Private but Biased? Exploring Fairness in Federated Recommendation

Anonymous authors
Paper under double-blind review

Abstract

Bias and fairness are central concerns in machine learning, particularly in recommendation systems that may reinforce gender, age, or occupation stereotypes. In parallel, federated recommendation has emerged as a privacy-preserving alternative to centralized systems, particularly in cross-device settings where data remains on user devices. Despite extensive studies of bias in centralized recommendation, its behavior under federated training remains largely underexplored. Indeed, the transition to a decentralized architecture introduces additional sources of statistical skew and uneven representation across users, making the impact of federated learning on bias dynamics unclear. Furthermore, most existing bias mitigation techniques rely on sharing sensitive user or item attributes, which conflicts with the privacy constraints inherent to federated learning. In this work, we investigate how federated training influences the emergence of gender and user activity bias in cross-device federated recommendation systems. We further adapt an existing bias mitigation approach to the federated setting and propose a privacy-aware framework for bias mitigation that does not require sharing sensitive attributes. Our results show that, under certain conditions, federated training can introduce less bias than its centralized counterpart. Across datasets and model families, we observe a consistent reduction in gender-based bias under FL settings, while popularity (activity) bias exhibits model-dependent behavior and can increase in graph-based user-expansion methods. Moreover, we demonstrate that effective bias mitigation is feasible in federated recommendation while preserving user privacy. Our Source code, datasets and trained models are available at `https://anonymous.4open.science/r/Bias-and-cross-device-Federated-recommendation-50D5/`

1 Introduction

Recommender systems are no longer evaluated solely by their predictive accuracy. Beyond performance, a growing body of work has shown that recommendation models can systematically amplify social and behavioral biases, affecting both users and content providers Boratto et al. (2022); Ekstrand et al. (2018); Wei & He (2022). These biases manifest in forms such as popularity amplification, exposure imbalance, feedback loops, and disparities linked to sensitive user attributes Boratto et al. (2022). As a result, fairness-aware recommendation has emerged as a major research direction, with numerous methods proposed to mitigate biases related to gender, age, or user activity Ekstrand et al. (2018); Kheya et al. (2025); Boratto et al. (2022). However, most of these approaches assume centralized access to user interactions and global group statistics. The increasing adoption of Federated Learning (FL) for recommendation fundamentally alters this setting: training is decentralized across clients Wang et al. (2024), sensitive attributes remain local, and global statistics are no longer directly observable. As a result, the assumptions underlying many existing fairness interventions no longer hold.

Federated recommendation can be instantiated in two main paradigms: *cross-silo FL* Kalloori & Klingler (2021), where a limited number of organizations collaboratively train a shared model, and *cross-device FL* Sun et al. (2024), where each client corresponds to an individual user device. In this paper, we focus on the cross-device setting, which introduces distinctive bias dynamics compared to centralized training. Indeed, in cross-device FL each client updates the model solely based on its own interaction history, and the

global model is formed by aggregating gradients from a dynamically sampled subset of users. Because user interaction data is inherently heterogeneous and often correlated with demographic or behavioral attributes, local updates may reflect skewed group-specific patterns. Moreover, client participation is intermittent and non-uniform, meaning that certain user groups may be over- or under-represented in different training rounds. These dynamics introduce additional sources of imbalance beyond those present in centralized optimization. Whether decentralization amplifies or attenuates disparities across sensitive attributes is therefore not theoretically obvious and depends on factors such as client sampling, aggregation mechanisms, and local data heterogeneity.

Although fairness in recommendation has been extensively studied in centralized settings, and recent work has begun to explore fairness-aware methods in federated recommendation Liu et al. (2022); Agrawal et al. (2024), a systematic understanding of how decentralization itself affects bias remains limited. Existing studies typically focus on adapting specific mitigation techniques to federated training, without directly comparing federated and centralized models under aligned architectures and optimization settings. As a result, it remains unclear whether observed differences in bias stem from the learning paradigm, the model design, or the mitigation strategy employed. Furthermore, many fairness interventions are formulated under the assumption that sensitive attributes and global group statistics are centrally accessible Kheya et al. (2025), an assumption that does not hold in cross-device federated environments. Consequently, the relationship between federated optimization and bias dynamics is not yet well characterized, and the transferability of centralized mitigation methods to federated settings remains an open question.

In this work, we focus on disparities associated with sensitive user attributes, specifically gender and user activity. Gender is a widely studied demographic attribute in fairness-aware recommendation Kheya et al. (2025), while user activity—often defined relative to global interaction statistics Xuan et al. (2025); Ji et al. (2022)—captures disparities between highly active and less active users. In cross-device federated settings, these attributes introduce distinct challenges: sensitive information remains local to user devices, and global group statistics cannot be directly computed without additional coordination mechanisms. Consequently, mitigation strategies developed for centralized training cannot be directly applied in decentralized environments. To address this gap, we conduct a systematic comparison of bias in centralized and cross-device federated recommendation under closely aligned model families, including MF/MetaMF Koren et al. (2009); Lin et al. (2020b), NCF/FedNCF He et al. (2017); Perifanis & Efraimidis (2022), VAE/EFVAE Liang et al. (2018); Zhang et al. (2024c), and LightGCN with FedPerGNN He et al. (2020); Wu et al. (2022). Using established bias metrics such as Category Coverage (CC) and Category Discounted Cumulative Gain (CDCG) Kheya et al. (2025), we evaluate whether transitioning to federated optimization alters the magnitude or direction of sensitive-attribute disparities. Furthermore, we adapt an existing in-training bias mitigation strategy to the cross-device setting through privacy-compatible aggregation mechanisms that avoid direct sharing of sensitive attributes. This design enables us to assess not only how bias behaves under federated training, but also how mitigation strategies transfer across centralized and decentralized paradigms.

Our main contributions are summarized as follows:

- A systematic comparison of gender and user activity bias in centralized and cross-device federated recommendation under aligned model families (MF/MetaMF, NCF/FedNCF, VAE/EFVAE, and LightGCN-based variants) and evaluation protocols.
- A reproduction of the results from Kheya et al. (2025) under a cross device federated scenario.
- A privacy-compatible adaptation of the in-training bias mitigation method of Kheya et al. (2025) to the cross-device federated setting.
- An ablation study analyzing how federated optimization parameters and mitigation hyperparameters influence the emergence and evolution of bias.

2 Related work

This section reviews prior work on fairness in AI, bias in recommendation systems, and federated recommendation. We focus on studies most relevant to consumer-side bias and its mitigation in cross-device federated settings.

2.1 Bias and Fairness in AI

Bias in machine learning refers to the unfair inclinations affecting the model's outcomes. This can arise from (i) the data, when it contains historical patterns reflecting existing stereotypes; (ii) from humans, through biases during the data collection, development, or deployment process; and (iii) from the models themselves, when design choices lead to skewed predictions Mehrabi et al. (2021); Kheya et al. (2024). Recent research in the field of fairness in AI Models has focused heavily on the application of fairness definitions introduced by Hardt et al. (2016), such as Equalized Odds Dablain et al. (2022); Yang et al. (2023); Zhang et al. (2018b); Bharti et al. (2023); Yu et al. (2024); Mishler et al. (2021), Equal Opportunity Geyik et al. (2019); Zhang et al. (2018b); Huang et al. (2022), and demographic parity Geyik et al. (2019); Denis et al. (2024); Pereira Barata et al. (2024); Li et al. (2021a); Zhang et al. (2018a); Kheya et al. (2025); Rosenblatt & Witter (2023) (among others), to develop methods to mitigate biases in such models. These mitigation schemes focus on ensuring group-level fairness, for instance, ensuring users of different sensitive attributes get a similar number of positive outcomes. Based on such definitions, practitioners can design bias mitigation strategies that operate at different stages in the AI pipeline: pre-processing, which involoves modifying the data itself to reduce bias Bellamy et al. (2019); Chakraborty et al. (2020); Iosifidis et al. (2019); Rastegarpanah et al. (2019); in-processing, which directly penalizes the model during training to prevent it from learning biased representations Kheya et al. (2025); Yao & Huang (2017); Du et al. (2021a); Qi et al. (2022); and post-processing, which adjusts the output of the model once its done training to correct for bias Fu et al. (2020); Li et al. (2021a); Singh & Joachims (2018); Naghiaei et al. (2022).

2.2 Bias in Recommendation Systems

Bias in recommendation systems can manifest in several ways, and the mitigation strategies employed can be multi-sided. These include: provider-side fairness, which focuses on fair exposure of providers; consumer-side fairness, which aims to provide similar recommendations regardless of the users' sensitive attributes; and CP-fairness, which jointly accounts for fairness for both parties Burke (2017). As our work centers on consumer-side bias and fairness, this section provides an overview of the existing work in this field. Several studies have revealed inequalities in recommendations, with different outcomes for users with different demographic attributes (like gender age, occupation) Kheya et al. (2025); Ferraro et al. (2021); Yang et al. (2020); Boratto et al. (2022); Zhu et al. (2020; 2018); Wei & He (2022) as well as characteristics such as user popularity/activity Ekstrand et al. (2018); Abdollahpouri et al. (2017; 2021a;b); Lin et al. (2020a); Kowald et al. (2020).

2.3 Federated recommendation

Federated recommendation (FR) is a field of study that aim to develop recommendation sytems in FL settings. Since recommendation systems are designed to leverage users' data, like their preferences, personal information and historical interactions Nguyen et al. (2024a). FR Systems can be categorized into two broad paradigms, cross silo (or cross-platform) settings where multiple organizations collaborate to build a shared recommendation system without sharing their data, and Cross-device where each user represents a single device Sun et al. (2024). In cross-device setting, the recommendation system must be able to learn from user interactions across different devices, with only one user's interactions, rendering Collaborative filtering in particular challenging. For this reason, many centralized recommendation algorithms and techniques need to be adapted to FR settings to operate under such challenging scenarios. Such methods include, Matrix Factorisation (MF)Chai et al. (2021); Du et al. (2021b); Lin et al. (2020b), Neural Collaborative filtering (NCF) Perifanis & Efraimidis (2022); Ali et al. (2025), variational autoencoders Zhang et al. (2024c), Two-tower systems Zhang et al. (2024a) or GNNs Wu et al. (2022); Agrawal et al. (2024); Qu et al. (2023). Other

contributions on cross-device FR over the years range from local resource for training and communiction overhead Dogra et al. (2022); Muhammad et al. (2020); Nguyen et al. (2024b), exploring recommendation specific alternatives to FL methods such as client selection and aggregation Zhang et al. (2024b); Muhammad et al. (2020), security investigations Yin et al. (2024); Zhang et al. (2022; 2024d) to encryption and privacy preservation Zhang et al. (2024c); Nguyen et al. (2024b).

Previous work has shown how federated recommenders can exacerbate both user-group biases, including disparities based on activity and sensitive attributes Liu et al. (2022); Li et al. (2025). To mitigate such biases, existing approaches propose fairness-aware optimization objectives Li et al. (2025); Liu et al. (2022); Agrawal et al. (2024); Wang et al. (2025). While these works mitigate bias (following their adopted notion of fairness), they can fall short in (i) conducting a systematic comparison between multiple federated and centralized recommendation systems when evaluating consumer-side fairness; (ii) considering multiple sensitive attributes (like gender and activity); (iii) evaluating and mitigating bias in a granular manner by explicitly incorporating item categories (like genres for movies), into the fairness assessment.

3 Preliminaries and Problem Formulation

Notations: We consider a cross-device FR setting with an honest-but-curious server, where each client corresponds to a single user and holds private interaction data as well as sensitive attributes. These attributes may include demographic information (e.g., gender, age, occupation) or behavioral characteristics (e.g., activity level, popularity). The central server coordinates training but does not have access to raw user data or sensitive attributes.

Formally, let $U = \{u_1, \ldots, u_n\}$ denote the set of n users and $V = \{v_1, \ldots, v_m\}$ the set of m items. User–item interactions are represented by a rating matrix $R \in \mathbb{R}^{n \times m}$, where each entry r_{ij} indicates an interaction (e.g., rating, click, or purchase) between user u_i and item v_j. Each user $u \in U$ has a sensitive attribute $a_u \in A$ known only locally to u, with A a set of values for a sensitive attribute (e.g., $A = \{m, f\}$ for gender, or $A = \{\text{teen, adult, senior}\}$ for age, etc). Each item v is associated with a set of categories $C_v \subseteq C$, with C representing the set of all categories (e.g, Movie categories, Business types, Book Genres, etc.). Finally, let $\text{TopK}_u \subseteq V$ denote the set of top-K items recommended to user u by a given recommendation model.

Bias Metrics: To quantify bias in recommendation outcomes, we employ two exposure-based metrics: Category Coverage (CC) and Category Discounted Cumulative Gain (CDCG) Kheya et al. (2025). CC and CDCG are respectively defined as follows:

$$CC(c, U) = \frac{1}{|U|} \sum_{u \in U} \frac{1}{|\text{TopK}_u|} \sum_{v \in \text{TopK}_u} \frac{\mathbb{1}_{\{\{c\} \cap C_v \neq \emptyset\}}}{|C_v|} \tag{1}$$

$$CDCG(c, U) = \frac{1}{|U|} \sum_{u \in U} \frac{1}{|\text{TopK}_u|} \sum_{v \in \text{TopK}_u} \frac{\mathbb{1}_{\{\{c\} \cap C_v \neq \emptyset\}}}{|C_v| \pi} \tag{2}$$

where $\mathbb{1}$ represents the binary indicator, and π represents the rank of item v in the TopK_u.

These metrics measure the relative exposure allocated to item categories in recommendation lists and allow us to quantify disparities across sensitive user groups.

Problem Formulation: Our objective is to investigate how transitioning from centralized recommendation to cross-device federated recommendation affects bias in top-K recommendations, particularly gender bias and user activity bias. Let M denote a recommendation model, instantiated either as a centralized model M_{cent} or a federated model M_{fed}. Let $B(M)$ represent a bias metric computed using CC or CDCG. We define the bias difference between learning paradigms as:

$$\Delta B = B(M_{\text{fed}}) - B(M_{\text{cent}}). \tag{3}$$

Our goal is to determine whether federated optimization amplifies or attenuates sensitive-attribute disparities relative to centralized training, and to identify the conditions under which this occurs (e.g., model architecture, aggregation method, client sampling). Furthermore, we adapt an in-training group-level fairness regularization method from centralized settings to the federated context, ensuring that sensitive attributes remain local to clients. We evaluate the effectiveness of this mitigation strategy and compare bias levels across centralized and federated models before and after mitigation. Our key research questions are:

1. How does bias in federated recommendation compare to centralized recommendation?
2. What factors influence the bias difference ΔB?
3. How effective are bias mitigation strategies in the cross-device federated setting?

4 Theoretical intuitions of Bias in FR

This section analyzes how the transition from centralized to federated training can alter bias dynamics in recommender systems. Rather than providing a formal proof, we outline key mechanisms through which federated optimization may either amplify or attenuate existing disparities. The goal is to provide an interpretable reasoning framework that supports the empirical observations and informs the design of mitigation strategies.

Effect of Client Heterogeneity and aggregation on Bias: A fundamental difference between centralized and federated training is that, in FL, the global update results from aggregating gradients computed on locally skewed client data. In cross-device recommendation, where each client corresponds to a single user, this skewness reflects user-level demographic or behavioral characteristics. For demographic attributes (e.g., gender or age), each client update represents one individual. The global update thus becomes a mixture of group-specific gradients. If participation probabilities differ across groups, the aggregated model may be influenced disproportionately by overrepresented users. For instance, if a fraction $frac$ of sampled clients belongs to one group and $1 - frac$ to another, deviations of α from the true population proportion can shift optimization toward the dominant group's interaction patterns. However, because aggregation occurs at the user level and across randomly sampled subsets, the resulting global update can also dilute persistent majority dominance observed in centralized training, particularly when centralized optimization overfits to globally prevalent patterns. Behavioral attributes, in contrast, are defined by interaction patterns rather than inherent characteristics. In cross-device FL, heterogeneity appears in the size of local datasets. Under aggregation schemes like FedAvg, where updates are weighted by local data size, highly active users can exert greater influence. Yet centralized training similarly over-represents active users through interaction-level sampling. In practice, user-level aggregation may partially decouple interaction frequency from update frequency, potentially attenuating activity-driven disparities.

Effect of User Expansion on Behavioral Bias: In federated recommendation frameworks that incorporate user expansion (e.g., FedPerGNN), clients augment their local data with information from other users who share at least one interaction. While this increases effective sample support, it can structurally favor already active users. Highly active users are more likely to appear in multiple expansion sets, causing their interactions to be repeatedly propagated across clients. When such users are selected in a training round, their updates reflect not only their own interactions but also the aggregated influence of others' expansions, amplifying their contribution to the global model. This mechanism differs from item-side popularity bias: the amplification occurs at the user level, increasing the relative influence of already active users rather than merely reinforcing popular items. As a result, user activity bias may intensify even if other disparities diminish. Consistent with this intuition, our experiments suggest that while user expansion can reduce certain demographic disparities (e.g., gender bias), it may simultaneously exacerbate activity-based behavioral bias.

Effect of Performance Degradation on Observed Bias: In cross-device federated recommendation settings, models often exhibit lower overall ranking performance compared to centralized training due to

limited local data, partial client participation, and constrained communication. This performance degradation can lead to a superficial reduction in measured bias: the predicted scores across items become flatter, and the Top-K recommendations for each user may include a wider variety of items, including less popular ones. While this can reduce metrics such as CC or CDCG, it is important to note that this reduction does not reflect genuine bias mitigation, but rather a side effect of diminished discrimination power in the model. Empirical studies in cross-device federated recommendation have observed that lower-quality models may exhibit lower concentration on popular categories, yet at the cost of reduced recommendation relevance.

5 Privacy aware mitigation of bias

In this section, we explore and adapt to the FR setting, a mitigation strategy based on fairness loss regularization introduced in Kheya et al. (2025). Following that work, we use the Category Coverage metric defined in Equation 1. However, other metrics could also be used to compute the loss, as our approach is, in principle, generalizable to other training-loss-based regularization methods. In order to compute Equation 1 and the fairness regularization term Kheya et al. (2025) on the server, we require estimates of (i) the category exposure scores aggregated over users belonging to each sensitive attribute value, and (ii) the number of users in each attribute group. We denote these aggregated quantities by $\mathbf{S}$ and $\mathbf{q}$, respectively.

First, we compute a category exposure measure of user u_i for category c_z denoted $\mathbf{x}_u$, which represents the part of CC in Equation 1 that can be computed locally, and is defined as:

$$x_{u,c} = \frac{1}{|\mathrm{TopK}_u|} \sum_{v \in \mathrm{TopK}_u} \frac{\mathbb{1}_{\{\{c\} \cap C_v \neq \emptyset\}}}{|C_v|}. \tag{4}$$

This quantity measures the fraction of the recommendation list allocated to category c_z, normalized by the number of categories associated with each item. The resulting vector:

$$\mathbf{x}_u = (x_{u,c_1}, \ldots, x_{u,c_{|C|}}) \in \mathbb{R}^{|C|} \tag{5}$$

represents the category exposure profile of user u_i. To enable group-level fairness estimation without revealing sensitive attributes, each user u constructs the necessary statistics for each attribute $a \in A$ locally, namely, $n_{u,a}$, the local count of sensitive attribute a (aka, 1 or 0 in a cross device FR) and $\mathbf{s}_{u,a}$ the category exposures across all categories for user u with respect to the attribute a, such as:

$$\mathbf{s}_{u,a} = \begin{cases} \mathbf{x}_u & \text{if } a_u = a, \\ \mathbf{0}_{|\mathbf{C}|} & \text{otherwise,} \end{cases} \quad q_{u,a} = \begin{cases} 1 & \text{if } a_u = a, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Thus, each client produces for all $a \in A$ a matrix for category exposures $\mathbf{S}_u \in \mathbb{R}^{|A| \times |C|}$ and a one hot encoding for the attribute count $\mathbf{q}_u \in \mathbb{R}^{|A|}$, but never transmits the attribute a_u or the categories exposures scores for u explicitly.

We mainly rely on secure aggregation to transmit and protect these local statistics. Secure aggregation is a cryptographic mechanism that guarantees that the server can only observe aggregated sums of client inputs and learns no information about any individual client's values beyond what is implied by these sums Bonawitz et al. (2017). Thus, the server receives the global statistics, such as:

$$\mathbf{S} = \sum_{u \in U} \mathbf{S}_u, \quad \mathbf{q} = \sum_{u \in U} \mathbf{q}_u, \tag{7}$$

without having access to the individual local statistics $\mathbf{S}_u$ and $\mathbf{q}_u$[1].

Operationally, there exist multiple secure aggregation protocols that can achieve this outcome Zhang et al. (2025). One example is pairwise masking Bonawitz et al. (2017) where each client masks its value with

[1]In practice, we collect $\mathbf{q}$ only once at the beginning of the training, since the statistics are unlikely to change during the process. This can decrease communication costs and the risks of revealing the information through adaptive subsampling attacks.

random noise that cancels out across clients. Specifically, let's consider $\mathbf{q}_u$ stored locally for each user u (will be the same process for $\mathbf{S}_u$), in order to transmit a global sum $\mathbf{q}$ across all clients, each user u constructs a mask:

$$\mathbf{m}_{u_i} = \mathbf{q}_{u_i} + \sum_{u_j \in U, u_j \neq u_i} \mathbf{r}_{u_i,u_j},$$

where $\mathbf{r}_{u_i,u_j}$ are pairwise random masks satisfying $\mathbf{r}_{u_i,u_j} = -\mathbf{r}_{u_i,u_j}$. When the server aggregates all messages:

$$\sum_{u \in U} \mathbf{m}_u = \sum_{u \in U} \mathbf{q}_u,$$

Since all random masks cancel out, the server can recover the sum of client inputs $\mathbf{q}$ but cannot infer any individual contribution[2].

Now that the server has access to the global aggregated statistics, it computes the group-wise CC for every group and category:

$$\widehat{CC}_{a,c} = \frac{\mathbf{S}_{a,c}}{\mathbf{q}_a}, \tag{8}$$

where $\mathbf{S}_{a,c}$ denotes the aggregated exposure of category c for the group of users with sensitive attribute value a. The server then computes the global fairness signal $\mathcal{L}_{\text{fair}}$ as a sum of the global group disparity across all categories, for all pairs of sensitive attributes $(a_i, a_j) \in A$:

$$\mathcal{L}_{\text{fair}} = \sum_{\substack{a_i,a_j \in A \\ a_i \neq a_j}} \sum_{c \in C} |\widehat{CC}_{a_i,c} - \widehat{CC}_{a_j,c}|, \tag{9}$$

The server then broadcasts $\mathcal{L}_{\text{fair}}$ (or a smooth transformation thereof) to all participating clients. Each client incorporates the broadcast fairness signal into its local training objective. The overall local objective becomes:

$$\mathcal{L}_u = (1-\alpha)\mathcal{L}_{\text{rec},u} + \alpha\mathcal{L}_{\text{fair}}, \tag{10}$$

where $L_{\text{rec},u}$ is the standard recommendation loss like MSELoss or BPRLoss and $\alpha \in [0,1]$ controls the trade-off between accuracy and fairness. We must note that if each client performs multiple local optimization steps, which is often the case in FL architectures. $\mathcal{L}_{\text{fair}}$ would be reused for multiple epochs, however our ablation study in Section 6.4.3 shows that this has little impact on the bias reduction and performance of the system.

Potential limits of secure aggregation protections Secure aggregation can be a sufficient privacy protection layer under multiple threat models. However, while it protects individual client inputs, the server has access to the exact aggregated outputs. In our general case, having access to exact aggregated outputs (exact global statistics) can be viewed as desirable for accurately computing the fairness signal. However, under more adversarial assumptions, the availability of exact aggregates may still enable inference about individual clients or sensitive groups. Here is an overview of some assumptions that may impact the privacy of the system:

- Differencing Attacks: In real-world scenarios, participation is typically dynamic, clients may join or leave between rounds due to connectivity or availability. From consecutive rounds of group-level statistics $(\mathbf{S}^{(t)}, \mathbf{q}^{(t)})$, the server may infer some information about users. For example, in the extreme case where a single client from a sensitive group drops out between two rounds, the difference between the aggregates may approximate that client's contribution. This risk increases when participation sets are small or variable.

[2]Other considerations for modern secure aggregation protocols incorporate mechanisms for handling client dropouts and communication failures Zhang et al. (2025). However, this falls beyond the scope of this paper; we can theoretically employ any secure aggregation method for transmitting sensitive attributes.

- Malicious Server and Adaptive Subsampling: if we assume a malicious server, instead of our honest-but-curious assumption, the server may attempt to isolate subsets of clients across multiple rounds to infer their attributes. More advanced secure aggregation variants can mitigate this risk Bell et al. (2023); So et al. (2023) and ensure that participation sets cannot be manipulated.
- Rare or Imbalanced Sensitive Attributes: When certain sensitive attributes correspond to small or rare groups, exact aggregated counts $\mathbf{q}$ may themselves reveal sensitive information. If only a few clients belong to a particular group $a \in A$ in a given round, the server may infer participation or demographic structure with high confidence. In the extreme case where $\mathbf{q} = 1$, the aggregate exposure vectors directly correspond to a single user's statistic, even though the server does not observe it explicitly. This issue becomes more pronounced when group sizes are highly imbalanced or fluctuate significantly across rounds. Even without explicit dropout attacks, this increases the risk of attribute inference.

Clipping and Statistical Noise While stronger secure aggregation variants can reduce certain attack surfaces, they can be costly in terms of communication costs and may be insufficient in certain scenarios. As such, adding statistical noise to the local statistics can provide an additional layer of protection at the output level. By perturbing $\mathbf{S}$ and $\mathbf{q}$ before or during aggregation, the system prevents exact reconstruction of individual contributions through temporal differencing and reduces the precision of demographic inference in small or rare groups. As such, we investigate adding differential privacy-style noise as an optional layer of security that can be applied under specific scenarios. To bound the influence of any single client, each vector $\mathbf{s}_{u,a}$ is clipped to a fixed norm:

$$\mathbf{s}_{u,a} = \frac{\mathbf{s}_{u,a}}{\max(1, \|\mathbf{s}_{u,a}\|_2 / S)}, \tag{11}$$

where S is a predefined clipping threshold. Then, each client perturbs its statistics using additive noise:

$$\tilde{\mathbf{s}}_{u,a} = \mathbf{s}_{u,a} + \mathcal{N}(0, \sigma_s^2 I), \quad \tilde{q}_{u,a} = q_{u,a} + \mathcal{N}(0, \sigma_n^2), \tag{12}$$

where σ_s and σ_n control the privacy-utility trade-off.

Handling Globally defined group attributes Contrary to demographic attributes such as gender or age, which are directly known to the user and stored locally, some fairness attributes are defined with respect to the global data distribution. A typical example is user activity, which is often defined as belonging to the top $n\%$ of users with the highest number of interactions Li et al. (2021b); Xuan et al. (2025), using a median split Ji et al. (2022), or other dataset-specific heuristics. In a cross-device federated setting, each client can compute its own number of interactions p_u, but does not know its relative rank within the global population. Therefore, before applying group-based fairness regularization, the system must first estimate the global distribution of interaction counts in a privacy-preserving manner. Although activity and p_u are not always protected attributes, in cross-device settings, they often constitute private behavioral data. Interaction counts are potentially identifying, can be stable across rounds, and may correlate with demographic attributes (e.g., age, socio-economic status). To address this, we can use a secure aggregation federated analytics process to privately estimate a global popularity threshold t corresponding to the top $n\%$ of users before training. Instead of sharing exact interaction counts, clients contribute only aggregated statistics through secure aggregation, allowing the server to approximate the $(1-n)$-quantile of the interaction distribution without accessing individual p_u values. Intuitively, this can be achieved by first obtaining a coarse estimate of the global distribution and then refining the threshold within a reduced interval using a small number of additional aggregated queries. Once the global threshold t is determined, it is broadcast to clients, which locally assign themselves to the active or non-active group based on whether $p_u \geq t$.

6 Experiments

In this section, we empirically evaluate the impact of FL on bias and fairness in FR. Furthermore, we test the mitigation strategy adapted from Kheya et al. (2025) in Section 5. Finally, we conduct an in depth ablation study to determine the impact on fairness of various elements.

Table 1: Datasets' statistics.

Dataset	#Users	#Items	#Interactions	Sparsity	#Item Categories
MovieLens-100k (ML-100k) Harper & Konstan (2015)	943	1,682	100,000	93.69%	18
MovieLens-1M (ML-1m) Harper & Konstan (2015)	6,040	3,706	1,000,209	95.53%	18
Yelp Mansoury et al. (2019)	1,316	1,272	97,991	99.72%	21

6.1 Datasets and experimental setup

Datasets: We conduct experiments on three widely-used recommendation system datasets, namely MovieLens-100k (ML-100k), MovieLens-1M (ML-1m) Harper & Konstan (2015), and Yelp Mansoury et al. (2019) (see Table 1 for dataset statistics). To ensure a fair comparison between centralized and cross-device federated settings, we apply the same data processing pipeline across all experiments. Specifically, we filter users and items using a k-core filtering with k=5 interactions, to reduce extreme sparsity and cold-start artifacts, and split each dataset into train/validation/test sets (with the ratios 70/10/20) using a temporal stratified split Meng et al. (2020) for MovieLens and a random stratified split for Yelp. Furthermore, for models requiring implicit feedback (eg, LightGCN), we filter interactions with a rating lower than 2.5 and binarize them (an item is considered interacted with if the user has a positive interaction), and construct the candidate set for ranking evaluation using standard negative sampling.

Baselines: We evaluate the following models: Matrix Factorization (MF) Koren et al. (2009), Neural Collaborative Filtering (NCF) He et al. (2017), LightGCN He et al. (2020), and a variational autoencoder for collaborative filtering (VAE) Liang et al. (2018). In the federated setting, we compare representative federated baselines and federated variants of these models under a cross-device simulation where each client corresponds to a single user. In particular, we include MetaMF Lin et al. (2020b) and FedNCF Perifanis & Efraimidis (2022) as alternatives to MF and NCF, respectively, FedPerGNN Wu et al. (2022) instantiated with a LightGCN local model[3], and EFVAE Liang et al. (2018) as a federated VAE-style method. For all federated methods, training proceeds over communication rounds with a central server coordinating model aggregation; clients locally optimize their objective on-device using their private interactions[4] and then send updates to the server.

Evaluation metrics: We evaluate both recommendation quality and bias outcomes. Recommendation accuracy is reported using nDCG@50 and HR@50. For evaluating fairness, we measure two forms of bias, *gender bias*, where users are grouped according to the available gender attribute, and *user popularity bias*, where users are grouped according to their activity levels (active vs non-active users). The disparities are computed across these groups using CC@50 (See Equation 1) and CDCG@50 (See Equation 2) computed from the ranked lists of recommended items across categories[5] produced by each model.

6.2 Comparing bias and mitigation

We first compare centralized training against federated training to characterize how decentralization affects both recommendation utility and bias. For each backbone (MF, NCF, LightGCN, and VAEF), we report results for: (i) centralized training without mitigation, (ii) federated training without mitigation (FedMF, FedNCF, FedPerGNN (LightGCN), and EFVAE), and (iii) the corresponding centralized and federated variants augmented with bias mitigation (separately for gender and popularity). This design allows us to isolate whether changes in bias are driven primarily by the learning paradigm (centralized vs. federated), the model family (shallow vs. neural vs. graph-based vs. generative), or by the mitigation procedure itself.

[3]FedPerGNN was originally proposed with explicit feedback models using GCNs or GATs, for the purposes of this comparison, we used LightGCN as a local model to enable the comparison with the centralized variant. In order to do so we enabled the learning and aggregation of expanded embeddings within FedPerGNN. Because since LightGCN doesn't have any GNN parameter beside the user and item embedding, the model didn't converge well with the aggregation of item embeddings alone.

[4]FedperGNN also includes a user expansion phase where local interactions are privately expanded by relying on a third-party trusted server

[5]The Categories are movie categories for MovieLens and Store categories for Yelp

Table 2: Comparison of performance (NDCG@50 and HR@50) and Gender/User Activity bias (CC@50 and CDCG@50) between Centralized and Decentralized settings and mitigation strategy for 4 families of models

			nDCG@50↑	HR@50↑	CC@50↓ Gend	CDCG@50↓ Gend	CC@50↓ Activity	CDCG@50↓ Activity
ML-100k								
MF	No mitigation	Cent	0.1947	0.9226	0.0476	0.0127	0.2004	0.0444
		Fed	0.0644	0.5779	0.0157	0.0057	0.1404	0.0415
	Gender mitigation	Cent	0.1760	0.9067	0.0565	0.0151	0.1982	0.0512
		Fed	0.0659	0.5885	0.0205	0.0065	0.1637	0.0444
	Activity mitigation	Cent	0.1760	0.9067	0.0565	0.0151	0.1982	0.0512
		Fed	0.0682	0.5960	0.0172	0.0054	0.1661	0.0454
NCF	No mitigation	Cent	0.2351	0.9438	0.0974	0.0303	0.1003	0.0270
		Fed	0.1468	0.7620	0.0116	0.0024	0.0171	0.0054
	Gender mitigation	Cent	0.1978	0.9183	0.0467	0.0110	0.1680	0.0531
		Fed	0.0741	0.5848	0.0097	0.0027	0.0173	0.0053
	Activity mitigation	Cent	0.1575	0.8908	0.0420	0.0097	0.1556	0.0340
		Fed	0.0741	0.5848	0.0097	0.0027	0.0173	0.0053
LightGCN	No mitigation	Cent	0.2421	0.9502	0.0896	0.0268	0.1257	0.0307
		Fed	0.2240	0.9130	0.0120	0.0026	0.1752	0.0322
	Gender mitigation	Cent	0.2148	0.9152	0.0800	0.0233	0.0884	0.0274
		Fed	0.2200	0.9120	0.0055	0.0016	0.1875	0.0323
	Activity mitigation	Cent	0.2298	0.9384	0.0816	0.0246	0.1360	0.0195
		Fed	0.2203	0.9173	0.0092	0.0020	0.1803	0.0298
VAEF	No mitigation	Cent	0.1501	0.8918	0.0403	0.0106	0.1798	0.0446
		Fed	0.2286	0.9152	0.0166	0.0060	0.1252	0.0330
	Gender mitigation	Cent	0.1503	0.8887	0.0484	0.0112	0.1974	0.0522
		Fed	0.2328	0.9279	0.0187	0.0059	0.1131	0.0275
	Activity mitigation	Cent	0.1496	0.8865	0.0474	0.0111	0.1844	0.0474
		Fed	0.2378	0.9247	0.0248	0.0065	0.1091	0.0304
ML-1m								
MF	No mitigation	Cent	0.1408	0.8387	0.1283	0.0372	0.1383	0.0396
		Fed	0.0759	0.6266	0.0077	0.0028	0.1200	0.0399
	Gender mitigation	Cent	0.1042	0.7803	0.0676	0.0184	0.0922	0.0242
		Fed	0.0619	0.5855	0.0044	0.0013	0.1692	0.0407
	Activity mitigation	Cent	0.1269	0.8177	0.0461	0.0131	0.1593	0.0415
		Fed	0.0621	0.5839	0.0042	0.0013	0.1735	0.0418
NCF	No mitigation	Cent	0.1467	0.9169	0.2460	0.0671	0.0828	0.0226
		Fed	0.1172	0.7667	0.0043	0.0012	0.0079	0.0020
	Gender mitigation	Cent	0.1211	0.7925	0.0597	0.0140	0.1425	0.0366
		Fed	0.1127	0.7601	0.0024	0.0008	0.0134	0.0025
	Activity mitigation	Cent	0.1643	0.9020	0.3177	0.0879	0.1262	0.0344
		Fed	0.1147	0.7700	0.0045	0.0011	0.0121	0.0030
LightGCN	No mitigation	Cent	0.1670	0.9033	0.3127	0.0850	0.1257	0.0338
		Fed	0.1831	0.8972	0.0295	0.0066	0.1849	0.0461
	Gender mitigation	Cent	0.1396	0.8583	0.2623	0.0708	0.1615	0.0439
		Fed	0.1825	0.8957	0.0263	0.0060	0.1842	0.0414
	Activity mitigation	Cent	0.1403	0.8647	0.2684	0.0720	0.1582	0.0418
		Fed	0.1811	0.8945	0.0283	0.0069	0.1763	0.0437
VAEF	No mitigation	Cent	0.1235	0.8002	0.0837	0.0220	0.1571	0.0425
		Fed	0.1686	0.9148	0.0084	0.0020	0.1524	0.0409
	Gender mitigation	Cent	0.1236	0.7957	0.0889	0.0224	0.2396	0.0597
		Fed	0.1625	0.9146	0.0084	0.0026	0.1200	0.0328
	Activity mitigation	Cent	0.1237	0.8025	0.0811	0.0215	0.1728	0.0405
		Fed	0.1659	0.9071	0.0108	0.0024	0.1243	0.0342
Yelp								
MF	No mitigation	Cent	0.1134	0.8457	0.0365	0.0078	0.0388	0.0152
		Fed	0.0340	0.4962	0.0130	0.0036	0.0209	0.0086
	Gender mitigation	Cent	0.0680	0.6983	0.0071	0.0018	0.0474	0.0069
		Fed	0.0313	0.4848	0.0072	0.0021	0.0215	0.0073
	Activity mitigation	Cent	0.0913	0.7903	0.0264	0.0073	0.0787	0.0207
		Fed	0.0348	0.4992	0.0135	0.0036	0.0173	0.0067
NCF	No mitigation	Cent	0.1687	0.9301	0.0409	0.0112	0.0801	0.0204
		Fed	0.0864	0.7623	0.0035	0.0010	0.0065	0.0018
	Gender mitigation	Cent	0.0875	0.7788	0.0087	0.0021	0.0512	0.0129
		Fed	0.0832	0.7608	0.0021	0.0007	0.0043	0.0013
	Activity mitigation	Cent	0.0875	0.7789	0.0087	0.0021	0.0513	0.0130
		Fed	0.0869	0.7585	0.0026	0.0007	0.0128	0.0024
LightGCN	No mitigation	Cent	0.1933	0.9635	0.0495	0.0135	0.0604	0.0142
		Fed	0.0904	0.7720	0.0040	0.0011	0.0680	0.0103
	Gender mitigation	Cent	0.1816	0.9521	0.0452	0.0122	0.0693	0.0151
		Fed	0.0910	0.7705	0.0030	0.0008	0.0594	0.0105
	Activity mitigation	Cent	0.1877	0.9552	0.0461	0.0124	0.0676	0.0148
		Fed	0.0904	0.7758	0.0026	0.0008	0.0432	0.0098
VAEF	No mitigation	Cent	0.0917	0.7888	0.0097	0.0019	0.1011	0.0336
		Fed	0.1612	0.9159	0.0136	0.0044	0.0318	0.0102
	Gender mitigation	Cent	0.0912	0.7774	0.0056	0.0012	0.0883	0.0277
		Fed	0.1736	0.9339	0.0206	0.0053	0.0312	0.0092
	Activity mitigation	Cent	0.0888	0.7644	0.0052	0.0016	0.0588	0.0228
		Fed	0.1729	0.9415	0.0203	0.0053	0.0337	0.0094

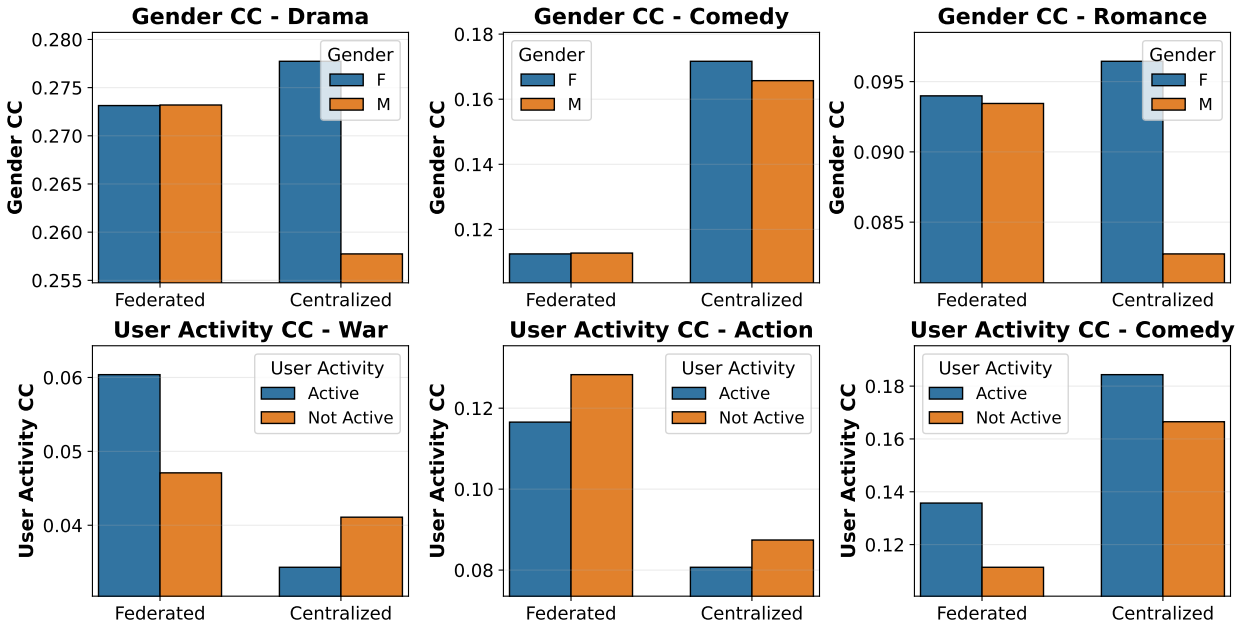


Figure 1: Comparison of CC values for sensitive attributes between Centralized (LightGCN) and Federated (FedPerGNN with LightGCN) settings across 3 categories on ML-100k

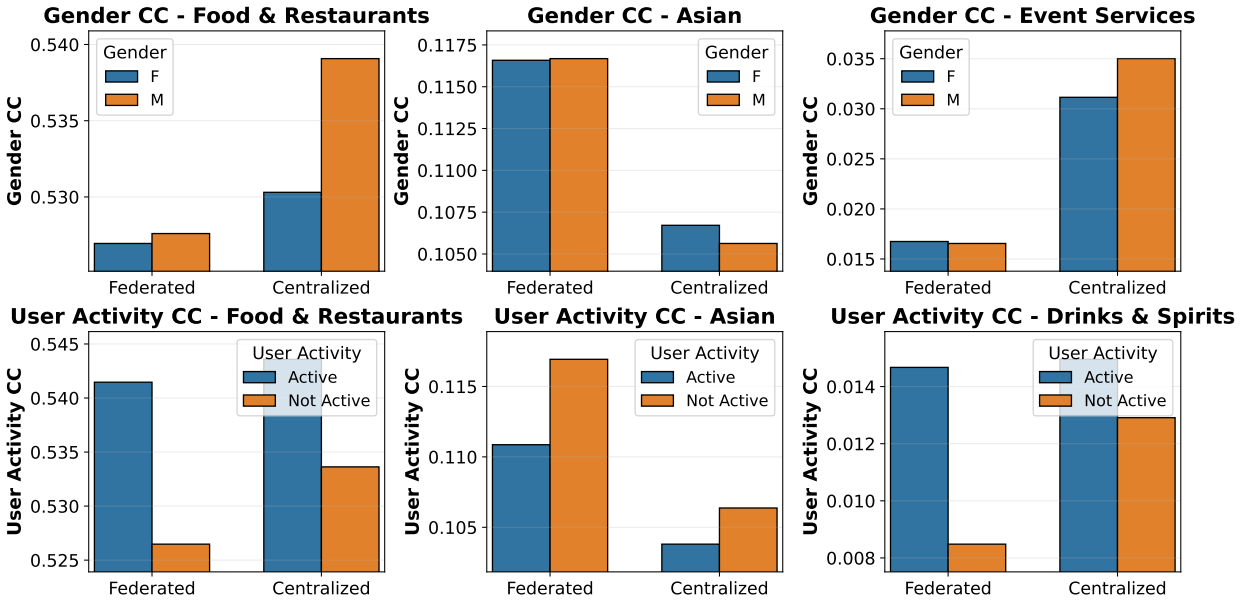


Figure 2: Comparison of CC values for sensitive attributes between Centralized (LightGCN) and Federated (FedPerGNN with LightGCN) settings across 3 categories on Yelp

For mitigation, we evaluate two target objectives: reducing gender-based exposure disparity and reducing popularity-based exposure disparity. In the centralized setting, mitigation is applied in the form of a loss regularization term Kheya et al. (2025). We then adapt the same principle to the federated setting with a privacy-aware design presented in Section 5. For privacy preservation, we only rely on secure aggregation, the impact of added noise for differential privacy would be explored in the ablation study. We report results for three conditions per model family: *No mitigation*, *Gender mitigation*, and *Popularity mitigation*, each evaluated for both centralized (Cent) and federated (Fed) training.

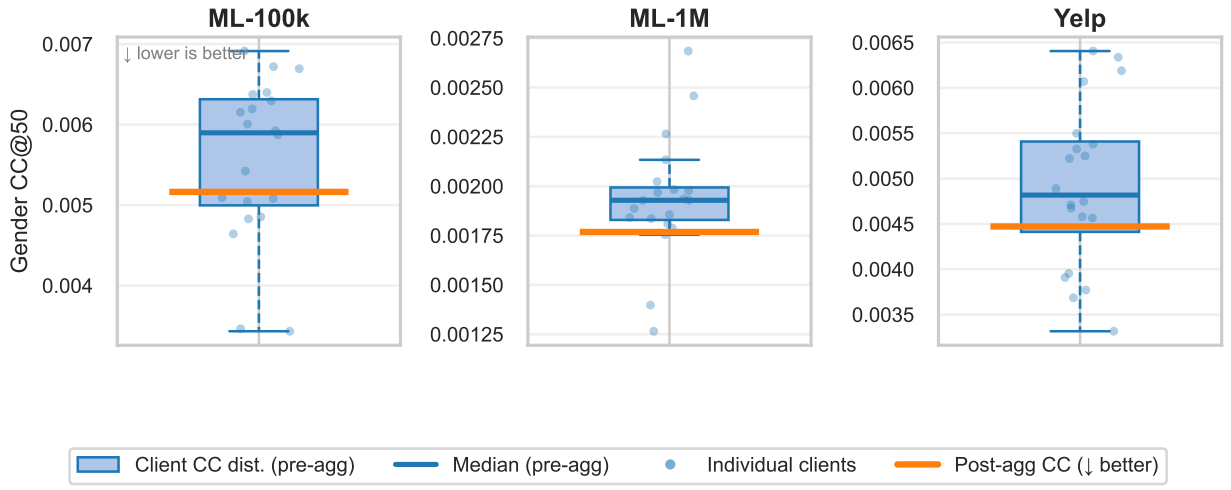


Figure 3: Distribution of gender bias (CC@50) across clients before aggregation (blue box-and-whisker plot, with jittered dots representing individual clients) and the resulting global CC@50 after aggregation (orange stripe) for FedPerGNN. The consistent position of the orange stripe below the client-level average demonstrates that FL aggregation typically reduces gender bias beyond what individual clients achieve on average. Also, we can note that the local models themselves exhibit lower amount of bias than their centralized counterpats.

The results are shown in Table 2. Overall, we observe mixed results, when moving from centralized to federated settings. On one hand, we can see a clear and consistent reduction of gender bias across all datasets and models. This result is surprising, considering the general intuition that FL exacerbates bias Liu et al. (2022); Li et al. (2025). On the other hand, the effect on user activity bias is more mixed; outcomes depend on specific models and datasets. Specifically, we can see that there is a sharp increase in activity bias on LightGCN, which was expected due to the effects of the expansion algorithm as hypothesized previously. To better visualize this change in bias dynamics and how they emerge across categories, we show the CC scores for 3 different categories for gender and activity bias for LightGCN on ML-100k and Yelp on both centralized and FL settings shown in Figure 2. We can see clearly the decrease of gender bias when moving to FL settings, to the point where there is hardly any bias at all. But we can also see an increase in activity bias that matches our observations from Table 2. Another result we can see from the table is that the effect of the mitigation strategy is considerably smaller in FL than in centralized settings, with some instances of bias barely decreasing. This might be due to the low bias present in the model in the first place, especially with respect to gender bias. Alternatively, the effects of the aggregation might cancel the changes from the fairness regularization. However, on the effects of the mitigation on the nDCG measure, we can see little to no negative effect, with the exception of NCF on ml-100k. In some instances, applying the mitigation seems to have even improved the performance of the system, for instance, for VAEF on Yelp and ML-100k and MF on ML-100k.

6.3 Investigations on the bias results in Federated systems

The results in Table 2 and Figure 2 show a consistent and somewhat unexpected reduction in gender bias (and in some case activity bias) when moving from centralized to FL settings. To better understand the mechanisms driving this phenomenon, we design a set of controlled experiments to evaluate our two main hypothesis:

- The bias reduction arise because the subset of clients participating in each round creates a particular exposure distribution.
- The averaging of heterogeneous local updates during aggregation suppresses bias signals.

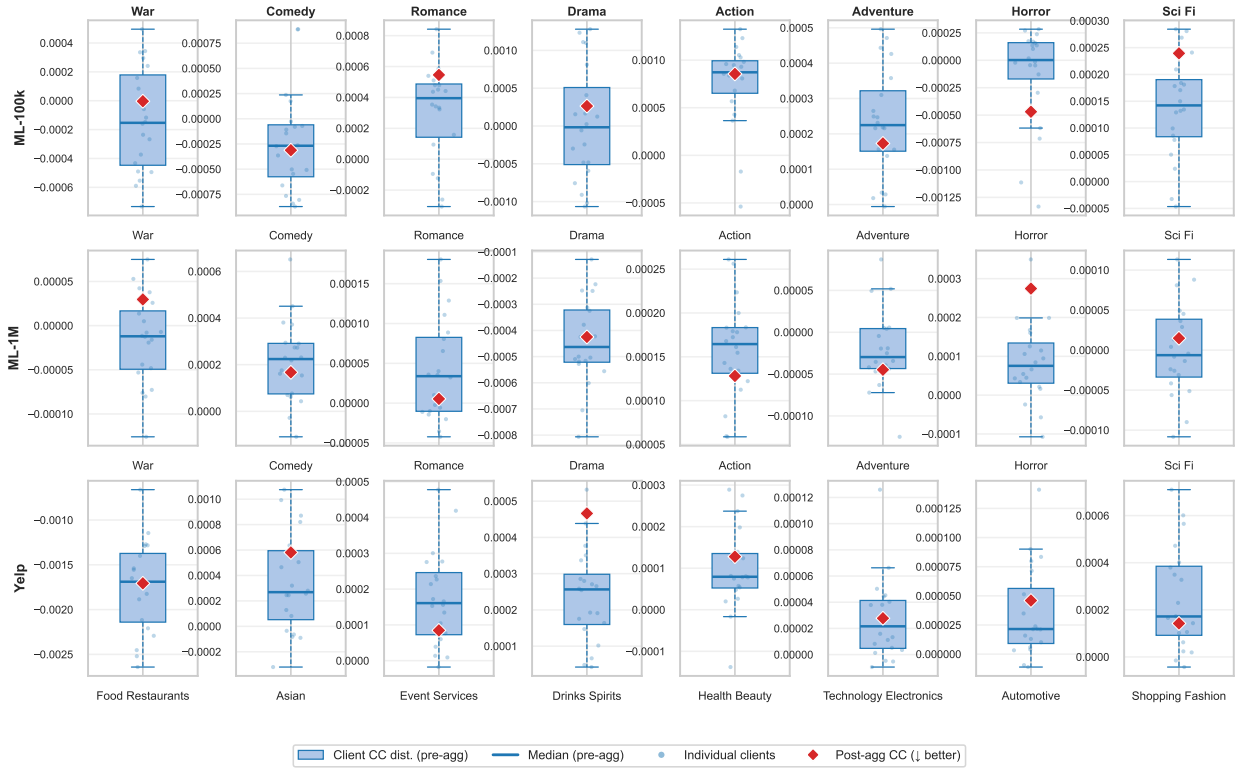


Figure 4: Per-category gender bias (CC@50) for eight selected categories on FedPerGNN, shown in a grid (rows = datasets, columns = category). For each category, the blue box-and-whisker plot with jittered dots depicts the distribution of per-client CC@50 values before aggregation, and the red diamond marks the post-aggregation CC@50 on the global test set.

- Smaller amounts of data during the local training in each client naturally prevent the learning of bias signals.

To answer this, we include in the FL setup a round-level replay system that captures model states and fairness metrics for each selected client before aggregation, then after the aggregation. This lets us attribute changes in bias to either the local training phase (capturing selection effects) or the aggregation phase separately.

To isolate the contribution of model aggregation to the observed bias reduction, we examine in Figures 3 how fairness metrics evolve between the post-local and post-aggregation checkpoints within a single FL round. If aggregation itself partially or fully suppresses bias, we would expect the global model produced by FedAvg to exhibit lower disparity than the average of the participating clients' local models, even when the local models have individually increased their bias during local training. On the other hand Figure 4 shows a similar comparison with the distribution of CC scores per categories across eight different categories in movieLens and Yelp. Our results confirm that aggregation consistently contributes to bias reduction. Across models and datasets, the post-aggregation global model exhibits lower gender bias than the mean of the participating clients' post-local models, suggesting that the parameter-space averaging inherent to FedAvg acts as an implicit regularizer on the exposure disparity captured by our fairness metrics. However, we should note that even the clients with the highest CC score in Figure 3 have a lower scores than centralized alternatives. As such, while aggregation consistently help supressing bias, our results seem to indicate that the local training in each client naturally capture less bias signals.

To further understand the impact of selected users in the cancelation of bias we introduce our next set of experiments which involves directly controling client participation in a given round by fixing the demographic composition of the selected client pool, using either:

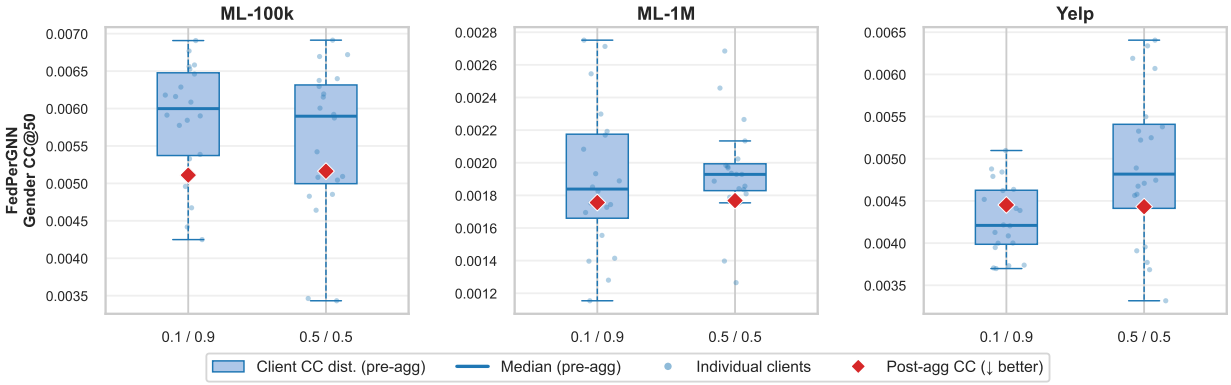


Figure 5: Gender bias (CC@50) before aggregation (blue dashed line) and after aggregation (orange solid line) as the ratio of male to female clients selected per round is skewed in the left plot to 10% females and 90% males. The results indicate that bias signals are not primarily cancelled through opposite signals from different groups

- Group categories selection: selecting users based on group membership rations (For example selecting only male users for a round of training).
- Stereotypical users selection: selecting users based on the stereotypicality of users within fairness groups.

Cancellation of Bias with Group categories selection

A natural hypothesis for why aggregation reduces bias is that bias signals that emerge during the local training across clients belonging to different demographic groups produce local model updates with opposing bias signatures, and that these opposing signals cancel in expectation when averaged. This hypothesis would also explain the smaller reduction in bias on user activity groups compared to gender. We test this by sampling users based on their respective gender using a ratio of 90% male and 10% female, and compare it to the standard 50% male and 50% female shown in Figure 5. If the bias signals get cancelled across different demographic groups, then we would expect to see no reduction or even an amplification of bias signals with the skewed demograpic ratio. However, our results show that the demographic composition of the selected client pool has only a small effect on bias reduction. While the skewed distribution on Yelp does increase CC score, this wasn't generalizable on MovieLense, where it seems that the model trained on skewed gender distribution is less biased, this might be due to the model not learning bias signals from the opposite group, reducing the CC score.

Cancellation of Bias with Stereotypical users selection

Gender and demographic attributes, however, are not the only axis along which client heterogeneity may drive bias cancellation. An alternative source of heterogeneity can be the degree to which a user's interaction history can be considered stereotypical, or in other words, how strongly a user's interaction patterns reflect the patterns associated with their given demographic group. For example, a male user interacting mostly with Romance movies, might be considered non-stereotypical with respect to the existing patterns in MovieLens-100k. The main intuition is that a user with a highly stereotypical interaction history will produce a local model update that reinforces group-level exposure disparities, whereas a user with a non-stereotypical history will produce updates that cut against those disparities. If aggregation averages both kinds of updates, the resulting global model may again exhibit reduced net bias even in the absence of an explicit fairness objective. To test this, we compute a stereotypicality score from which we determine stereotypical and non-stereotypical users in respect to each fairness group, then we sample users based on their membership in the stereotypical or non-stereotypical groups using ratios similarly to the Group categories selection experiments.

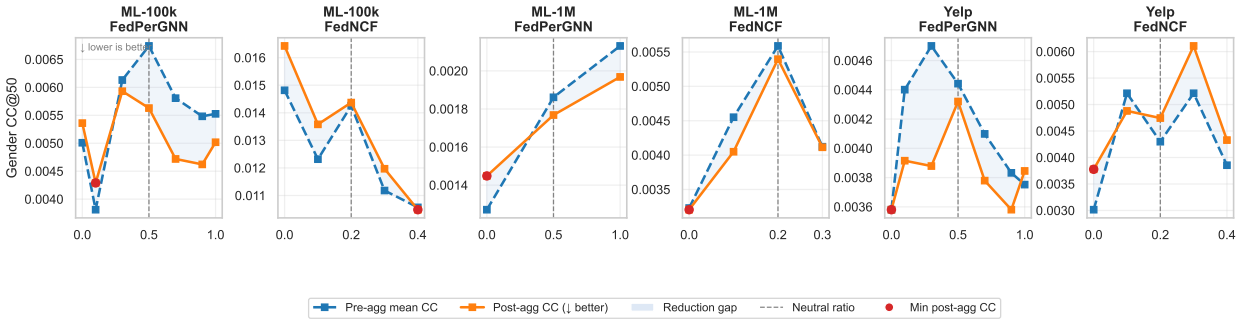


Figure 6: Gender bias (CC@50) before and after aggregation as the proportion of stereotypically behaving clients (i.e., those whose preferences align with gender-group norms) varies from 0.0 to 1.0, where 0.0 indicates rounds composed exclusively of non-stereotypical users and 1.0 indicates rounds composed exclusively of stereotypical users. For FedNCF, a baseline of 60% of selected clients are neutral users, neither stereotypical nor non-stereotypical, with the remaining 40% allocated to the varying stereotypical/non-stereotypical ratio.

To compute the stereotypicality score, we first derive a group-level category preference profile for each fairness group $g \in A$. Let $U_g \subseteq U$ denote the set of users belonging to group g, and let $R_u \subseteq V$ denote the set of items interacted with by user u. For each category $c \in C$, we define the mean group-category affinity as:

$$\mu_{g,c} = \frac{1}{|U_g|} \sum_{u \in U_g} \frac{\sum_{v \in R_u} \mathbb{1}_{\{c \in C_v\}}}{|R_u|}. \tag{13}$$

This yields a group preference vector $\boldsymbol{\mu}_g = (\mu_{g,c})_{c \in C} \in \mathbb{R}^{|C|}$ for each group g, encoding the average category exposure across all members of that group. For each user $u \in U_g$, we similarly construct a personal category proportion vector $\mathbf{p}_u = (p_{u,c})_{c \in C}$, where:

$$p_{u,c} = \frac{\sum_{v \in R_u} \mathbb{1}_{\{c \in C_v\}}}{|R_u|}. \tag{14}$$

The stereotypicality score of user u with respect to their group g is then defined as the KL divergence from their personal profile to the group profile:

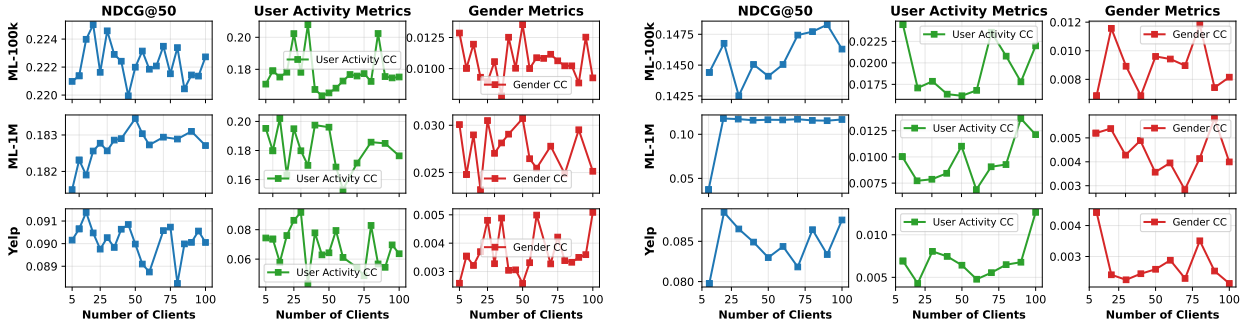
$$\mathcal{S}(u, g) = D_{\mathrm{KL}}(\mathbf{p}_u \,\|\, \boldsymbol{\mu}_g) = \sum_{c \in C} p_{u,c} \log \frac{p_{u,c}}{\mu_{g,c}}. \tag{15}$$

A lower value of $\mathcal{S}(u, g)$ indicates that user u's interaction history closely mirrors the average preference profile of their group, and is therefore considered more stereotypical. Users in the bottom τ% of the score distribution within their group are labeled stereotypical, while users in the top τ% are labeled non-stereotypical, with $\tau = 10$ by default.

Our results shown in Figure 6 show that both the selection mechanism and the aggregation step contribute to bias reduction in FL settings, with stereotypicality composition of the client pool providing a complementary and largely independent source of the cancellation effect. Runs dominated by high-stereotypicality clients converge to higher bias in general, while mixed or low-stereotypicality pools reproduce the suppression effect seen in default training. Together, these experiments provide converging evidence that the bias reduction observed in federated recommender systems is not an artifact of a single mechanism but arises from the interplay of whom the system selects and how it aggregates their contributions.

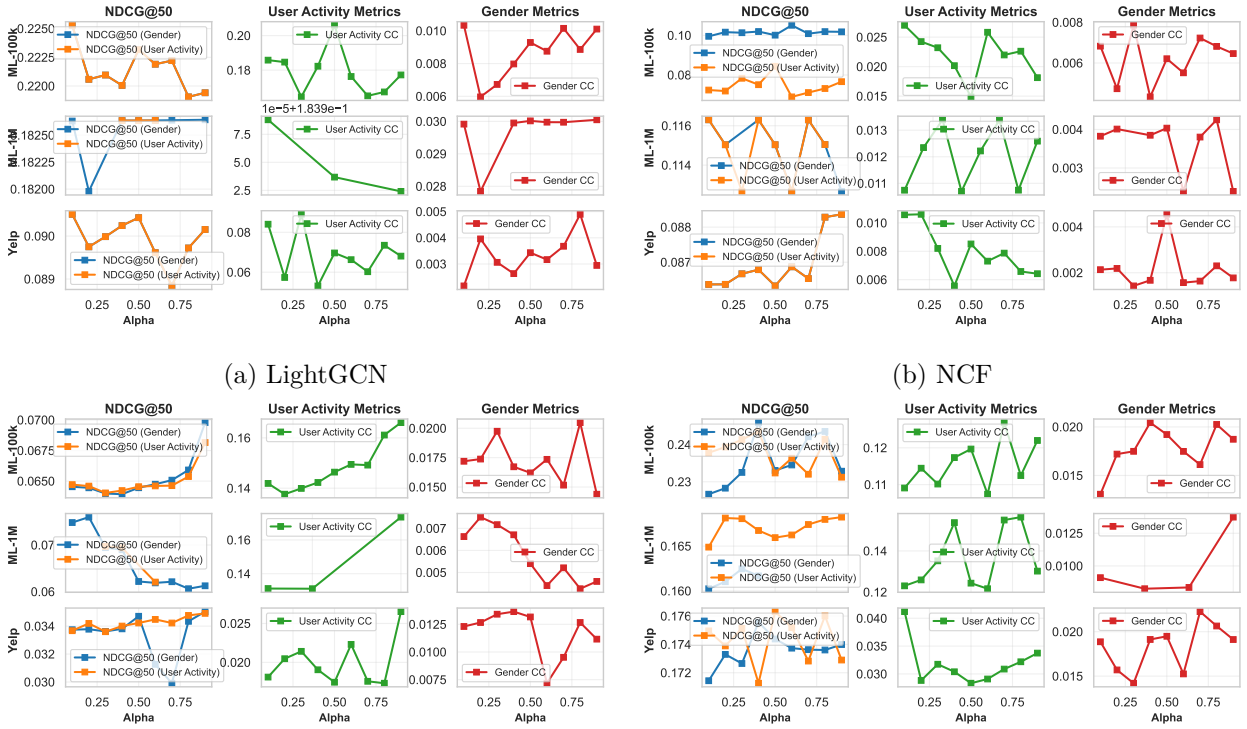
6.4 Hyperparameter Analysis

To better understand the sensitivity of federated bias and mitigation outcomes to the key system and privacy parameters, we perform additional experiments where we vary one factor at a time while keeping all others fixed, and report the resulting changes in both utility (nDCG@50) and bias (CC@50). Figures 7-10 summarize the results on multiple datasets.



(a) LightGCN (b) NCF

Figure 7: Impact of different numbers of selected clients per round for FedPerGNN (LightGCN) and FedNCF(NCF).



(a) LightGCN (b) NCF

(c) MF (d) VAE

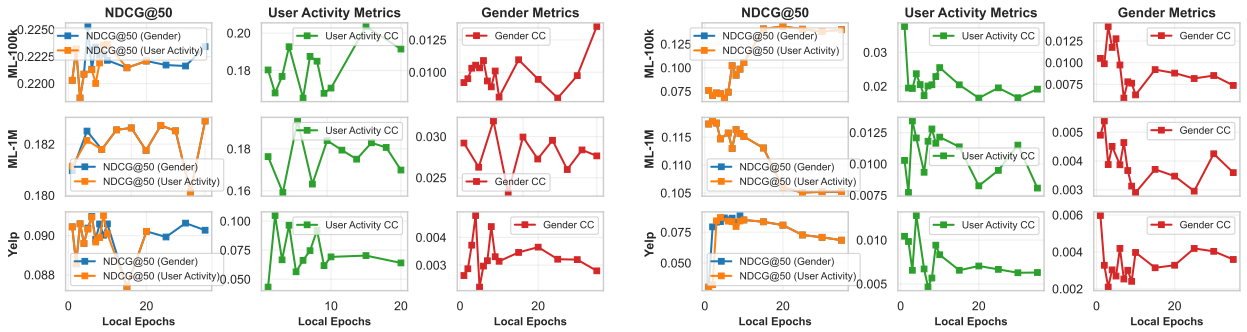
Figure 8: Comparison between different values of α.

6.4.1 Impact of sampled clients on the bias

We evaluate the impact of the number of selected clients per round shown in Figure 7. We can see that while the number of selected clients impacts performance up to a certain number of clients, it has a smaller effect on bias, the effect is variable, and while we can observe a small increase overall, it is not very significant.

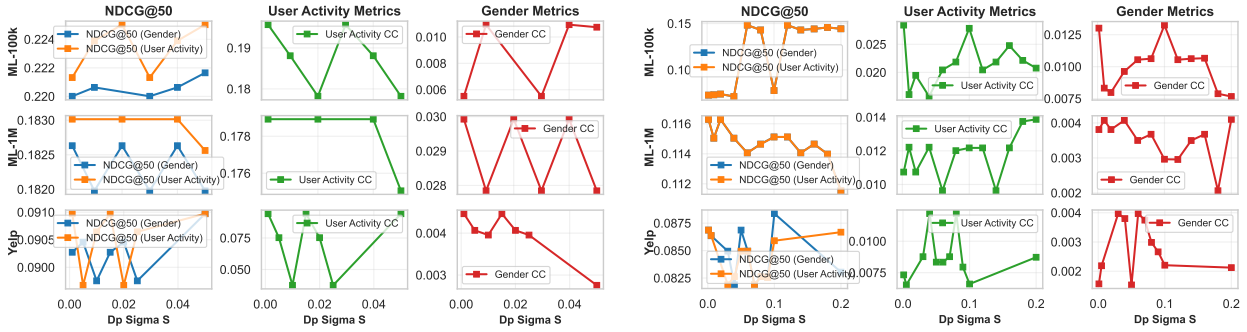
6.4.2 Penalty strength α.

We vary the strength of the mitigation penalty α (Figure 8) to quantify the fairness-utility trade-off and identify regimes where mitigation is either too weak to affect exposure disparities or too strong and harms



(a) LightGCN (b) NCF

Figure 9: Comparison between different numbers of local epochs for FedPerGNN (LightGCN) and FedNCF(NCF).



(a) LightGCN (b) NCF

Figure 10: Impact of added noise layer for local statistics computations FedPerGNN (LightGCN) and FedNCF(NCF).

ranking performance. However, we can see that α has little to no impact on both the performance and bias scores of the system, this result is in complete opposition to what we see in the original study in centralized settings Kheya et al. (2025). This observation, coupled to the smaller reduction in bias by the regulizer, might suggest that the aggregation process disturbs the impact of $\mathcal{L}_{fair}$ in Equation 10.

6.4.3 Number of local epochs

One concern that we might have with the privacy-aware mitigation strategy is that the fairness loss $\mathcal{L}_{\text{fair}}$ only gets updated once every round at the server level, which corresponds to n local epochs. This is one of the major differences between the results of the original regularization term and the privacy-aware variant. In Figure 9, we show the effect of the bias mitigation varying the number of local epochs from 1 to 10, with 1 being equivalent to FedSGD, where the communication is done after each loss, and should give the exact same results to the original loss computation, assuming no added noise and clipping. Since both EFVAE and MetaMF work with an FedSGD assumption, they only have a single local epoch per round, as such we ignore them for this comparison. Interestingly, we observe both an increase in performance and a slight decrease in bias, with added interactions, meaning that computing the bias once every few optimization steps can improve performance. In future work, this might be an interesting hypothesis to test in centralized settings.

6.4.4 Noise scale for private statistics.

To test the stability of the mitigation strategy to noise, we vary the amount of noise injected into the local statistics used by the privacy-aware mitigation (Figure 10). Because of limitations of space in this paper, we only show the plots for LightGCN and NCF. This ablation characterizes how privacy constraints (stronger noise) affect the stability of the global mitigation signal and the resulting bias reduction. We see that a small amount of noise does not affect the bias reduction; however, bigger perturbations have a more considerable impact on the bias. And as expected additonal noise on statistics doesn't affect the performance of the system very much, since $\mathcal{L}_{fair}$, whether representative of global statistics, can be viewed as a penalty to the model loss $\mathcal{L}_{rec}$. Additionally, the smoothing effect described previously may further limit the impact of a noisy $\mathcal{L}_{fair}$ on the model performance.

7 Conclusions

This paper provides a systematic investigation of how transitioning from centralized to cross-device FR settings alters user-side bias dynamics. By comparing closely aligned model families (MF/MetaMF, NCF/FedNCF, VAE/EFVAE, and LightGCN-based FedPerGNN) under identical evaluation protocols, we isolate the effect of the learning paradigm itself on exposure disparities related to gender and user activity. Our empirical findings challenge a common intuition that FL inherently amplifies bias. Across datasets and architectures, we observe a consistent reduction in gender-based exposure disparities under FL training. Through controlled pre- and post-aggregation experiments, we show that this reduction arises from two complementary mechanisms: local training on small, private datasets naturally captures fewer bias signals than centralized optimization, and the parameter-space averaging inherent to FedAvg acts as an implicit regularizer on group-level exposure disparity. Furthermore, we show that the stereotypicality composition of the client pool is an independent and complementary driver of bias reduction: rounds dominated by highly-stereotypical clients converge to higher bias, while balanced or low-stereotypicality pools reproduce the suppression effect. In contrast, activity-based bias exhibits model-dependent behavior: while some architectures show attenuation, graph-based expansion mechanisms (e.g., FedPerGNN) can substantially amplify activity disparities by structurally increasing the influence of already active users. We further show that centralized fairness regularization can be adapted to cross-device federated settings through privacy-compatible aggregation of group-level sufficient statistics. While the mitigation strategy remains effective in several regimes, its impact is consistently weaker than in centralized training, and its sensitivity to the penalty strength α is substantially reduced. This suggests that aggregation dynamics and delayed global fairness signals can smooth or attenuate the effect of regularization, highlighting that fairness objectives designed for centralized optimization do not transfer transparently to federated environments. Overall, our results indicate that the relationship between FR and bias is nuanced and architecture-dependent. Current understanding of fairness mechanisms in cross-device FR remains incomplete, both theoretically and empirically. Future work should develop formal models of bias dynamics under stochastic client participation and weighted aggregation, extend analysis to additional sensitive attributes and architectures, and design fundamentally new fairness-aware optimization strategies that explicitly account for decentralization and the aggregation process.

References

Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. Controlling Popularity Bias in Learning-to-Rank Recommendation. In *RecSys*, RecSys '17, pp. 42–46, New York, NY, USA, August 2017. Association for Computing Machinery. ISBN 978-1-4503-4652-8. doi: 10.1145/3109859.3109912.

Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. User-centered evaluation of popularity bias in recommender systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '21, pp. 119–129, New York, NY, USA, 2021a. Association for Computing Machinery. ISBN 9781450383660. doi: 10.1145/3450613.3456821. URL `https://doi.org/10.1145/3450613.3456821`.

Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward Malthouse. User-centered Evaluation of Popularity Bias in Recommender Systems. In *UMAP*, UMAP '21, pp. 119–129, New York, NY, USA, June 2021b. Association for Computing Machinery. ISBN 978-1-4503-8366-0. doi: 10.1145/3450613.3456821.

Nimesh Agrawal, Anuj Kumar Sirohi, Sandeep Kumar, and Jayadeva. No Prejudice! Fair Federated Graph Neural Networks for Personalized Recommendation. In *AAAI*, volume 38, pp. 10775–10783, March 2024. doi: 10.1609/aaai.v38i10.28950.

Waqar Ali, Muhammad Ammad-Ud-Din, Xiangmin Zhou, Yan Zhang, and Jie Shao. Communication-Efficient Federated Neural Collaborative Filtering with Multi-Armed Bandits. *ACM Trans. Recomm. Syst.*, 4(1):2:1–2:28, July 2025. doi: 10.1145/3651168.

James Bell, Adrià Gascón, Tancrède Lepoint, Baiyu Li, Sarah Meiklejohn, Mariana Raykova, and Cathie Yun. ACORN: Input Validation for Secure Aggregation. In *USENIX*, pp. 4805–4822, 2023. ISBN 978-1-939133-37-3.

R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15, 2019. doi: 10.1147/JRD.2019.2942287.

Beepul Bharti, Paul Yi, and Jeremias Sulam. Estimating and controlling for equalized odds via sensitive attribute predictors. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pp. 37173–37192, Red Hook, NY, USA, December 2023. Curran Associates Inc.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical Secure Aggregation for Privacy-Preserving Machine Learning. In *SIGSAC*, CCS '17, pp. 1175–1191, New York, NY, USA, October 2017. Association for Computing Machinery. ISBN 978-1-4503-4946-8. doi: 10.1145/3133956.3133982.

Ludovico Boratto, Gianni Fenu, Mirko Marras, and Giacomo Medda. Consumer Fairness in Recommender Systems: Contextualizing Definitions and Mitigations. In *Advances in Information Retrieval*, pp. 552–566, Cham, 2022. Springer International Publishing. ISBN 978-3-030-99736-6. doi: 10.1007/978-3-030-99736-6_37.

Robin Burke. Multisided Fairness for Recommendation, July 2017.

Di Chai, Leye Wang, Kai Chen, and Qiang Yang. Secure Federated Matrix Factorization. *IEEE Intelligent Systems*, 36(5):11–20, September 2021. ISSN 1941-1294. doi: 10.1109/MIS.2020.3014880.

Joymallya Chakraborty, Suvodeep Majumder, Zhe Yu, and Tim Menzies. Fairway: a way to build fair ml software. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2020, pp. 654–665, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370431. doi: 10.1145/3368089.3409697. URL `https://doi.org/10.1145/3368089.3409697`.

Damien Dablain, Bartosz Krawczyk, and Nitesh Chawla. Towards A Holistic View of Bias in Machine Learning: Bridging Algorithmic Fairness and Imbalanced Learning, July 2022.

Christophe Denis, Romuald Elie, Mohamed Hebiri, and François Hu. Fairness guarantees in multi-class classification with demographic parity. *Journal of Machine Learning Research*, 25(130):1–46, 2024. ISSN 1533-7928.

M. Dogra, B.P. Meher, P.V. Mani, and H.-K. Min. Memory Efficient Federated Recommendation Model. In *ICSC*, pp. 139–142, 2022. doi: 10.1109/ICSC52841.2022.00028.

Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Hassan Awadallah, and Xia Hu. Fairness via representation neutralization. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021a. Curran Associates Inc. ISBN 9781713845393.

Yongjie Du, Deyun Zhou, Yu Xie, Jiao Shi, and Maoguo Gong. Federated matrix factorization for privacy-preserving recommender systems. *Applied Soft Computing*, 111:107700, November 2021b. ISSN 1568-4946. doi: 10.1016/j.asoc.2021.107700.

Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Conference on Fairness, Accountability and Transparency*, pp. 172–186. PMLR, January 2018.

Andres Ferraro, Xavier Serra, and Christine Bauer. Break the Loop: Gender Imbalance in Music Recommenders. In *CHIIR*, CHIIR '21, pp. 249–254, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8055-3. doi: 10.1145/3406522.3446033.

Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, and Gerard de Melo. Fairness-aware explainable recommendation over knowledge graphs, 2020.

Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 2221–2231, New York, NY, USA, July 2019. Association for Computing Machinery. ISBN 978-1-4503-6201-6. doi: 10.1145/3292500.3330691.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3323–3331, Red Hook, NY, USA, December 2016. Curran Associates Inc. ISBN 978-1-5108-3881-9.

F. Maxwell Harper and Joseph A. Konstan. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.*, 5(4):19:1–19:19, December 2015. ISSN 2160-6455. doi: 10.1145/2827872.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural Collaborative Filtering. In *WWW*, WWW '17, pp. 173–182, Republic and Canton of Geneva, CHE, April 2017. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4913-0. doi: 10.1145/3038912.3052569.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, YongDong Zhang, and Meng Wang. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *SIGIR*, pp. 639–648, New York, NY, USA, July 2020. ACM. ISBN 978-1-4503-8016-4. doi: 10.1145/3397271.3401063.

Jonathan Huang, Galal Galal, Mozziyar Etemadi, and Mahesh Vaidyanathan. Evaluation and Mitigation of Racial Bias in Clinical Machine Learning Models: Scoping Review. *JMIR Medical Informatics*, 10(5): e36388, May 2022. doi: 10.2196/36388.

Vasileios Iosifidis, Besnik Fetahu, and Eirini Ntoutsi. Fae: A fairness-aware ensemble framework. In *2019 IEEE International Conference on Big Data (Big Data)*, pp. 1375–1380, 2019. doi: 10.1109/BigData47090.2019.9006487.

Yitong Ji, Aixin Sun, Jie Zhang, and Chenliang Li. Do Loyal Users Enjoy Better Recommendations? Understanding Recommender Accuracy from a Time Perspective. In *SIGIR*, pp. 92–97, August 2022. doi: 10.1145/3539813.3545124.

Saikishore Kalloori and Severin Klingler. Horizontal cross-silo federated recommender systems. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, pp. 680–684, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384582. doi: 10.1145/3460231.3478863. URL `https://doi.org/10.1145/3460231.3478863`.

Tahsin Alamgir Kheya, Mohamed Reda Bouadjenek, and Sunil Aryal. The Pursuit of Fairness in Artificial Intelligence Models: A Survey, March 2024.

Tahsin Alamgir Kheya, Mohamed Reda Bouadjenek, and Sunil Aryal. Unmasking Gender Bias in Recommendation Systems and Enhancing Category-Aware Fairness. In *Web Conference*, WWW '25, pp. 5127–5138. Association for Computing Machinery, April 2025. ISBN 979-8-4007-1274-6. doi: 10.1145/3696410.3714528.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8):30–37, August 2009. ISSN 1558-0814. doi: 10.1109/MC.2009.263.

Dominik Kowald, Markus Schedl, and Elisabeth Lex. The Unfairness of Popularity Bias in Music Recommendation: A Reproducibility Study. In *ECIR*, pp. 35–42, Berlin, Heidelberg, April 2020. Springer-Verlag. ISBN 978-3-030-45441-8. doi: 10.1007/978-3-030-45442-5_5.

Peng Li, Xinru Zhu, Xiaoshan Li, and Baofeng Huo. Research on mitigating popularity bias in federal recommendation based on users' behavior. *The Journal of Supercomputing*, 81(4):616, March 2025. ISSN 1573-0484. doi: 10.1007/s11227-025-07144-7.

Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. User-oriented fairness in recommendation. In *Proceedings of the Web Conference 2021*, WWW '21, pp. 624–632, New York, NY, USA, 2021a. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3449866. URL https://doi.org/10.1145/3442381.3449866.

Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. User-oriented Fairness in Recommendation. In *Web Conference*, WWW '21, pp. 624–632, New York, NY, USA, June 2021b. Association for Computing Machinery. ISBN 978-1-4503-8312-7. doi: 10.1145/3442381.3449866.

Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. Variational Autoencoders for Collaborative Filtering. In *WWW*, WWW '18, pp. 689–698, Republic and Canton of Geneva, CHE, April 2018. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-5639-8. doi: 10.1145/3178876.3186150.

Kun Lin, Nasim Sonboli, Bamshad Mobasher, and Robin Burke. Calibration in Collaborative Filtering Recommender Systems: A User-Centered Analysis. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, HT '20, pp. 197–206, New York, NY, USA, July 2020a. Association for Computing Machinery. ISBN 978-1-4503-7098-1. doi: 10.1145/3372923.3404793.

Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Dongxiao Yu, Jun Ma, Maarten de Rijke, and Xiuzhen Cheng. Meta Matrix Factorization for Federated Rating Predictions. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pp. 981–990, New York, NY, USA, July 2020b. Association for Computing Machinery. ISBN 978-1-4503-8016-4. doi: 10.1145/3397271.3401081.

Shuchang Liu, Yingqiang Ge, Shuyuan Xu, Yongfeng Zhang, and Amelie Marian. Fairness-aware Federated Matrix Factorization. In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, pp. 168–178, New York, NY, USA, September 2022. Association for Computing Machinery. ISBN 978-1-4503-9278-5. doi: 10.1145/3523227.3546771.

Masoud Mansoury, Bamshad Mobasher, Robin Burke, and Mykola Pechenizkiy. Bias Disparity in Collaborative Recommendation: Algorithmic Evaluation and Comparison, August 2019.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6):115:1–115:35, July 2021. ISSN 0360-0300. doi: 10.1145/3457607.

Zaiqiao Meng, Richard McCreadie, Craig Macdonald, and Iadh Ounis. Exploring Data Splitting Strategies for the Evaluation of Recommendation Models. In *RecSys*, RecSys '20, pp. 681–686, New York, NY, USA, September 2020. Association for Computing Machinery. ISBN 978-1-4503-7583-2. doi: 10.1145/3383313.3418479.

Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. Fairness in Risk Assessment Instruments: Post-Processing to Achieve Counterfactual Equalized Odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 386–400, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445902.

Khalil Muhammad, Qinqin Wang, Diarmuid O'Reilly-Morgan, Elias Tragos, Barry Smyth, Neil Hurley, James Geraci, and Aonghus Lawlor. FedFast: Going Beyond Average for Faster Training of Federated Recommender Systems. In *SIGKDD*, KDD '20, pp. 1234–1242, New York, NY, USA, August 2020. Association for Computing Machinery. ISBN 978-1-4503-7998-4. doi: 10.1145/3394486.3403176.

Mohammadmehdi Naghiaei, Hossein A. Rahmani, and Yashar Deldjoo. Cpfair: Personalized consumer and producer fairness re-ranking for recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 770–779, 2022. ISBN 9781450387323.

Luong Vuong Nguyen, Quoc-Trinh Vo, and Thi-Thu-Hong Phan. A Survey of Recommendation Systems: Datasets, Evaluation Methods, and Application Domains. In *Intelligent Systems Design and Applications*, pp. 311–322, Cham, 2024a. Springer Nature Switzerland. ISBN 978-3-031-64779-6. doi: 10.1007/978-3-031-64779-6_30.

Ngoc-Hieu Nguyen, Tuan-Anh Nguyen, Tuan Nguyen, Vu Tien Hoang, Dung D. Le, and Kok-Seng Wong. Towards Efficient Communication and Secure Federated Recommendation System via Low-rank Training. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pp. 3940–3951, New York, NY, USA, May 2024b. Association for Computing Machinery. ISBN 979-8-4007-0171-9. doi: 10.1145/3589334.3645702.

António Pereira Barata, Frank W. Takes, H. Jaap van den Herik, and Cor J. Veenman. Fair tree classifier using strong demographic parity. *Machine Learning*, 113(5):3305–3324, May 2024. ISSN 1573-0565. doi: 10.1007/s10994-023-06376-z.

Vasileios Perifanis and Pavlos S. Efraimidis. Federated Neural Collaborative Filtering. *Knowledge-Based Systems*, 242:108441, April 2022. ISSN 0950-7051. doi: 10.1016/j.knosys.2022.108441.

Tao Qi, Fangzhao Wu, Chuhan Wu, Peijie Sun, Le Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. Profairrec: Provider fairness-aware news recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pp. 1164–1173, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3532046. URL https://doi.org/10.1145/3477495.3532046.

Liang Qu, Ningzhi Tang, Ruiqi Zheng, Quoc Viet Hung Nguyen, Zi Huang, Yuhui Shi, and Hongzhi Yin. Semi-decentralized Federated Ego Graph Learning for Recommendation. In *The Web Conference*, WWW '23, pp. 339–348, New York, NY, USA, April 2023. ACM. ISBN 978-1-4503-9416-1. doi: 10.1145/3543507.3583337.

Bashir Rastegarpanah, Krishna P. Gummadi, and Mark Crovella. Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pp. 231–239, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359405. doi: 10.1145/3289600.3291002. URL https://doi.org/10.1145/3289600.3291002.

Lucas Rosenblatt and R. Teal Witter. Counterfactual fairness is basically demographic parity. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14461–14469, Jun. 2023. doi: 10.1609/aaai.v37i12.26691. URL https://ojs.aaai.org/index.php/AAAI/article/view/26691.

Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2219–2228, 2018. ISBN 9781450355520.

Jinhyun So, Ramy E. Ali, Başak Güler, Jiantao Jiao, and A. Salman Avestimehr. Securing secure aggregation: Mitigating multi-round privacy leakage in federated learning. In *AAAI*, volume 37 of *AAAI'23/IAAI'23/EAAI'23*, pp. 9864–9873. AAAI Press, February 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i8.26177.

Zehua Sun, Yonghui Xu, Yong Liu, Wei He, Lanju Kong, Fangzhao Wu, Yali Jiang, and Lizhen Cui. A Survey on Federated Recommendation Systems. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2024. ISSN 2162-2388. doi: 10.1109/TNNLS.2024.3354924.

Lingyun Wang, Hanlin Zhou, Yinwei Bao, Xiaoran Yan, Guojiang Shen, and Xiangjie Kong. Horizontal Federated Recommender System: A Survey. *ACM Comput. Surv.*, 56(9):240:1–240:42, May 2024. ISSN 0360-0300. doi: 10.1145/3656165.

Zhihao Wang, He Bai, Wenke Huang, Duantengchuan Li, Jian Wang, and Bing Li. Federated Recommendation with Explicitly Encoding Item Bias. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(12):12792–12800, April 2025. ISSN 2374-3468. doi: 10.1609/aaai.v39i12.33395.

Tianxin Wei and Jingrui He. Comprehensive Fair Meta-learned Recommender System. In *CIKM*, KDD '22, pp. 1989–1999, New York, NY, USA, August 2022. Association for Computing Machinery. ISBN 978-1-4503-9385-0. doi: 10.1145/3534678.3539269.

Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Tao Qi, Yongfeng Huang, and Xing Xie. A federated graph neural network framework for privacy-preserving personalization. *Nature Communications*, 13(1):3091, June 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-30714-9.

Yueqing Xuan, Kacper Sokol, Mark Sanderson, and Jeffrey Chan. Evaluating and Addressing Fairness Across User Groups in Negative Sampling for Recommender Systems. In *CIKM*, CIKM '25, pp. 3720–3729, New York, NY, USA, November 2025. Association for Computing Machinery. ISBN 979-8-4007-2040-6. doi: 10.1145/3746252.3761263.

Jenny Yang, Andrew A. S. Soltan, David W. Eyre, and David A. Clifton. Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. *Nature Machine Intelligence*, 5 (8):884–894, August 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00697-3.

Ke Yang, Joshua R. Loftus, and Julia Stoyanovich. Causal intersectionality for fair ranking, June 2020.

Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. *Advances in neural information processing systems*, 30, 2017.

Ming Yin, Yichang Xu, Minghong Fang, and Neil Zhenqiang Gong. Poisoning Federated Recommender Systems with Fake Users. In *Proceedings of the ACM on Web Conference 2024*, WWW '24, pp. 3555–3565, New York, NY, USA, May 2024. Association for Computing Machinery. ISBN 979-8-4007-0171-9. doi: 10.1145/3589334.3645492.

Zhe Yu, Joymallya Chakraborty, and Tim Menzies. FairBalance: How to Achieve Equalized Odds With Data Pre-Processing. *IEEE Trans. Softw. Eng.*, 50(9):2294–2312, September 2024. ISSN 0098-5589. doi: 10.1109/TSE.2024.3431445.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 335–340, New York, NY, USA, 2018a. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278779. URL `https://doi.org/10.1145/3278721.3278779`.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 335–340, New York, NY, USA, December 2018b. Association for Computing Machinery. ISBN 978-1-4503-6012-8. doi: 10.1145/3278721.3278779.

Chunxu Zhang, Guodong Long, Hongkuan Guo, Xiao Fang, Yang Song, Zhaojie Liu, Guorui Zhou, Zijian Zhang, Yang Liu, and Bo Yang. Federated Adaptation for Foundation Model-based Recommendations. In *IJCAI*, pp. 5453–5461, Jeju, South Korea, August 2024a. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/603.

Chunxu Zhang, Guodong Long, Tianyi Zhou, Zijian Zhang, Peng Yan, and Bo Yang. GPFedRec: Graph-Guided Personalization for Federated Recommendation. In *SIGKDD*, KDD '24, pp. 4131–4142, New York, NY, USA, August 2024b. ACM. ISBN 979-8-4007-0490-1. doi: 10.1145/3637528.3671702.

Lu Zhang, Qian Rong, Xuanang Ding, Guohui Li, and Ling Yuan. EFVAE: Efficient Federated Variational Autoencoder for Collaborative Filtering. In *CIKM*, CIKM '24, pp. 3176–3185, New York, NY, USA, October 2024c. Association for Computing Machinery. ISBN 979-8-4007-0436-9. doi: 10.1145/3627673.3679818.

Shijie Zhang, Hongzhi Yin, Tong Chen, Zi Huang, Quoc Viet Hung Nguyen, and Lizhen Cui. PipAttack: Poisoning Federated Recommender Systems for Manipulating Item Promotion. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, pp. 1415–1423, New York, NY, USA, February 2022. Association for Computing Machinery. ISBN 978-1-4503-9132-0. doi: 10.1145/3488560.3498386.

Shijie Zhang, Wei Yuan, and Hongzhi Yin. Comprehensive Privacy Analysis on Federated Recommender System Against Attribute Inference Attacks. *IEEE Trans. on Knowl. and Data Eng.*, 36(3):987–999, March 2024d. ISSN 1041-4347. doi: 10.1109/TKDE.2023.3295601.

Xing Zhang, Yuexiang Luo, and Tianning Li. A Review of Research on Secure Aggregation for Federated Learning. *Future Internet*, 17(7):308, July 2025. ISSN 1999-5903. doi: 10.3390/fi17070308.

Ziwei Zhu, Xia Hu, and James Caverlee. Fairness-Aware Tensor-Based Recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pp. 1153–1162, New York, NY, USA, October 2018. Association for Computing Machinery. ISBN 978-1-4503-6014-2. doi: 10.1145/3269206.3271795.

Ziwei Zhu, Jianling Wang, and James Caverlee. Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems. In *SIGIR*, SIGIR '20, pp. 449–458, New York, NY, USA, July 2020. Association for Computing Machinery. ISBN 978-1-4503-8016-4. doi: 10.1145/3397271.3401177.