049* | 0.1714* | 0.0163* |
| EBPR | 1.2160 | 0.4677* | 0.2650* | 1.7682 | 0.3442 | 0.0844 | **3.4056*** | 0.1521* | 0.0146* |
| pUEBPR | 1.2939* | 0.4491* | 0.2587* | 1.8148* | 0.3341 | 0.0822* | 3.3446 | 0.1531* | 0.0137* |
| UEBPR | **1.4699*** | **0.4127*** | **0.2414*** | **1.8716*** | **0.3222*** | 0.0800* | 3.3843* | **0.1478** | **0.0130*** |

## 7.7 Explainability as Debiasing or *Explainable Debiasing*

EBPR's superior accuracy with no apparent tradeoff with explainability suggests an inherent popularity debiasing mechanism that is a byproduct of adding explainability. This is certainly possible because the explainability term $E_{ui_+}(1 - E_{ui_-})$, when multiplied into the ranking accuracy loss, captures finer detail about an item's rating from the item's neighbors in addition to the item's own rating. This term has therefore ended up counteracting the bias of very popular items by relying on their neighborhoods. In fact, the explainability weighting term is expected to pull very popular items down, similarly to propensity debiasing. However what the proposed explainability term, ends up doing, in contrast to propensity debiasing, is succeeding in the estimation of propensity, more accurately and in a local way, namely by using the neighborhood around each item, and not solely the item itself. The advantage of the explainability term is also that it takes into account the local neighborhood to provide a rationale for both positive and negative interactions. Indeed the explainability score is not only providing intuitive quantitative explanation scores for output predictions, but also providing a rationale for debiasing, effectively providing what can be considered an *explainable local debiasing* strategy for each item. Next, we investigate this powerful idea for local explainable propensity estimation by evaluating and comparing the models in terms of Novelty (EFD), Popularity (Avg_Pop) and Diversity (Div). We summarize our results in Table 4. For all datasets and for almost all evaluation metrics, the explainable model EBPR outperformed the vanilla BPR, thus supporting our aforementioned claims of popularity debiaing with explainability weighting. Moreover, adding the exposure debiasing (moving from BPR to UBPR or moving from EBPR to pUEBPR then UEBPR) almost always improves the popularity bias metrics. This demonstrates a relationship between exposure bias and popularity bias where mitigating the former consequently mitigates the latter. Finally, UEBPR showed the best popularity debiasing overall on all the datasets. The considerably high debiasing performance of UEBPR is likely due to its down-weighting of the items with popular *neighborhoods*, in addition to the popular items, hence allowing the less popular items to be discovered. We plan to investigate this further in future work.

## 8 CONCLUSION

We proposed a novel explainable pairwise ranking loss with a corresponding MF-based model called Explainable Bayesian Personalized Ranking. We theoretically quantified the additional exposure bias resulting from the explainability, and proposed an IPS-based unbiased estimator for the ideal loss. We tested our proposed approaches on three recommendation tasks and presented an extensive discussion about the advantages of the proposed explainability extension; as well as the impact of the debiasing, for varying data sparsities and varying neighborhood sizes. Finally, we studied the popularity-debiasing properties of the proposed methods in terms of Novelty,

Popularity, and Diversity; and unveiled an inherent popularity debiasing stemming from the neighborhood interactions. Our findings are informative and motivate further research because our proposed EBPR model yielded the best performance overall with no significant trade-off between explainability and accuracy. Moreover, we showed how combining explainability and exposure debiaing yields powerful popularity debiasing through the proposed UEBPR loss. Finally, our results point towards EBPR and pUEBPR being the top performers that offer the best tradeoff between accuracy, explainability and debiasing capacity. However, despite their competitive performance, our proposed approaches may suffer from the vanishing gradient problem in extremely sparse settings.

## ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation grant IIS-1549981.

## REFERENCES


[1] Behnoush Abdollahi and Olfa Nasraoui. 2016. Explainable matrix factorization for collaborative filtering. In *Proceedings of the 25th International Conference Companion on World Wide Web*. 5–6.

[2] Behnoush Abdollahi and Olfa Nasraoui. 2017. Using explainability for constrained matrix factorization. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 79–83.

[3] Wentian Bao, Hong Wen, Sha Li, Xiao-Yang Liu, Quan Lin, and Keping Yang. 2020. GMCM: Graph-based Micro-behavior Conversion Model for Post-click Conversion Rate Estimation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2201–2210.

[4] Mustafa Bilgic and Raymond J Mooney. 2005. Explaining recommendations: Satisfaction vs. promotion. In *Beyond Personalization Workshop, IUI*, Vol. 5. 153.

[5] Iván Cantador, Peter Brusilovsky, and Tsvi Kuflik. 2011. 2nd Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011). In *Proceedings of the 5th ACM conference on Recommender systems* (Chicago, IL, USA) *(RecSys 2011)*. ACM, New York, NY, USA.

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

[7] Jiawei Chen, Can Wang, Sheng Zhou, Qihao Shi, Jingbang Chen, Yan Feng, and Chun Chen. 2020. Fast Adaptively Weighted Matrix Factorization for Recommendation with Implicit Feedback.. In *AAAI*. 3470–3477.

[8] Jiawei Chen, Can Wang, Sheng Zhou, Qihao Shi, Yan Feng, and Chun Chen. 2019. Samwalker: Social recommendation with informative sampling strategy. In *The World Wide Web Conference*. 228–239.

[9] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. 335–344.

[10] Xu Chen, Yongfeng Zhang, Hongteng Xu, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2018. Visually explainable recommendation. *arXiv preprint arXiv:1801.10288* (2018).

[11] Weiyu Cheng, Yanyan Shen, Linpeng Huang, and Yanmin Zhu. 2019. Incorporating Interpretability into Latent Factor Models via Fast Influence Analysis. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 885–893.

[12] Ludovik Coba, Panagiotis Symeonidis, and Markus Zanker. 2019. Personalised novel and explainable matrix factorisation. *Data & Knowledge Engineering* 122 (2019), 142–158.

[13] Khalil Damak, Olfa Nasraoui, and William Scott Sanders. 2021. Sequence-based Explainable Hybrid Song Recommendation. *Frontiers in Big Data* 4 (2021), 57.

[14] Robin Devooght, Nicolas Kourtellis, and Amin Mantrach. 2015. Dynamic matrix factorization with priors on unknown values. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 189–198.

[15] Aristides Gionis, Piotr Indyk, and Rajeev Motwani. 1999. Similarity search in high dimensions via hashing. In *Vldb*, Vol. 99. 518–529.

[16] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.

[17] Winston Haynes. 2013. Tukey's test. *Encyclopedia of systems biology* (2013), 2303–2304.

[18] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[19] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[20] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. 2016. Fast matrix factorization for online recommendation with implicit feedback. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 549–558.

[21] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. 241–250.

[22] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*. Ieee, 263–272.

[23] Amir H Jadidinejad, Craig Macdonald, and Iadh Ounis. 2020. Using Exploration to Alleviate Closed Loop Effects in Recommender Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2025–2028.

[24] Sami Khenissi, Boujelbene Mariem, and Olfa Nasraoui. 2020. Theoretical Modeling of the Iterative Properties of User Discovery in a Collaborative Filtering Recommender System. In *Fourteenth ACM Conference on Recommender Systems*. 348–357.

[25] Sami Khenissi and Olfa Nasraoui. 2020. Modeling and counteracting exposure bias in recommender systems. *arXiv preprint arXiv:2001.04832* (2020).

[26] Last.FM. [n.d.]. hetrec2011-lastfm-2k. https://grouplens.org/datasets/hetrec-2011/.

[27] Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. 345–354.

[28] Yanen Li, Jia Hu, ChengXiang Zhai, and Ye Chen. 2010. Improving one-class collaborative filtering by incorporating rich user information. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 959–968.

[29] Dawen Liang, Laurent Charlin, James McInerney, and David M Blei. 2016. Modeling user exposure in recommendation. In *Proceedings of the 25th international conference on World Wide Web*. 951–961.

[30] Huafeng Liu, Jingxuan Wen, Liping Jing, Jian Yu, Xiangliang Zhang, and Min Zhang. 2019. In2Rec: Influence-based interpretable recommendation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1803–1812.

[31] Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. 2018. Entire space multi-task model: An effective approach for estimating post-click conversion rate. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1137–1140.

[32] Olfa Nasraoui and Patrick Shafto. 2016. Human-algorithm interaction biases in the big data cycle: A markov chain iterated learning framework. *arXiv preprint arXiv:1608.07895* (2016).

[33] Rong Pan and Martin Scholz. 2009. Mind the gaps: weighting the unknown in large-scale one-class collaborative filtering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 667–676.

[34] Rong Pan, Yunhong Zhou, Bin Cao, Nathan N Liu, Rajan Lukose, Martin Scholz, and Qiang Yang. 2008. One-class collaborative filtering. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 502–511.

[35] Bashir Rastegarpanah, Mark Crovella, and Krishna P Gummadi. 2017. Exploring explanations for matrix factorization recommender systems. (2017).

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

[37] Yuta Saito. 2019. Unbiased Pairwise Learning from Implicit Feedback. In *NeurIPS 2019 Workshop on Causal Machine Learning*.

[38] Yuta Saito, Suguru Yaginuma, Yuta Nishino, Hayato Sakata, and Kazuhide Nakata. 2020. Unbiased Recommender Learning from Missing-Not-At-Random Implicit Feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 501–509.

[39] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. 285–295.

[40] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. *arXiv preprint arXiv:1602.05352* (2016).

[41] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. 2017. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *Proceedings of the eleventh ACM conference on recommender systems*. 297–305.

[42] Wenlong Sun, Sami Khenissi, Olfa Nasraoui, and Patrick Shafto. 2019. Debiasing the human-recommender system feedback loop in collaborative filtering. In *Companion Proceedings of The 2019 World Wide Web Conference*. 645–651.

[43] Nava Tintarev and Judith Masthoff. 2007. A survey of explanations in recommender systems. In *2007 IEEE 23rd international conference on data engineering workshop*. IEEE, 801–810.

[44] Saúl Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. 109–116.

[45] Shuo Wang, Hui Tian, Xuzhen Zhu, and Zhipeng Wu. 2018. Explainable Matrix Factorization with Constraints on Neighborhood in the Latent Space. In *International Conference on Data Mining and Big Data*. Springer, 102–113.

[46] Xiang Wang, Yaokun Xu, Xiangnan He, Yixin Cao, Meng Wang, and Tat-Seng Chua. 2020. Reinforced Negative Sampling over Knowledge Graph for Recommendation. In *Proceedings of The Web Conference 2020*. 99–109.

[47] Hong Wen, Jing Zhang, Yuan Wang, Fuyu Lv, Wentian Bao, Quan Lin, and Keping Yang. 2020. Entire Space Multi-Task Modeling via Post-Click Behavior Decomposition for Conversion Rate Prediction. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2377–2386.

[48] Yahoo! [n.d.]. Yahoo! Webscope dataset ydata-ymusic-rating-study-v1.0-train. http://research.yahoo.com/Academic_Relations.

[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] Hsiang-Fu Yu, Mikhail Bilenko, and Chih-Jen Lin. 2017. Selection of negative samples for one-class matrix factorization. In *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 363–371.

[51] Wenhao Zhang, Wentian Bao, Xiao-Yang Liu, Keping Yang, Quan Lin, Hong Wen, and Ramin Ramezani. 2020. Large-scale Causal Approaches to Debiasing Post-click Conversion Rate Estimation with Multi-task Learning. In *Proceedings of The Web Conference 2020*. 2775–2781.

[52] Yongfeng Zhang. 2015. Incorporating phrase-level sentiment analysis on textual reviews for personalized recommendation. In *Proceedings of the eighth ACM international conference on web search and data mining*. 435–440.

[53] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 83–92.