Fairness-aware Federated Matrix Factorization

Shuchang Liu
Rutgers University
New Brunswick, NJ, USA
shuchang.syt.liu@rutgers.edu

Yingqiang Ge
Rutgers University
New Brunswick, NJ, USA
yingqiang.ge@rutgers.edu

Shuyuan Xu
Rutgers University
New Brunswick, NJ, USA
shuyuan.xu@rutgers.edu

Yongfeng Zhang
Rutgers University
New Brunswick, NJ, USA
yongfeng.zhang@rutgers.edu

Amélie Marian
Rutgers University
New Brunswick, NJ, USA
amelie.marian@rutgers.edu

ABSTRACT

Achieving fairness over different user groups in recommender systems is an important problem. The majority of existing works achieve fairness through constrained optimization that combines the recommendation loss and the fairness constraint. To achieve fairness, the algorithm usually needs to know each user's group affiliation feature such as gender or race. However, such involved user group feature is usually sensitive and requires protection. In this work, we seek a federated learning solution for the fair recommendation problem and identify the main challenge as an algorithmic conflict between the global fairness objective and the localized federated optimization process. On one hand, the fairness objective usually requires access to all users' group information. On the other hand, the federated learning systems restrain the personal data in each user's local space. As a resolution, we propose to communicate group statistics during federated optimization and use differential privacy techniques to avoid exposure of users' group information when users require privacy protection. We illustrate the theoretical bounds of the noisy signal used in our method that aims to enforce privacy without overwhelming the aggregated statistics. Empirical results show that federated learning may naturally improve user group fairness and the proposed framework can effectively control this fairness with low communication overheads.

KEYWORDS

Recommendation System, Fairness, Federated Learning

ACM Reference Format:
Shuchang Liu, Yingqiang Ge, Shuyuan Xu, Yongfeng Zhang, and Amélie Marian. 2022. Fairness-aware Federated Matrix Factorization. In *Sixteenth ACM Conference on Recommender Systems (RecSys '22), September 18–23, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3523227.3546771

RecSys '22, September 18–23, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9278-5/22/09…$15.00
https://doi.org/10.1145/3523227.3546771

1 INTRODUCTION

Ensuring that decision systems provide fair results towards different users is a problem that has recently garnered considerable attention. In particular, the concept of "user group fairness" focuses on making sure that the under-represented or minority group of the users are not receiving worse outcomes than others [11, 16, 31]. In recommender systems, this is an especially important problem that has been the subject of recent work [2, 8, 46]. This focus is motivated by increasing awareness of the issue and ethical demands from users. For example, a job recommendation system that provides much more accurate results to users of one gender while neglecting or hurting the other users would be unfair. In this case, the gender attribute groups the users and and the group fairness is described by how the recommender system treat users differently according to their gender. In general, a fairness-aware recommender system should have the ability to achieve a certain degree of fairness on the recommendation quality among all user groups.

In reality, a complication is that user group features that require fairness control (such as gender and age) are likely to be sensitive at the same time. In the worst case, not many users are willing to share this protected group information with the system or other users, which makes centralized fairness control mechanically impossible. Often, the reason for a user not sharing his/her group membership is precisely the fear of being unfairly treated after revealing this information. And it is urgently needed a fairness-aware recommendation solution that protects user sensitive data.

To achieve user data protection, a reasonable idea in machine learning that has recently attracted considerable attention is the federated learning (FL) [26] technique. It allows each user's personal information to stay on its local device without being shared, and only communicates the model parameters of the machine learning model instead of the user's raw data between user devices and the central server. Though the communication of model parameters may still partially leak user information, one can overcome these potential threats with the help of encryption tools [27]. Existing works such as FedMF have shown the effectiveness of federated learning for recommendation models without fairness constraints [7, 25]. These algorithms usually represent each user as a participant, and the objective function is naturally separable by users, which is well-suited for the federated optimization process. Despite of the effectiveness of these methods on achieving accurate recommendations, there is no existing work exploring the user group fairness in the scope of federated learning on recommendation task.

In this work, we consider a general fairness metric that measures whether different user groups are treated equally by the recommendation model. Then we identify that the main challenge for this problem is the intrinsic conflict between the fairness learning goal and the data protection mechanism of FL. More specifically, the optimization of the fairness metric usually requires the access of all users' group features, so it is difficult to avoid sharing this information between the central server and users. However, the federated learning framework may need to protect this information from being shared, causing an intrinsic contradiction which has also been recognized in other machine learning tasks that involves federated fair learning [48]. As a result, we need an alternative that can effectively control the recommendation fairness even when all user group memberships are kept private in local spaces. Fortunately, the fairness metric only needs the aggregated group statistics rather than individual information of each single user. This opens for us the option of applying differential privacy (DP) techniques, which employs noisy signals to disguise the real information of users while keeping the aggregated statistics accurate. We build upon the aforementioned idea and propose a fairness-aware federated matrix factorization (F2MF) framework, and summarize our contribution as follows:

- We formally identify the conflict between user group fairness and federated learning in the recommendation problem and propose an effective solution framework (F2MF) for different attribute sharing scenarios.
- We show that the optimization of a loss-based fairness metric derives a simple algorithm that nicely fits into FL systems and potentially controls other performance-based fairness metrics.
- We further give two theoretical bounds of the added noises of the differential privacy module such that it can effectively disguise user information without overwhelming the aggregation process.
- Our observation also suggests that federated learning may naturally improve the fairness of recommendation between user groups, but the fairness become harder to control.

In the following sections, we first discuss the related work in section 2, and then describe the fairness and federated learning in recommendation as well as the aforementioned intrinsic conflict in section 3. We illustrate the loss-based fairness metric and derive our solution F2MF along with its alternatives for partially and totally private scenarios in section 4. We describe supporting experiments in section 5 and conclude our work in section 6.

2 RELATED WORK

2.1 Federated Learning in Recommendation

While some pioneer works have studied the privacy issue in recommendation systems [15], Federated Learning techniques are the first to emphasize the importance of leaving protected data in users' local spaces. There are two general scenarios of FL-based recommendation in terms of how participants are connected: they either connect to a central server forming a star-shape communication scheme [3, 25, 28]; or they form a decentralized connected network with no central server [12, 42, 43]. Our work belongs to the first scenario where each user trains a recommendation model with its local data and the central service aggregates the uploaded model parameters from users. We further illustrate the general paradigm of this type of FL scheme in section 3.2. In terms of how the user privacy is protected, existing federated recommendation systems mostly adopt encryption methods [7, 27] or differential privacy methods [37] upon communication of model parameters. These approaches are complementary to our work, since we address a FL solution to the fairness objective and assume that only the user group membership requires protection. To our knowledge, this is the first work that touches the fairness-aware recommendation problem under the FL setting.

2.2 Fairness in Recommendation

There have been growing interests on studying fairness in recommender systems as they are deeply and profoundly intertwined with people's daily lives [19, 32, 35]. Several recent works have already found various types of unfairness in recommendations, such as gender and race [5, 8, 33], item popularity [2, 21–23] and user feedback [16, 31, 34], etc. Primarily, fairness can be summarized into two paradigms based on the algorithmic definition: individual fairness and group fairness. Individual fairness requires that individuals who are similar in their features should be treated similarly, while group fairness requires that the protected groups should be treated similarly to the advantaged group or the populations as a whole. Besides, the relevant solutions to achieve fairness in ranking and recommendation can be roughly divided into three categories: pre-processing, in-processing, and post-processing algorithms [20, 32, 35]. The pre-processing methods usually aim to minimize the bias in data before the model training process. This includes fairness-aware sampling or balancing methodologies to increase coverage to minorities, repairing methodologies to ensure label correctness, and removal of disparate impact [20]. The in-processing methods aim at encoding fairness as part of the objective function, typically as a regularizer, and mitigate the bias during training [1, 4]. Our work falls into this category. Finally, post-processing methods tend to modify the presentations of the results, e.g., re-ranking through linear programming [31, 41, 45] or multi-armed bandit [6].

2.3 Federated Fair Learning

Most of the existing fair learning methods require full access to the dataset which naturally conflicts with the privacy-preserving nature of FL, since FL assumes that each participant maintains their own data proportion and may be reluctant to share the raw information [48]. Recent work [9, 14, 18] have addressed this intrinsic conflict and proposed general solutions to classification problems with different group fairness constraints. Our solution starts from a class of group fairness in the recommendation problem and takes the insights of the these methods to derive a federated learning solution. In addition, there exists other definitions of fairness discussed specifically under federated learning setting, including the participant performance fairness [30, 38] that pursues uniform accuracy across participants, and the collaboration fairness [44] that rewards participants based on their contribution. Both of these types of fairness are essentially special cases of individual fairness while our work is discussing group-wise fairness.

3 PRELIMINARIES

In this work, we consider the in-processing solution family (as described in section 2.2) that integrate the fairness metric into the objective function. This section gives the general description of this centralized formulation and formally illustrate the intrinsic conflict when it is optimized under general FL framework.

3.1 Recommendation with Differentiable Fairness Objective

We denote the set of N users as $\mathcal{U}$ and set of M items as $\mathcal{I}$ and define user group fairness in terms of the recommendation performance — a type of fairness that enforces equalized odds [17, 24], and consider optimization-based approaches that define the (un)fairness objective as the difference of the group-average performances:

$$\mathcal{L}_{\text{fair}}(G_0, G_1, \mathcal{F}) = \left| \frac{1}{|G_0|} \sum_{u \in G_0} \mathcal{F}(u) - \frac{1}{|G_1|} \sum_{u \in G_1} \mathcal{F}(u) \right|^{\rho} \quad (1)$$

Note that Eq.(1) is a bi-group metric where G_0, G_1 are two mutually exclusive user groups (e.g. active/inactive), $\rho \in \{1, 2\}$ determines the smoothness (similar to L1 or L2 norm), and $\mathcal{F}$ calculates the user-wise recommendation performance (e.g. F1 or NDCG). Since this metric tells how the recommendation model unfairly discriminates the two groups, so a smaller $\mathcal{L}_{\text{fair}}$ indicates a better model fairness. Similarly, one can define a more general multi-group metric with the number of group $K > 2$:

$$\mathcal{L}_{\text{fair}}(G, \mathcal{F}) = \frac{1}{K} \sum_{G_i} \left| \frac{1}{|G_i|} \sum_{u \in G_i} \mathcal{F}(u) - \frac{1}{K - |G_i|} \sum_{u \notin G_i} \mathcal{F}(u) \right|^{\rho} \quad (2)$$

Then, the overall objective of the fairness-aware recommendation can be formulated as the optimization problem that minimizes Eq.(3), which uses λ to trade-off the recommendation loss $\mathcal{L}_{rec}$ (e.g. BPR [40]) and the unfairness metric:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{fair}} \quad (3)$$

When setting $\lambda > 0$, minimizing Eq.(3) would minimize the unfairness along with the training process. It is worth mentioning that there are multiple choices for $\mathcal{F}(u)$ including F1, NDCG, and other reasonable learning-to-rank metrics. However, most of these metrics are non-differentiable and one might turn to constraint optimization [41] methods which are not well-suited for FL systems. Furthermore, it is plausible to find accurate and differentiable approximations (e.g. α-NDCG [10]) for some of these metrics, but it is hard to find a single metric that is correlated with other performance-based fairness metrics such that optimizing one is equivalent to optimizing all.

3.2 Federated Learning for Recommendation

In a horizontal federated learning environment, we have each participant protecting a subset of the data. This setting naturally accommodates most recommender systems since we can regard each user as a participant and apply local optimization on his/her history with local demographic features. Note that we are not assuming private user interaction history since it is more reasonable to consider it as a joint property of both the user and the service provider. In this work, we only assume that some of the user group features (e.g., age, gender, etc.) are sensitive and non-shareable. Notice that the objective Eq.(3) without the second unfairness term has property $\mathcal{L}_{\text{rec}} = \sum_u \mathcal{L}_{\text{rec}}^{(u)}$ which is already separable by users. Then, one can adopt the general federated learning paradigm illustrated as Algorithm 1, which is a distributed SGD with a user-separable loss function. The sampling step in line 4 is an algorithmic simulation of the participant dropout (e.g. connection loss) in practice. Typical aggregation functions for line 10 include but are not limited to FedAvg [36] and FedProx [29]. For simplicity of this paper, we adopt FedAvg (i.e. $\Theta^{(t+1)} \leftarrow \Theta^{(t)} + \frac{\beta}{|\mathcal{U}_{\text{subset}}|} \sum_u \nabla\Theta^{(t)}|u$), and focus on the effect of integrating fairness objective into federated learning for recommendation model. The coefficient $\beta > 0$ is the step size applied to the mitigated gradient.

Algorithm 1 General Horizontal FL for Recommendation

```
1:  procedure FedRec
2:      Input: Initial model Θ^(0)
3:      while Not Converged, in epoch t do
4:          Sample a subset of user U_subset ⊆ U
5:          for u ∈ U_subset do ▷ Compute in local space in parallel
6:              Download Θ^(t) to local space of u.
7:              ∇Θ^(t)|u ← local optimization of u based on L_rec^(u).
8:              Upload ∇Θ^(t)|u to the central server.
9:          end for
10:         Update Θ^(t+1) ←Aggregate_{∀u∈U_subset}(∇Θ^(t)|u).
11:     end while
12: end procedure
```

3.3 Natural Conflicts between Fairness and Federated Learning

Different from $\mathcal{L}_{\text{rec}}$, the fairness objective $\mathcal{L}_{\text{fair}}$ is a group-level metric and is NOT directly separable by users. Specifically, it requires the knowledge of group membership (e.g. $u \in G_0$ or $u \in G_1$) and the performance information $\mathcal{F}(u)$ of all users, so each user has to communicate this information with the central server when applying federated optimization. This mechanically contradicts the federated learning setting where user information is protected in local spaces. This critical issue could be quite common since the group features that require fairness control are likely to be sensitive as well (e.g. gender) and thus, not all users are willing to reveal this information. Besides, when engaging distributed local optimization across users, the calculation and back-propagation of $\mathcal{L}_{\text{fair}}$ require each user to wait for all other users' performance information in order to compute its local gradient. This potentially induces an impractical communication cost, and the situation is even worse when local optimization involves multiple learning steps in each epoch. In this paper, we aim to find a solution to the intrinsic conflict and avoid excessive additional communication.

4 FEDERATED RECOMMENDATION WITH FAIRNESS CONTROL

4.1 User Group Fairness under Federated Learning

As we have discussed in section 3.1, given the global fairness-aware objective defined as Eq.(1), we first adopt the substitution $\mathcal{F}(u) = -\mathcal{L}_{\text{rec}}^{(u)}$ instead of using approximation for any performance-based loss function. Here we list three advantages using this loss-based fairness objective: 1) it is empirically correlated with other performance-based metrics in our observation (with details in section 5) and intuitively, better recommendation performance correspond to less error in the learning objective; 2) the local recommendation loss is differentiable and separable by user, and thus, it nicely fits into the federated learning process as we will discuss in section 4.1; 3) it achieves fairness control with a simple modification of the gradient updates which involves little communication overhead.

For simplicity of expressions, denote the group statistics as $A = \frac{1}{|G_0|}\sum_{u\in G_0}\mathcal{F}(u)$ and $B = \frac{1}{|G_1|}\sum_{u\in G_1}\mathcal{F}(u)$. Then, we can derive the corresponding gradient of Eq.(3) with respect to each user's local model parameters with $\mathcal{F}(u) = -\mathcal{L}_{\text{rec}}^{(u)}$:

$$\begin{aligned}\nabla\Theta_u &= \frac{\partial}{\partial\Theta_u}\mathcal{L}_{\text{rec}}^{(u)} + \lambda\frac{\partial}{\partial\Theta_u}\mathcal{L}_{\text{fair}} \\ \frac{\partial}{\partial\Theta_u}\mathcal{L}_{\text{fair}} &= -C\,|A-B|^{\rho-1}\,\frac{\partial}{\partial\Theta_u}\mathcal{L}_{\text{rec}}^{(u)}\end{aligned} \tag{4}$$

where $C = \rho(-1)^{\mathbb{1}(A<B)}(-1)^{\mathbb{1}(u\notin G_0)}$, which indicates that $C > 0$ when u belongs to a group with superior performance (i.e. $A < B \wedge u \notin G_0$ or $A > B \wedge u \in G_0$), and $C \le 0$ otherwise (i.e. $A < B \wedge u \in G_0$ or $A > B \wedge u \notin G_0$). Then, combining the two loss terms in Eq.(4), we get:

$$\nabla\Theta_u = D\frac{\partial}{\partial\Theta_u}\mathcal{L}_{\text{rec}}^{(u)}, \text{ where } D = 1 - \lambda C\,|A-B|^{\rho-1} \tag{5}$$

This means that the resulting loss-based fairness objective end up scaling the gradient of $\mathcal{L}_{\text{rec}}^{(u)}$ by a scalar $D = 1-\lambda C|A-B|^{\rho-1}$ which has a simple intuitive explanation: When $\lambda > 0$, the scalar D would slow down the learning of the user u when the user belongs to the group with superior performance ($C > 0 \Rightarrow D < 1$). Otherwise, it would speed up the learning with $D \ge 1$. In other words, the low-performance group needs to learn faster and the high-performance group needs to learn slower in order to produce a better group-level fairness. In the general multi-group version with Eq.(2), the gradient is similar to Eq.(5) but the scalar D uses the average of $|A-B|^{\rho-1}$ for all Bs that come from other groups. Note that this nice intuitive explanation is taking the advantage of the setting $\mathcal{F}(u) = -\mathcal{L}_{\text{rec}}^{(u)}$. As we have described in section 3.1, there exist other feasible choices of $\mathcal{F}(u)$, but they might not have a simple derivation as Eq.(5). Besides, Θ_u represents the model parameters that are updated by user u and the derivation of Eq.(5) is model agnostic.

As the backbone model of our solution, we consider federated matrix factorization (FedMF) which has been proven effective under FL. Then the user-wise local parameters Θ_u consists of the user u's embedding and the item embeddings that are used in the training of u's embedding. We denote this solution as **F**airness-aware **F**ederated **MF** (F2MF). In this work, we focus on the effectiveness of fairness control with FedMF, and keep skeptical of how other advanced recommendation models would behave under federated system.

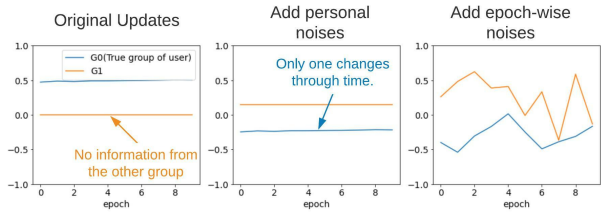

Figure 1: Example of adding user-wise and epoch-wise noises as Eq.(6) for a given user of group G_0.

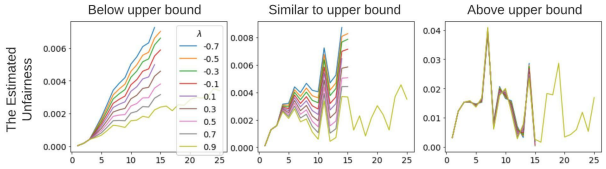

Figure 2: Example of choosing different σ for F2MF on ML-1M data with user activity level as group feature.

4.2 Communication of Group Statistics

During federated learning, the gradient computation of the $\mathcal{L}_{\text{rec}}^{(u)}$ is already feasible for local optimization, but this is not the case for the scalar D. Specifically, the value of A and B requires knowledge from other users but the federated system forbids direct sharing of this information. This is the algorithmic cause of the intrinsic conflict and the additional communication mentioned in section 3.3. To avoid revealing the user's group membership when it needs protection, we use differential privacy (DP) techniques to make each user's true information hard to infer. Specifically, we allow each user to collaboratively update the value of A and B and add noise signals when uploading information that potentially reveals group memberships. In our DP module, each user u will upload the following information:

$$\begin{aligned}\nabla A_{\text{sum}}|u &= \mathbb{1}(u\in G_0)\mathcal{F}_u + \epsilon_{1,u} + \epsilon_{A,t} \\ \nabla B_{\text{sum}}|u &= \mathbb{1}(u\in G_1)\mathcal{F}_u + \epsilon_{2,u} + \epsilon_{B,t} \\ \nabla A_{\text{count}}|u &= \mathbb{1}(u\in G_0) + \epsilon_{3,u} \\ \nabla B_{\text{count}}|u &= \mathbb{1}(u\in G_1) + \epsilon_{4,u}\end{aligned} \tag{6}$$

so that the required statistics are $A = \sum_u \nabla A_{\text{sum}}/\sum_u \nabla A_{\text{count}}$ and $B = \sum_u \nabla B_{\text{sum}}/\sum_u \nabla B_{\text{count}}$. The proposed method involves two type of noise signals: the personalized noise ($\epsilon_{1,u}, \epsilon_{2,u}, \epsilon_{3,u}, \epsilon_{4,u} \sim \mathcal{N}(0,\sigma^2)$) and the epoch-wise noise ($\epsilon_{A,t}, \epsilon_{B,t} \sim \mathcal{N}(0,\sigma^2)$). The personalized noise signals are fixed once initialized in the user's local space and do not change across epochs. Additionally, as shown in Figure 1, after adding personalized noise, the user performance of the true group changes over time but the other group stays the same. This would also expose the group membership. As a remedy, we include epoch-wise noise signals $\epsilon_{A,t}, \epsilon_{B,t}$ so that one cannot infer the membership information based on the value changes over time. The values of ∇A_{count} and ∇B_{count} never change across epochs so they do not need to apply epoch-wise noises.

Algorithm 2 F2MF with User Group Fairness

1: **procedure** F2MF_U_Group
2: **Input:** Initial model $\Theta^{(0)}$
3: Initialize $A^{(0)}, B^{(0)} \leftarrow 1$.
4: Choose σ for the random noise.
5: Each user $u \in \mathcal{U}$ initialize its own $\epsilon_{1,u}, \epsilon_{2,u}, \epsilon_{1,u}, \epsilon_{2,u}$ based on σ.
6: **while** Not Converge, in round t **do**
7: Sample a subset of user $\mathcal{U}_{\text{subset}} \subseteq \mathcal{U}$
8: **for** $u \in \mathcal{U}_{\text{subset}}$ in parallel **do**
9: Download $\Theta^{(t-1)}, A^{(t-1)}, B^{(t-1)}$ to local space of u.
10: Random sample of ϵ_{A_t} and $\epsilon_{B,t}$.
11: $\nabla\Theta, \nabla A_{\text{sum}}, \nabla B_{\text{sum}}, \nabla A_{\text{count}}, \nabla B_{\text{count}} \leftarrow$ local optimization of u based on Eq.(4) and Eq.(6).
12: Upload all updates in line 10 to the central server.
13: **end for**
14: // Aggregation on central server
15: Update $\Theta^{(t)} \leftarrow$Aggregation$(\nabla\Theta|u, \forall u \in \mathcal{U}_{\text{subset}})$.
16: $A^{(t)} \leftarrow \frac{\sum_{u\in\mathcal{U}_{\text{subset}}} \nabla A_{\text{sum}}|u}{\sum_{u\in\mathcal{U}_{\text{subset}}} \nabla A_{\text{count}}|u}$
17: $B^{(t)} \leftarrow \frac{\sum_{u\in\mathcal{U}_{\text{subset}}} \nabla B_{\text{sum}}|u}{\sum_{u\in\mathcal{U}_{\text{subset}}} \nabla B_{\text{count}}|u}$
18: **end while**
19: **end procedure**

Based on the weak law of large numbers, the aggregated noises from all users will have $\Pr(|\lim_{N\to\infty} \bar{\epsilon} - \mathbb{E}[\epsilon]| < \delta) = 1$ for some close-to-zero positive value δ, and it holds for all noise signals $\epsilon \in \{\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_{A,t}, \epsilon_{B,t}\}$. In other words, even if a large noise deviates each user's uploaded response far from its real value, the aggregated statistics tend to be close to the ground-truth as long as each group has a sufficient number of users. This technique is a simple application of the differential privacy [13] which enables the system to learn from the entire population as a whole while protecting the privacy of each individual. Note that only including epoch-wise noise is not a robust choice due to the law of large numbers. The detailed procedure of the overall framework is summarized by algorithm 2. For the multi-group case, group statistics of all groups are synchronized for line 9, line 10, and line 15-16.

4.3 Choosing Standard Deviation of Noises

In the **partially private** scenario, only a small number of users deny the sharing of their group memberships, while most users agree with sharing it. Then the framework can systematically ignore users that deny uploads of this information, and for all other users that share this information, there is no need to apply the noise, so we can simply set $\sigma = 0$ resulting in $\epsilon = 0$. We denote this solution for the partially private scenario as Free-sharing Federated Fairness Recommendation (**F3MF**) which is a special case of F2MF. Without the noise signal, the resulting process is always accurate in calculating A and B for the available users, and also accurate for the corresponding gradient calculation in Eq.(4). Readers may also notice that the denial of user uploads might be related to the user's performance or group feature, but in this work, we are assuming "missing at random" for the missing uploads of users since there would not be a significant change of the group differences when the missing cases are rare.

In the more reasonable **totally private** scenario where most users require protection of the group membership, F2MF should enforce differential privacy by including the noise terms in Eq.(6) but with two statistical constraints:

(The **lower bound**): On one hand, for the purpose of privacy protection, σ must be *sufficiently large* such that the probability of correct inference of the user's group feature in the central server is low. A typical inference rule would be guessing the user's group by the largest uploaded performance. We hope to lower the confidence of this rule and make it close to random guess so that $\Pr(Z > 0) < 0.5 + \delta_1$ for some small positive constant δ_1, which derives:

$$\sigma \geq \frac{-\mathcal{F}_u}{\sqrt{2}\Phi^{-1}(0.5 - \delta_1)} = \frac{\mathcal{F}_u}{\sqrt{2}\Phi^{-1}(0.5 + \delta_1)} \tag{7}$$

where $\Phi(x)$ is the normal cumulative density function (the cumulative density function of the standard normal distribution $\mathcal{N}(0, 1)$). This lower bound indicates that setting a sufficiently large σ will significantly weaken the inference rule to be close to a random guess and reduce the attacker's confidence.

(The **upper bound**): On the other hand, for the purpose of correct calculation of A and B, σ should also be *sufficiently small* so that the aggregated noise cannot easily dominate the differences of the aggregated sum or count. Formally, we want the chances that the aggregated noise accounting for more than H (e.g. representing the difference of group-wise performance) portion of the ground-truth is less than δ_2, then we can find that:

$$\sigma \leq H|\bar{X}_{\text{actual}}|\sqrt{N\delta_2} \tag{8}$$

where N is the number of involved users and $|\bar{X}_{\text{actual}}|$ is the true value of the average performance which can be obtained from empirical studies. Notice that smaller values of H, δ, and N indicate a more tightened bound. To further illustrate the noise dominance, we give a showcase in the rightmost plot in Figure 2 where σ is too large and the actual group performance becomes negligible compared to the aggregated noise.

Finally, in order to select σ in a valid range, one should ensure that the lower bound is no larger than the upper bound (i.e. $1/(\sqrt{2}\Phi^{-1}(0.5 + \delta_1)) \leq H|\bar{X}_{\text{actual}}|\sqrt{N\delta_2}$. And one can achieve a larger feasible region of σ by relaxing the setting of H, δ_1, and δ_2 or increasing the number of users N. We give the details of the derivations in appendix A.

4.4 Delayed Information of Statistics

In order to avoid excessive additional communication of the statistics in Eq.(5) in the middle of the gradient calculation of each local optimization step, our solution allows the central server to store and use aggregated statistics from the previous round/epoch. At the beginning of a user's local round, the local space first synchronizes this one-round-behind information as line 9 in algorithm 2, then upload the updated information to the central server for the next round as line 15-18. In our experiments, the delayed information still works and effectively serves as a guide to control fairness, since the performance converges and the two consecutive rounds tend to have similar statistics when the training is stable.

Dataset	$\lvert\mathcal{U}\rvert$	$\lvert\mathcal{I}\rvert$	#record	sparsity	user feature	#group
ML-1M	6,022	3,043	995,154	0.9457	gender	2
					activity	2
					age	5
Movies	5,515	13,509	484,141	0.9935	activity	2

Table 1: Dataset Summary. "activity" of user is defined based on number of interactions, the top 20% are "active", others are "inactive".

5 EXPERIMENTS

5.1 Experimental Setting

Dataset: In our experiments, we include two public real-world datasets MovieLens-1M[1] (ML-1M), and Amazon-Movies[2] (Movies) [39]. The properties of these datasets are summarized in Table 1. For all datasets, we first filter n-core data, and then adopt 80%-10%-10% split for each user history based on temporal order and remove unseen items in the validation and test set to the training set. For the selection of user group features, we first consider a synthetic attribute — activity level of users for both datasets same as existing literature [16, 31], and then we also include two given attributes gender and age of users in ML-1M data to show the behavior of fairness on different types of user group features. In our experiments, we consider both partially private and totally private scenarios for all three selected user group features to verify the effective learning process across settings of our method. Yet, we remind readers that in practice, one may consider activity level as a shareable feature, user age as partially shareable, and user gender as totally private.

Models and Baselines: We consider the **FedMF** [7] model (ignoring encryption) as the backbone recommendation model and include **MF** [40] model as the centralized counterpart. We implemented the **F2MF** model that integrates FedMF model into our fairness-aware learning framework, and include **F3MF** for the partially private scenario as described in section 4.3. We also include a centralized counterpart for F2MF, denoted as **FairMF**, which optimizes Eq.(3) in a centralized environment without federated learning. For all federated environment, we set dropout rate = 0.1 (i.e. $|\mathcal{U}_{\text{subset}}|/|\mathcal{U}| = 0.9$) and one local learning step per user per epoch. Note that MF is different from all other solutions for its stochastic mini-batch training process, while all other models apply user-wise gradient descent. For all models, we adopt BPR [40] loss for $\mathcal{L}_{\text{rec}}^{(u)}$ and use 1 negative item per user-item interaction during training, 100 negative items per user for validation, and all items per user for the test set. We provide implementation details with the source code [3].

Evaluation: For recommendation performance, we choose **Recall**, **F1**, and **NDCG** as the evaluation metrics for top-k recommendation on the 10% test set where $k \in \{1, 5, 10, 50\}$. A higher score for any of these metrics would be an indicator of a better recommendation model. We train each model until its recommendation performance (i.e. NDCG@50) converges on the validation set and select the model with the best performances. For fairness evaluation, we observe that $\rho = 2$ is smooth but usually contributes trivial changes to the optimization when $|A - B|$ is small. Thus, we directly use Eq.(1) with $\rho = 1$ and larger value indicates greater **Unfairness**. For the choices of $\mathcal{F}(u)$ in the unfairness metric, we include the recommendation loss, F1, NDCG, and Recall in order to observe how different metrics may be correlated in terms of the accuracy-fairness trade-off. For reproducibility, we provide source code in supplementary.

[1] https://grouplens.org/datasets/movielens/1m/
[2] https://nijianmo.github.io/amazon/index.html
[3] https://github.com/CharlieMat/FedFairRec.git

5.2 Effectiveness of Fairness Control

We first consider the user activity level as the group feature for both datasets to showcase the effectiveness of controlling fairness of recommendation with $\lambda \in [-0.7, 0.9]$. We include negative λ just to further observe the continuation of the trend of model unfairness behavior and recommendation accuracy. Intuitively, When applying negative λ, the learning objective will no longer suppress but encourage the unfairness. Main recommendation results are summarized in Figure 3 and main comparison of the fairness control in Figure 4. The FairMF can achieve relatively the same level of accuracy as MF in ML-1M and higher performance in Movies data. When tuning λ, FairMF tends to achieve the best performance around $\lambda = 0$ in the Movie dataset, around $\lambda = -0.3$ in ML-1M dataset on top-10 performances, and around $\lambda = +0.3$ in ML-1M dataset on top-50 performances. F2MF shows a similar pattern with peak accuracy around $\lambda = 0$ in ML-1M, but shows a more stable accuracy-fairness trade-off in Movie.

A threshold for stable control: For both FairMF and F2MF, when the absolute value $|\lambda|$ is larger than some certain threshold, the fairness control will drastically impact the recommendation accuracy, generating an unstable and inconsistent behavior. We point out the thresholds (if we observed one) as red circles in Figure 4 which shows how the estimated unfairness over epochs (the number of epochs depends on model convergence). We observe that when λ is larger than the threshold, the estimated unfairness is suppressed to almost zero and groups become indistinguishable in performance (extremely small H and thus extremely tight upper bound for σ). Meanwhile, a large absolute value of λ also over-amplifies the relative differences between group-wise gradients (after scaling by D). This results in frequent swaps of D and unstable training curves, as shown in the last row in Figure 4.

Controlling fairness: Taking activity-level as an example of group features, when we increase λ under the threshold, we can see that both FairMF and F2MF can **effectively and consistently** reduce the value of unfairness to almost 100%, as shown in the first and last column of Figure 4. And decrease λ to the negative region can increase the unfairness to 50%, but as shown in Figure 3 it does not always correspond to an increase of recommendation accuracy, indicating a potential "sweat point" for the accuracy-fairness trade-off. We observe similar behavior for FairMF in ML-1M when using gender and age as the group feature for fairness control. Specially for user age, the swapping behavior of D for FairMF is even more frequent and chaotic, indicating a harder control when there are more than two groups. This contracted feasible region may be related to the fact that some user groups can still be close to the

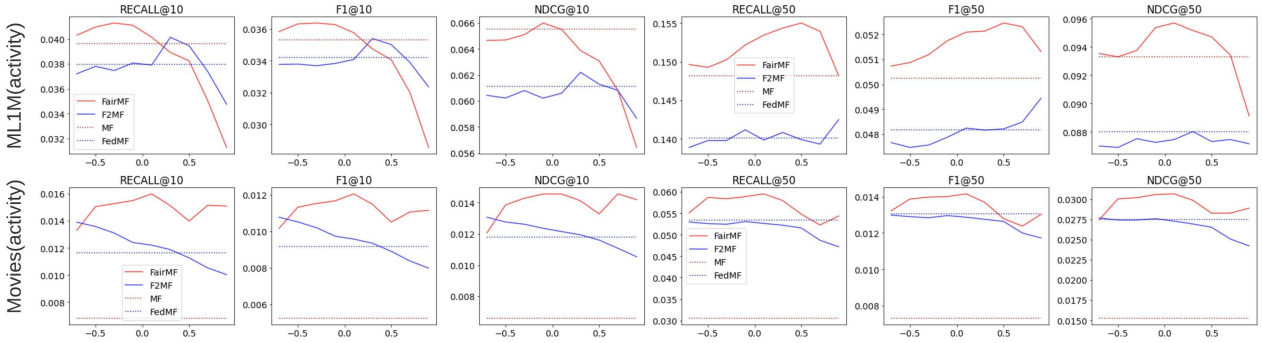

Figure 3: Recommendation performance when controlling fairness through λ (X-axis). F2MF and F3MF performs almost identically.

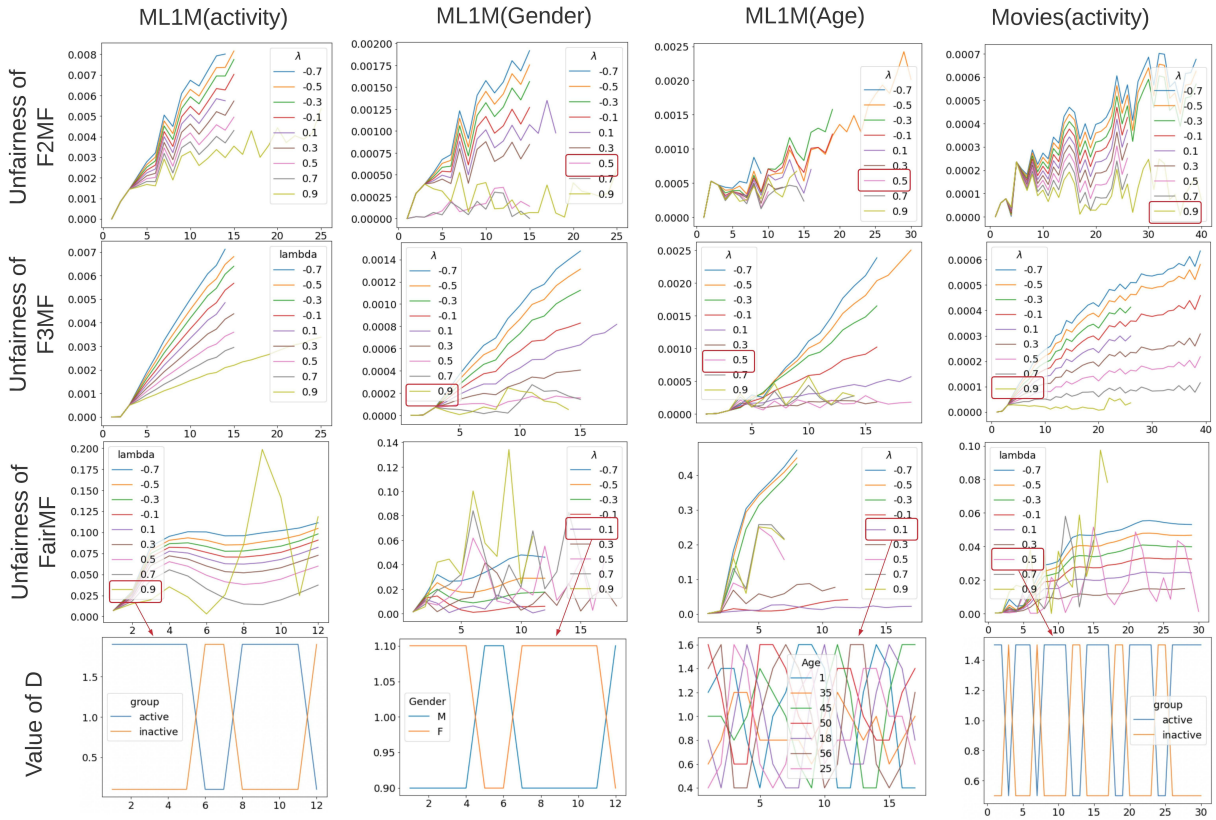

Figure 4: Estimated unfairness using $\mathcal{F}_u = 1 - \mathcal{L}_{\text{rec}}^{(u)}$ (first three rows) and the swap of D when setting large λ (last row). X-axis are epochs. Red circles represents the threshold where larger λ start causing unstable learning process. With increasing λ, curves becomes lower with stable D value until the threshold reached.

averaged performance even if the overall group difference is large, causing frequent swaps of D as shown in the last row of Figure 4.

5.3 Effect of Federated Learning

Note that we can consider FairMF with $\lambda = 0$ as a more precise centralized counterpart for FedMF since they both adopt user-wise training. With this notion, FedMF appears to be sub-optimal in recommendation accuracy compared to FairMF ($\lambda = 0$) as shown in Figure 3. On the other hand, in Figure 4, the absolute unfairness (Y-axis) of F2MF and F3MF (regardless of the value of λ) is usually significantly lower than FairMF and becomes almost zero across all datasets. This indicates that the federated learning process can drastically improve the fairness of the system. One reason is that the

FedAvg method aggregates user uploads with equal weights by default in horizontal FL, and this implicitly balances the performance of each user which eventually mitigates the group differences. In contrast, the accuracy-fairness trade-off becomes less obvious and harder to control because of the reduced magnitude of unfairness by FL [47].

5.4 Partially Private vs. Totally Private

In reality, one might consider the user activity level as a non-sensitive feature and consider gender and age features as sensitive ones. As we have discussed in section 4.3, we can apply F3MF for non-sensitive features which correspond to the partially private scenario and apply F2MF ($\sigma > 0$) for sensitive features under totally private scenarios. As we have explained in section 4.2, including a sufficiently large number of users in training will always make the averaged noise close to zero and the F2MF model will perform exactly like the F3MF model. Mathematically, if we consider each user in the dataset as equivalent to 10,000 users (i.e. 6,022 dataset users in ML-1M represents 60,220,000 users), so setting $\sigma = 1.0$ for each imaginary user is equivalent to setting $\sigma = 1/\sqrt{10000} = 0.01$ for each user in the dataset. And the resulting F2MF becomes indistinguishable with F3MF for recommendation accuracy. Their similar fairness control can also be observed in Figure 4 except for age feature in ML-1M where the noise in F2MF is indeed more influential on ML-1M (Age) because of the increased number of groups. Additionally, a system that is already fair may also be disturbed by this noise since group difference H is small and the upper bound is harder to suffice. In this case, we can still find a feasible σ that will not dominate the group difference in the cost of including more users in the training.

5.5 Correlations Between Metrics

Though we have shown effective fairness control on $\mathcal{F}(u) = -\mathcal{L}_{\text{rec}}^{(u)}$, the definition of Eq.(1) also allows other choices of $\mathcal{F}(u)$. Here, we show results for F1 and Recall with recommendation list size in $\{1, 10, 50\}$ on ML-1M as examples and plot the results for FairMF and F2MF as Figure 5 and Figure 6. When evaluating user gender group fairness, all selected metrics tend to improve on fairness when increasing λ (below the threshold mentioned in section 5.2) for both FairMF and F2MF. Yet, the model behavior is no longer consistent across metrics when using user activity level as the group feature, where F1@1, Recall@1, and Recall@10 tend to improve while other metrics show diverging group performances. This indicates that negative correlations between the loss and certain metrics do exist. For multi-group feature user age, we find that the recommendation accuracy of F2MF and F3MF are more stable than FairMF and the improvement of the fairness is more consistent (third row of Figure 5 and Figure 6), but it also becomes harder to control (fairness change < 10%) for the effect of FL. In general, it is hard to find a universal fairness metric that is consistent with all other metrics but one can usually control them towards a certain direction by tuning the loss-based metric.

5.6 Complexity

As we have discussed in section 4.4, the F2MF framework uses the same FedAvg communication protocol as FedMF, and the only extra information to communicate between central service and user devices is the statistics of user groups (line 9 and line 11 in Algorithm 2). In each local optimization, there is no extra loss term to calculate and the method only scales the gradient by the scalar D that can be calculated in $O(\#\text{group})$. In the view of each user, this corresponds to a time and space complexity that only depends on the number of user groups, which is asymptotically negligible compared to the transfer of model parameters. In the view of the central server, the overall communication and computational cost of each epoch induced by the fairness term is $O(NK)$, where N is the number users and K is the number of groups. Note that K is usually a small constant integer in practice, so this extra complexity is much smaller than that of the number of model parameters.

6 CONCLUSION

The fairness objective in recommender systems intrinsically conflicts with the federated learning paradigm. In this work, we have shown that one can integrate the learning goal of recommendation with a loss-based fairness metric and derive a simple and effective federated solution for fairness-aware recommendation. The solution induces little communication and computation overhead on the backbone FL of the recommendation model. We theoretically show the feasible parameter region of the DP module, and empirically show that it can effectively control the user group fairness in terms of the loss-based metric, which indirectly control other performance-based fairness metrics. While our method shed light on how one can solve federated fairness-aware recommendation during optimization phase, we believe it is worth further exploring the alternatives in the pre-processing and post-processing phases as well.

A APPENDIX: DERIVATION OF BOUNDS

A.1 The Lower Bound of σ: In the most extreme scenario, where the number of iteration approaches infinity and the performance converges, then one can statistically eliminate the epoch-wise noise: ($\lim_{T\to\infty}\sum_{t\in[1,T]}\epsilon_{A,t}/T = 0$ and $\lim_{T\to\infty}\sum_{t\in[1,T]}\epsilon_{B,t}/T = 0$) and figure out that $\nabla A_{\text{sum}} = \mathbb{1}(u \in G_0)\mathcal{F}_u + \epsilon_{1,u}, \nabla B_{\text{sum}} = \mathbb{1}(u \in G_1)\mathcal{F}_u + \epsilon_{2,u}, \nabla A_{\text{count}} = \mathbb{1}(u \in G_0) + \epsilon_{3,u}, \nabla B_{\text{count}} = \mathbb{1}(u \in G_1) + \epsilon_{4,u}$. Without loss of generality, assume that $u \in G_0$ and the rule of inference attack is $(\nabla A_{\text{sum}}|u > \nabla B_{\text{sum}}|u) \Rightarrow u \in G_0$. Then the confidence of a correct outsider inference (happens when observing $\nabla A_{\text{sum}}|u > \nabla B_{\text{sum}}|u$) is given by:

$$\Pr(Z \geq 0), \text{where } Z \sim \mathcal{N}(\mathcal{F}_u, 2\sigma^2) \tag{9}$$

where Z is the value of random variables $A_{\text{sum}} - B_{\text{sum}}$ and it always have positive mean $\mathcal{F}_u$ since we assume A_{sum} as the correct group G_0. For privacy protection, we aim for $\Pr(Z > 0) < 0.5 + \delta_1$ for some small positive δ_1, which derives:

$$\begin{aligned} &\Pr(Z \leq 0) \geq 0.5 - \delta_1 \\ \Leftrightarrow\ &\Pr\left(\frac{Z - \mathcal{F}_u}{\sqrt{2}\sigma} \leq \frac{-\mathcal{F}_u}{\sqrt{2}\sigma}\right) \geq 0.5 - \delta_1 \\ \Leftrightarrow\ &\Phi\left(\frac{-\mathcal{F}_u}{\sqrt{2}\sigma}\right) \geq 0.5 - \delta_1 \Leftrightarrow \frac{-\mathcal{F}_u}{\sqrt{2}\sigma} \geq \Phi^{-1}(0.5 - \delta_1) \\ \Leftrightarrow\ &\sigma \geq \frac{-\mathcal{F}_u}{\sqrt{2}\Phi^{-1}(0.5 - \delta_1)} = \frac{\mathcal{F}_u}{\sqrt{2}\Phi^{-1}(0.5 + \delta_1)} \end{aligned} \tag{10}$$

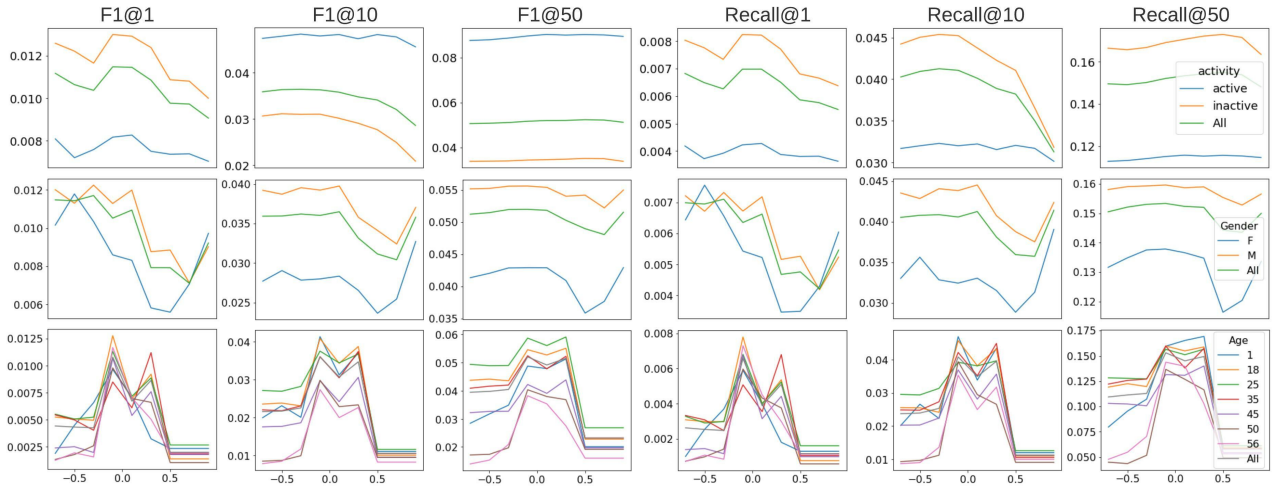

Figure 5: Group-wise performances of FairMF on ML-1M dataset. Each row correspond to a selected group feature. X-axis are values of λ and shared among rows, Y-axis are values of the corresponding metric. Performance metrics of different groups may contract or diverge when increasing λ, and becomes unstable after reaching the threshold.

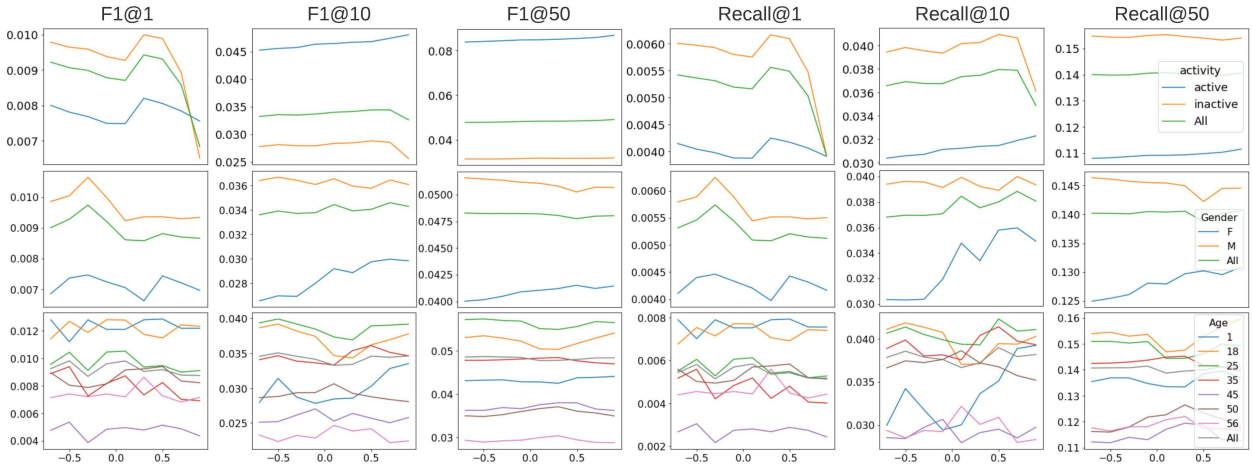

Figure 6: Group-wise performances of F2MF (and F3MF) on ML-1M dataset. Each row correspond to a selected group feature. X-axis are values of λ and shared among rows.

where $\Phi(x)$ is the cumulative density function of the standard normal distribution $\mathcal{N}(0,1)$. Similarly, for count information, we assume $\mathcal{F}_u \in (0,1]$ that is no larger than a count signal 1. Thus, we derive a larger lower bound $\sigma \geq 1/(\sqrt{2}\Phi^{-1}(0.5+\delta_1)) \geq \mathcal{F}_u/(\sqrt{2}\Phi^{-1}(0.5+\delta_1))$.

A.2 The Upper Bound of σ: Denote $\bar{X}$ as the average of one of the four values in Eq.(6) without epoch-aware noises ($\epsilon_{A,t}$ and $\epsilon_{B,t}$), then $\bar{X} \sim \mathcal{N}(\bar{X}_{\text{actual}}, \frac{\sigma^2}{N})$ which is an aggregation of Gaussian variables. Here we use the average value instead of the summation to better illustrate the relative influence of σ and the number of users N. Assume that the average performance/loss value or the count is within $(0,1]$, and the average ground-truth value $\bar{X}_{\text{actual}}$ (without noises) is not close to zero. Formally, we denote $H \in (0, 0.1)$ as the ratio between the difference and the absolute value of group-wise performances, so form some small constant δ_2, we want $\Pr(|\bar{X} - \bar{X}_{\text{actual}}| \geq H|\bar{X}_{\text{actual}}|) \leq \delta_2$. Note that the Chebyshev's Inequality gives a more strict upper bound for $Pr(|\bar{X} - \bar{X}_{\text{actual}}| \geq H|\bar{X}_{\text{actual}}|)$, so we can set:

$$\frac{\sigma^2}{NH^2|\bar{X}_{\text{actual}}|^2} \leq \delta_2 \Leftrightarrow \sigma \leq H|\bar{X}_{\text{actual}}|\sqrt{N\delta_2} \tag{11}$$

where $|\bar{X}_{\text{actual}}|$ can be obtained by empirical results.

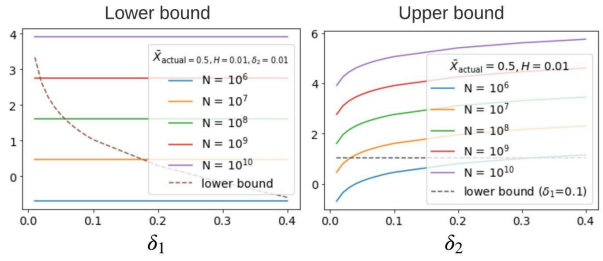

Figure 7: Y-axis is in log σ. Left panel gives the lower bounds (dotted lines) over δ_1. Right panel shows upper bounds (solid lines) over δ_2.

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] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286* (2019).

[3] Muhammad Ammad-Ud-Din, Elena Ivannikova, Suleiman A Khan, Were Oyomno, Qiang Fu, Kuan Eeik Tan, and Adrian Flanagan. 2019. Federated collaborative filtering for privacy-preserving personalized recommendation system. *arXiv preprint arXiv:1901.09888* (2019).

[4] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD*.

[5] Jesús Bobadilla, Raúl Lara-Cabrera, Ángel González-Prieto, and Fernando Ortega 0001. 2021. DeepFair: Deep Learning for Improving Fairness in Recommender Systems. *IJIMAI* 6, 6 (2021), 86–94. https://doi.org/10.9781/ijimai.2020.11.001

[6] L Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. 2019. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of the conference on fairness, accountability, and transparency*. 160–169.

[7] D. Chai, L. Wang, K. Chen, and Q. Yang. 2021. Secure Federated Matrix Factorization. *IEEE Intelligent Systems* 36, 05 (sep 2021), 11–20. https://doi.org/10.1109/MIS.2020.3014880

[8] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the Impact of Gender on Rank in Resume Search Engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174225

[9] Lingyang Chu, Lanjun Wang, Yanjie Dong, Jian Pei, Zirui Zhou, and Yong Zhang. 2021. FedFair: Training Fair Models In Cross-Silo Federated Learning. *arXiv preprint arXiv:2109.05662* (2021).

[10] Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 659–666.

[11] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. https://doi.org/10.1145/3097983.3098095

[12] Erika Duriakova, Elias Z Tragos, Barry Smyth, Neil Hurley, Francisco J Peña, Panagiotis Symeonidis, James Geraci, and Aonghus Lawlor. 2019. PDMFRec: a decentralised matrix factorisation with tunable user-centric privacy. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 457–461.

[13] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*. Springer, 1–19.

[14] Yahya H. Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and Salman Avestimehr. 2021. FairFed: Enabling Group Fairness in Federated Learning. (2021). arXiv:cs.LG/2110.00857

[15] Arik Friedman, Bart P Knijnenburg, Kris Vanhecke, Luc Martens, and Shlomo Berkovsky. 2015. Privacy aspects of recommender systems. In *Recommender systems handbook*. Springer, 649–688.

[16] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, and Gerard de Melo. 2020. Fairness-aware explainable recommendation over knowledge graphs. In *Proceedings of the 43rd SIGIR*. 69–78.

[17] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, and Gerard de Melo. 2020. Fairness-Aware Explainable Recommendation over Knowledge Graphs *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 69–78. https://doi.org/10.1145/3397271.3401051

[18] Borja Rodríguez Gálvez, Filip Granqvist, Rogier van Dalen, and Matt Seigel. 2021. Enforcing fairness in private federated learning via the modified method of differential multipliers. In *NeurIPS 2021 Workshop Privacy in Machine Learning*.

[19] Ruoyuan Gao, Yingqiang Ge, and Chirag Shah. 2022. FAIR: Fairness-aware information retrieval evaluation. *Journal of the Association for Information Science and Technology* (2022).

[20] Ruoyuan Gao and Chirag Shah. 2021. Addressing Bias and Fairness in Search Systems. In *Proceedings of the 44th International ACM SIGIR (SIGIR '21)*. 4. https://doi.org/10.1145/3404835.3462807

[21] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, and Yongfeng Zhang. 2021. Towards Long-Term Fairness in Recommendation. In *Proceedings of the 14th WSDM (WSDM '21)*. 445–453. https://doi.org/10.1145/3437963.3441824

[22] Yingqiang Ge, Juntao Tan, Yan Zhu, Yinglong Xia, Jiebo Luo, Shuchang Liu, Zuohui Fu, Shijie Geng, Zelong Li, and Yongfeng Zhang. 2022. Explainable Fairness in Recommendation. *arXiv preprint arXiv:2204.11159* (2022).

[23] Yingqiang Ge, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. 2022. Toward Pareto Efficient Fairness-Utility Trade-off in Recommendation through Reinforcement Learning. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 316–324.

[24] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.

[25] Amir Jalalirad, Marco Scavuzzo, Catalin Capota, and Michael Sprague. 2019. A simple and efficient federated recommender system. In *Proceedings of the 6th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*. 53–58.

[26] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492* (2016).

[27] Dongsheng Li, Chao Chen, Qin Lv, Li Shang, Yingying Zhao, Tun Lu, and Ning Gu. 2016. An algorithm for efficient privacy-preserving item-based collaborative filtering. *Future Generation Computer Systems* 55 (2016), 311–320.

[28] Mu Li, Ziqi Liu, Alexander J Smola, and Yu-Xiang Wang. 2016. Difacto: Distributed factorization machines. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. 377–386.

[29] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *In proceeding of Machine Learning and System* (2020).

[30] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. 2020. Fair Resource Allocation in Federated Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=ByexElSYDr

[31] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented Fairness in Recommendation. In *Proceedings of WWW 2021*.

[32] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. 2022. Fairness in Recommendation: A Survey. *arXiv preprint arXiv:2205.13619* (2022).

[33] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2021. Towards personalized fairness based on causal notion. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1054–1063.

[34] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2021. Towards Personalized Fairness Based on Causal Notion. In *Proceedings of the 44th SIGIR (SIGIR '21)*. 10. https://doi.org/10.1145/3404835.3462966

[35] Yunqi Li, Yingqiang Ge, and Yongfeng Zhang. 2021. Tutorial on Fairness of Machine Learning in Recommender Systems. In *Proceedings of the 30th CIKM*.

[36] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

[37] Xuying Meng, Suhang Wang, Kai Shu, Jundong Li, Bo Chen, Huan Liu, and Yujun Zhang. 2018. Personalized privacy-preserving social recommendation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[38] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. 2019. Agnostic federated learning. In *International Conference on Machine Learning*. PMLR, 4615–4625.

[39] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 188–197.

[40] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings*

of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI '09). 452–461.
[41] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD*.
[42] Amund Tveit. 2001. Peer-to-peer based recommendations for mobile commerce. In *Proceedings of the 1st international workshop on Mobile commerce*. 26–29.
[43] Paul Vanhaesebrouck, Aurélien Bellet, and Marc Tommasi. 2017. Decentralized collaborative learning of personalized models over networks. In *Artificial Intelligence and Statistics*. PMLR, 509–517.
[44] Guan Wang, Charlie Xiaoqian Dang, and Ziye Zhou. 2019. Measure contribution of participants in federated learning. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 2597–2604.
[45] Tao Yang and Qingyao Ai. 2021. Maximizing Marginal Fairness for Dynamic Learning to Rank. In *Proceedings of the Web Conference 2021*. 137–145.
[46] Sirui Yao and Bert Huang. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. In *Advances in Neural Information Processing Systems*.
[47] Yuchen Zeng, Hongxu Chen, and Kangwook Lee. 2021. Improving fairness via federated learning. *arXiv preprint arXiv:2110.15545* (2021).
[48] Zirui Zhou, Lingyang Chu, Changxin Liu, Lanjun Wang, Jian Pei, and Yong Zhang. 2021. Towards Fair Federated Learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 4100–4101.