CUGF: A Reliable and Fair Recommendation Framework

Nitin Bisht[1], Xiuwen Gong[1*], Guandong Xu[2, 1*]

[1]University of Technology, Sydney
[2]The Education University of Hong Kong
Nitin.Bisht@student.uts.edu.au, Xiuwen.Gong@uts.edu.au, gdxu@eduhk.hk

Abstract

Recommender systems (RS) play a crucial role in assisting decision-making but often suffer from either a lack of credibility or unfairness problems. A few recommendation models have endeavored to address the problem from only one aspect, and approaches to solving both problems remain to be explored. This paper aims to construct a generalized fairness-based recommendation framework that can also provide the credibility of recommendation models. Generally, we propose a reliable and fair recommendation framework called Conformalized User Group Fairness (CUGF) based on the inspiration of conformal prediction. Specifically, we construct dynamic prediction sets that are guaranteed to cover the true item with a user-pre-specified probability to ensure credibility while designing novel fairness metrics based on empirical risks to guarantee the fairness of users across different groups. Furthermore, we design a novel CUGF Algorithm to optimize the parameter γ that dominates the prediction sets and also the fairness. Besides, we conduct extensive experiments by applying CUGF on top of various recommendation models and representative datasets to validate its effectiveness with respect to recommendation performance (in terms of average set size) and fairness (in terms of the two defined fairness metrics), the results of which demonstrate the validity of the proposed framework.

Introduction

Recommendation models (Ricci, Rokach, and Shapira 2021; Aggarwal 2016; Jannach et al. 2010) usually recommend the items with the maximum or the top-k relevance scores in the models' output. Although it is important to recommend the most likely items, it is also very important to provide the credibility of recommendation models. Some researchers try to develop methods that provide the model confidence. However, these methods (Knyazev and Oosterhuis 2023; Cui et al. 2024) are heuristic modeling without statistical guarantee. Moreover, most existing recommendation models (Han et al. 2021; Kim and Suh 2019; Albora, Rossi Mori, and Zaccaria 2023) fail to consider the fairness problem, which is essential in ensuring the long-term success and credibility of digital platforms. For example, an unfair job recommendation platform might consistently overlook talented candidates from less renowned universities or regions, or a streaming service like Spotify where the music of mainstream artists is perpetually favored, leaving equally talented groups of indie musicians in the shadows. Such biases can alienate users, skew cultural representation, and ultimately drive people away from the platform. Recently, a few recommendation models (Zhu et al. 2018; Geyik, Ambler, and Kenthapadi 2019; Islam et al. 2021) try to solve the fairness problem. However, these models address specific fairness issues in certain fields but lack a generalized fairness-based recommendation framework that can also ensure model credibility.

Motivated by the above-mentioned problems, we would like to seek actionable uncertainty quantification for the RS to provide a reliable and fair recommendation framework. Specifically, we require a prediction set that reliably covers the true item with a high probability (e.g., 90%) rather than an estimate of the most likely outcomes, while guaranteeing all users' fairness. The objectives of our framework would be to 1) construct a prediction set that can cover the true item with a user-pre-defined probability while 2) guaranteeing the fairness of users across different groups.

Inspired by conformal prediction (CP) (Vovk, Gammerman, and Shafer 2005), a powerful statistical tool to quantify uncertainty, we propose a novel framework called Conformalized User Group Fairness (CUGF) to achieve the above-mentioned objectives. In general, we apply a conformal prediction framework to produce the prediction set on top of pre-built recommendation models following the general CP recipe. However, conformal prediction cannot be directly applied due to the following challenge: how will conformal prediction conformalize fairness of users in different groups to generate the prediction set? And it remains a doubt whether prediction sets still preserve the coverage guarantee.

To solve the above challenges, we create dynamic prediction sets for users in different groups to meet the fairness constraint. Specifically, we first define the recommendation problem and formulate it into the standard procedures of conformal prediction. We then design novel fairness metrics based on empirical risk corresponding to hit rate or NDCG to align with the design of the set predictor, which also makes the proposed framework more flexible by accommodating different types of loss functions. Furthermore, we de-

*Xiuwen Gong and Guandong Xu are the corresponding authors.

sign a novel optimization algorithm (i.e., CUGF Algorithm) for the proposed framework based on a greedy strategy to learn the optimal prediction sets that preserve the coverage guarantee. Lastly, we conduct comprehensive experiments on various datasets and recommendation models to validate the effectiveness of our proposed framework with respect to the recommendation performance (in terms of average set size) and fairness (in terms of the two fairness metrics).

Related Works

Recommendation. Recommender systems (Ko et al. 2022; Lu et al. 2015), have been thoroughly studied over past decades by developing recommendation models for different fields of application, such as e-commerce (Schafer, Konstan, and Riedl 1999), media streaming (Chang et al. 2017), social networks (He and Chu 2010) etc, which can help users make decisions via personalized content or product recommendations. To ensure long-term success of digital platforms, credibility and fairness of recommendation are two crucial factors today that are urgently needed to ensure the satisfaction of customers. Traditional recommendation models primarily focused on accuracy (Adomavicius and Tuzhilin 2005; Ricci et al. 2010), but there is an increasing recognition that model confidence—quantifying the reliability of the recommendations—is equally important. Researchers try to develop recommendation approaches that provide the model confidence/credibility. For example, Kweon et.al. (2024) enhances model confidence by dynamically adjusting recommendation strategies based on performance metrics, ensuring the model adapts to user interactions and maintains high predictive accuracy. Naghiaei et al. (2022) propose confidence-aware optimization-based re-ranking algorithm that accounts for calibration confidence based on user profile size. However, these methods are heuristic modeling without statistical guarantee. Meanwhile, some fairness-based recommendation models have been developed in recent years, which usually focus on a particular fairness issue in specific fields of application. For example, Li et al. (2021) developed a fairness algorithm for e-commerce platforms that ensures all groups of users receive equitable recommendations. Stratigi et al. (2017) focus on developing tools to recommend relevant and fair information to different groups of patients. Islam et al. (2021) develop a framework for reducing gender bias in recommending career-related sensitive items, such as jobs, academic concentrations, or courses of study. Although these methods alleviate the fairness problem to some extent, they fail to construct a generalized fairness-based recommendation framework, as also pointed out by Rahmani et al. (2022) that the fairness algorithms are mostly dataset-dependent and are effective only for a particular group of datasets. As a result, aligning with broader multi-objective approaches in machine learning (Liu, Tsang, and Muller 2017; Mao et al. 2020, 2022), there is an urgent need to develop a generalized recommendation framework that meets the fairness principles while ensuring the credibility of the recommendation model.

Conformal Prediction. Conformal prediction (CP) (Angelopoulos and Bates 2023; Romano, Patterson, and Candes 2019; Vovk, Gammerman, and Shafer 2005) is powerful statistical tool to quantify models' uncertainty and can provide finite sample guarantee for prediction sets, which is widely applied to improve the reliability of machine learning models in real-world applications, such as computer vision (Angelopoulos et al. 2021), natural language processing (Fisch et al. 2021) and large language models (Quach et al. 2023). Specifically, for a pre-defined error rate $\alpha \in (0,1)$ and n calibration points, CP generates prediction set $C_\alpha(X_{n+1})$ that is guaranteed to include true value Y_{n+1} with coverage guarantee of at least $1-\alpha$ for a new point X_{n+1} in Eq. (1):

$$\mathbb{P}\left(Y_{n+1} \in C_\alpha(X_{n+1})\right) \geq 1-\alpha. \tag{1}$$

In the next section, we would like to apply the paradigm of conformal prediction to build our framework on top of base recommendation models by taking them as black-box to improve their credibility while ensuring fairness for users across different groups.

The Proposed Framework

In this section, we propose a novel recommendation framework, named Conformalized User Group Fairness (CUGF), to construct prediction sets from any pre-trained recommendation models that are formally guaranteed to contain true items with a desired probability while guaranteeing the users' fairness across both groups.

To start with, we introduce some notations used in paper. Consider n items, denoted as $\mathbf{i} = (i^j)_{j\in\{1,2,\ldots,n\}}$, where each item i^j is an element of item space $\mathcal{I}$. Similarly, we use $\mathbf{u} = (u^k)_{k\in\{1,2,\ldots,m\}}$ to represent m users, where each user u^k belongs to the user space $\mathcal{U}$. For brevity, we use i and u to denote an item and a user, respectively, throughout the paper. In the user group fairness setting, we follow (Li et al. 2021) to divide the group information into advantaged group G_1 and disadvantaged group G_2, which are required to satisfy $G_1 \cap G_2 = \emptyset$ and $G_1 \cup G_2 = \mathcal{U}$. This ensures each user belongs to a unique group and the entire user set is covered. Besides, we use $m : \mathcal{U} \times \mathcal{I} \to [0,1]$ to represent recommendation model, which maps user u and item i to a recommendation score $m(u,i)$. Furthermore, we define set predictor for each user u as $T_\gamma(u)$, which is dominated by parameter γ in monotonically decreasing manner as follows:

$$\gamma_1 < \gamma_2 \implies T_{\gamma_2}(u) \subset T_{\gamma_1}(u). \tag{2}$$

To achieve the first objective, we first define the groupwise coverage $C(\gamma_G)$ as follows:

$$C(\gamma_G) := \mathbb{P}\left(i_{\text{true}} \in T_{\gamma_G}(u)\right) \geq 1-\alpha. \tag{3}$$

Here, G represents a group that user u belongs to; γ_G is the parameter set for all users throughout group G; i_{true} is the item that user u is most interested in or most relevant within recommendation models; $1-\alpha$ is a user pre-specified confidence level, or α is the coverage parameter such as 10%.

To generate valid prediction sets in Eq (3), we follow the general procedures of the conformal prediction framework and adaptive score strategy in (Romano, Sesia, and Candes 2020). However, we cannot directly use the score function based on the relevance output of the true item in (Romano, Sesia, and Candes 2020) as our constructed set should also

meet the fairness criteria that we define later. As a result, we define the core nonconformity score function by relaxing the relevance output of the true item to an adaptive threshold based on the parameter γ that controls the prediction sets in the following formulation:

$$S(u,i) = \mathbb{I}[m(u,i_{\text{true}}) \geq \gamma] \sum_{i \in I} m(u,i)\mathbb{I}[m(u,i) \geq \gamma] \tag{4}$$

Here, $\mathbb{I}[\cdot]$ is the indicator function; $m(u,i)$ is the relevance output from recommendation models.

Based on the above nonconformity score function, we further define the quantile function for group G following the general conformal prediction recipe (Angelopoulos and Bates 2021) as below:

$$Q_G := \text{Quantile}\left(\frac{\lceil (n+1)(1-\alpha) \rceil}{n}, \{S(u,i) \mid u \in G\}\right) \tag{5}$$

Here, n is the number of users in group G.

Subsequently, we can employ the following strategy to construct the set predictor $T_{\gamma_G}(u)$:

$$T_{\gamma_G}(u) = \{i \in I : S(u,i) \leq Q_G\} \tag{6}$$

To achieve the second objective of guaranteeing fairness for users throughout the advantaged and disadvantaged groups, we follow the general fairness framework in (Fu et al. 2020), that is, constraining disparity of recommendation performance (e.g., hit rate, NDCG) for two group users to be lower than a small threshold. To align with the design of the prediction set, we employ risks corresponding to hit rate or NDCG rather than a direct use of hit rate or NDCG. The definition of fairness metric $F(T_{\boldsymbol{\gamma}})$ can be formulated as follows:

$$F(T_{\boldsymbol{\gamma}}) := \left| R(T_{\gamma_{G_1}}(u)) - R(T_{\gamma_{G_2}}(u)) \right| \leq \eta \tag{7}$$

Here, $R(\cdot)$ denotes the expected risk which is defined later; γ_{G_1} and γ_{G_2} are the parameters corresponding to group G_1 and group G_2 respectively, and users in the same group share the same parameter; η is the user pre-specified threshold.

Furthermore, we define the above-mentioned risk as the following empirical risk:

$$R(T_{\gamma_G}(u)) := \frac{1}{n} \sum_{u \in G} L(i_{\text{true}}, T_{\gamma_G}(u)) \tag{8}$$

Here, n denotes the number of users in group G; $L(.)$ denotes the loss function with respect to the true item and prediction set for a user u.

The design of our fairness metric makes the proposed framework more flexible by accommodating different types of loss functions. For example, we expect to use two different fairness metrics corresponding to the loss functions with respect to hit rate and NDCG to evaluate user fairness in later experiments. Thus, we define the loss functions corresponding to hit rate and NDCG as below. We employ the 0-1 loss function for hit rate, i.e.,

$$L(i_{\text{true}}, T_{\gamma_G}(u)) = \begin{cases} 1 & \text{if } i_{\text{true}} \notin T_{\gamma_G}(u) \\ 0 & \text{if } i_{\text{true}} \in T_{\gamma_G}(u) \end{cases} \tag{9}$$

Algorithm 1: CUGF Algorithm

Define: Coverage and Fairness metrics as in eqs. (3) and (7) respectively.
Define: Loss functions as in eqs. (9) and (10) respectively.
Goal: Find the optimal $\boldsymbol{\gamma}$ that meet the coverage guarantee in eq. (3) and fairness metric in eq. (7).
Input: Calibration Dataset $\{(u_j, i \in I)\}_{j=1}^{m}$, recommendation models $m(u,i)$, coverage parameter α, fairness threshold η.
Output: Output $\gamma_{G_1}, \gamma_{G_2}$.
Initialize $\boldsymbol{\gamma} \leftarrow [\gamma_{G_1}, \gamma_{G_2}] \leftarrow \gamma_{\text{init}}$ ▷ Initialize control parameter vector $\boldsymbol{\gamma}$ for both groups;
while $\gamma_G < 1$ **do**
    Compute $S(u,i)$ given eq. (4) for each user in the calibration dataset;
    Compute Q_{G_1}, Q_{G_2} given eq. (5);
    Construct $T_{\gamma_G}(u)$ given eq. (6) for each user;
    Compute the two losses in eqs. (9) and (10), and then calculate R_{G_1}, R_{G_2} given eq. (8);
    if Fairness metric in eq. (7) is met **then**
        Output $\gamma_{G_1}, \gamma_{G_2}$
    else
        Update $\gamma_{G_1} \leftarrow \gamma_{G_1} - \Delta_1, \gamma_{G_2} \leftarrow \gamma_{G_2} - \Delta_2$
    end if
end while

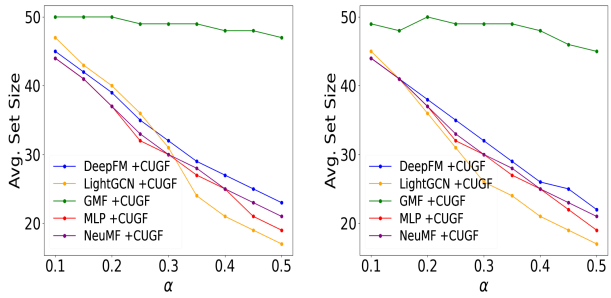
We define the following loss function for NDCG:

$$L(i_{\text{true}}, T_{\gamma_G}(u)) = \begin{cases} 1 - \frac{1}{\log_2(\text{rank}+1)} & \text{if } i_{\text{true}} \in T_{\gamma_G}(u) \\ 1 & \text{if } i_{\text{true}} \notin T_{\gamma_G}(u) \end{cases} \tag{10}$$

Here, rank refers to the order of the true item i_{true} located in the prediction set $T_{\gamma_G}(u)$.

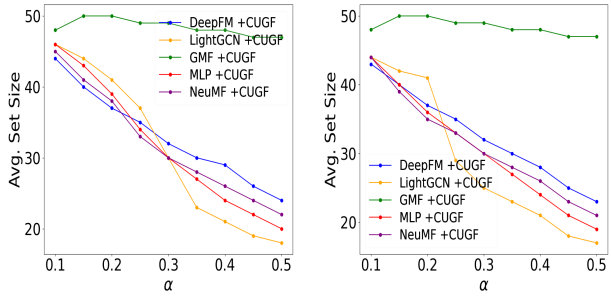
To this end, we complete building the modelling of the proposed framework, i.e., CUGF. To output the valid prediction sets that both meet the coverage and fairness guarantee by CUGF, we design a novel algorithm based on greedy strategy to optimize the parameter γ_G that controls the validity of set predictor. The complete procedures for implementing the proposed CUGF are summarized in Algorithm 1.

Recommendation After obtaining the optimal $\boldsymbol{\gamma}$ from Algorithm 1, we can do item recommendation for new customers. When a new user u_t comes, we first decide the group G that they belong to, and then apply the corresponding γ_G to calculate their non-conformity score $S(u_t,i)$ given Eq. (4), and output the prediction set $T_{\gamma_G}(u_t) = \{i \in I : S(u_t,i) \leq Q_G\}$ given Eq. (6). It is worth noting that this prediction set can meet both the coverage and fairness metrics strictly. For example, when the coverage parameter α and fairness parameter η are set to 0.2 and 0.2 respectively in Algorithm 1, the prediction set generated for the user u_t covers the true item that user u_t really interested in with a probability of 80% (confidence in recommendation performance), and the items in the prediction set are fairly recommended to the user u_t with a probability of 80% (confidence



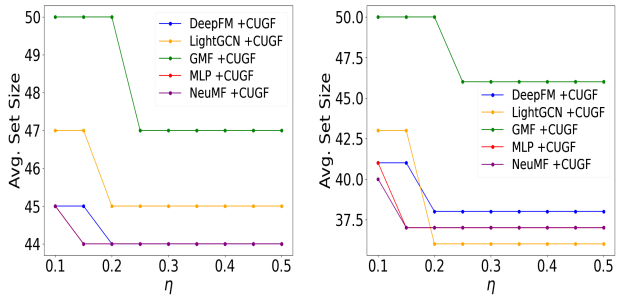
(a) $\eta = 0.15$ (b) $\eta = 0.20$

Figure 1: Performance analysis of the base models after applying the CUGF framework in terms of average set size with varying $\alpha = 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50$ on the **AmazonOffice** dataset grouped by **number of interactions** under different **fairness threshold** η.



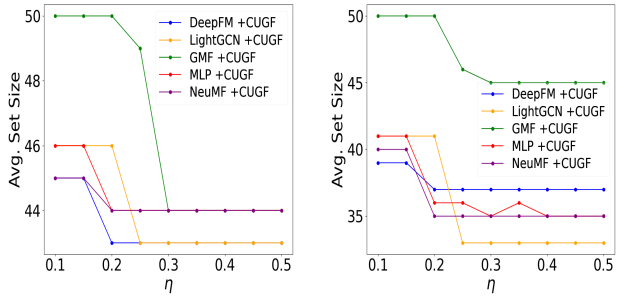
(a) $\eta = 0.15$ (b) $\eta = 0.20$

Figure 2: Performance analysis of the base models after applying the CUGF framework in terms of average set size with varying $\alpha = 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50$ on the **AmazonOffice** dataset grouped by **number of interactions & popular items interactions** under different **fairness threshold** η.



(a) $\alpha = 0.10$ (b) $\alpha = 0.20$

Figure 3: Performance analysis of the base models after applying the CUGF framework in terms of average set size with varying $\eta = 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50$ on the **MovieLens** dataset grouped by **number of interactions** under different **coverage parameter** α**.**



(a) $\alpha = 0.10$ (b) $\alpha = 0.20$

Figure 4: Performance analysis of the base models after applying the CUGF framework in terms of average set size with varying $\eta = 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50$ on the **MovieLens** dataset grouped by **number of interactions & popular items interactions** under different **coverage parameter** α.

in recommendation fairness).

To sum up, set predictors constructed by Algorithm 1 can modify any black-box recommendation models to output prediction sets for new customers that are strictly guaranteed to satisfy the desired coverage property as defined in Eq. (3) and fairness metric defined in Eq. (7). Moreover, when we set $\alpha = 1$, the CUGF framework degenerates into the state-of-the-art fairness-based recommendation method (Yao and Huang 2017; Ekstrand et al. 2018). When we set $\eta = 1$, the CUGF framework degenerates to the confidence-based recommendation models (Cui et al. 2024). When we set $\alpha = 1$ and $\eta = 1$, the CUGF framework degenerates to the base recommendation models (Rendle 2010; Gao et al. 2023).

Experiments

In this section, we conduct experiments to validate the effectiveness of the proposed framework (CUGF). We first evaluate the recommendation performance and the recommendation fairness under pre-specified coverage parameter α and fairness parameter η. Afterward, we conduct further analysis on the recommendation performance and fairness influenced by varying parameters (i.e., α and η).

Datasets

We generate four datasets via two different grouping methods on two publicly available datasets, i.e., AmazonOffice (eCommerce) (McAuley et al. 2015) and MovieLens (Movies)(Harper and Konstan 2015), to validate the proposed framework CUGF. We divide user groups into the advantaged group and the disadvantaged group. Specifically, we first apply the uniform strategy by assigning 50% of users to each group, and then dynamically adjust these groups to ensure that the minimum number of interactions of each user from the advantaged group is at least one more than the maximum number of interactions of any user from the disadvantaged group. We follow two ways for this adjustment:

Method	Group	AmazonOffice (no. of interactions)					AmazonOffice (no. of interactions + popular items interactions)				
		Loss (HR)	Loss (NDCG)	Avg. Set Size ↓	Diff. (HR) ↓	Diff. (NDCG) ↓	Loss (HR)	Loss (NDCG)	Avg. Set Size ↓	Diff. (HR) ↓	Diff. (NDCG) ↓
DeepFM + CUGF	1	0.170	0.595	38	0.008	**0.139**	0.166	0.594	37	0.079	**0.182**
	2	0.178	0.734				0.245	0.776			
LightGCN + CUGF	1	0.190	0.571	**36**	0.007	0.184	0.094	0.546	41	0.100	0.192
	2	0.196	0.756				0.195	0.736			
GMF + CUGF	1	0.001	0.509	50	**0.006**	0.197	0.001	0.504	50	0.010	0.194
	2	0.007	0.705				0.011	0.698			
MLP + CUGF	1	0.189	0.598	37	0.013	0.150	0.184	0.556	**31**	**0.008**	0.195
	2	0.202	0.747				0.192	0.751			
NeuMF + CUGF	1	0.193	0.595	38	0.012	0.143	0.196	0.578	35	0.012	0.184
	2	0.205	0.738				0.208	0.763			

Table 1: Recommendation performance comparisons in terms of average set size, and fairness comparisons in terms of Diff. (HR) and Diff. (NDCG) for five based models after applying our CUGF framework on **AmazonOffice** Dataset grouped by **no. of interactions** as well as **no. of interactions & popular items interactions** under $\alpha = 0.20$ and $\eta = 0.20$. Bold indicates the best result; underline indicates the second best.

Method	Group	MovieLens (no. of interactions)					MovieLens (no. of interactions + popular items interactions)				
		Loss (HR)	Loss (NDCG)	Avg. Set Size ↓	Diff. (HR) ↓	Diff. (NDCG) ↓	Loss (HR)	Loss (NDCG)	Avg. Set Size ↓	Diff. (HR) ↓	Diff. (NDCG) ↓
DeepFM + CUGF	1	0.034	0.407	38	0.031	**0.138**	0.036	0.389	37	**0.002**	**0.181**
	2	0.065	0.545				0.039	0.570			
LightGCN + CUGF	1	0.123	0.401	**36**	0.076	0.181	0.151	0.429	41	0.020	0.199
	2	0.200	0.582				0.171	0.628			
GMF + CUGF	1	0.053	0.369	50	**0.017**	0.200	0.036	0.364	50	0.035	0.198
	2	0.071	0.569				0.072	0.562			
MLP + CUGF	1	0.029	0.368	37	0.047	0.152	0.036	0.331	36	0.047	0.196
	2	0.076	0.520				0.083	0.527			
NeuMF + CUGF	1	0.037	0.370	37	0.030	0.149	0.024	0.337	**35**	0.021	0.183
	2	0.068	0.519				0.045	0.520			

Table 2: Recommendation performance comparisons in terms of average set size, and fairness comparisons in terms of Diff. (HR) and Diff. (NDCG) for five based models after applying our CUGF framework on **MovieLens** Dataset grouped by **no. of interactions** as well as **no. of interactions & popular items interactions** under $\alpha = 0.20$ and $\eta = 0.20$. Bold indicates the best result; underline indicates the second best.

one is based solely on the number of interactions (Li et al. 2021); the other considers the number of interactions and popular item interactions, where popular items are defined following Abdollahpouri et al. (2019). Due to the user-item interactions being limited in datasets, we follow (Chen et al. 2023) to implement negative sampling by selecting 50 non-interacted items for each user in the calibration, validation, and testing datasets, which ensures enough samples for empirical studies. Moreover, we split the held-out training data into the calibration data (60%), and testing data (40%).

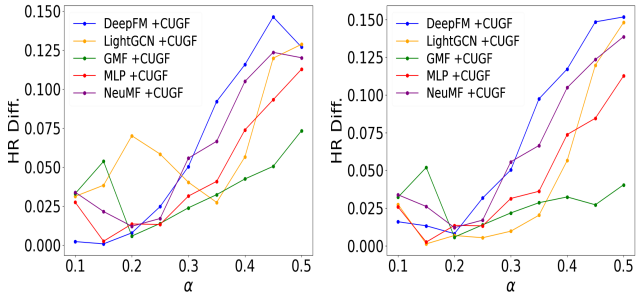
Base Models

We implement the proposed framework on top of the five base recommendation models, which are listed below.

- **DeepFM** (Guo et al. 2017): Integrates factorization machines and deep neural networks to learn feature interactions at different levels.
- **LightGCN** (He et al. 2020): Adapts Graph Convolutional Networks to recommendation systems, focusing on efficient learning.
- **GMF** (Koren, Bell, and Volinsky 2009): Utilizes linear interactions between user and item embeddings.
- **MLP** (Zhang et al. 2019): Employs deep learning to model complex user-item relationships.
- **NeuMF** (He et al. 2017): Combines Generalized Matrix Factorization and Multi-Layer Perceptron for capturing linear and non-linear interactions.

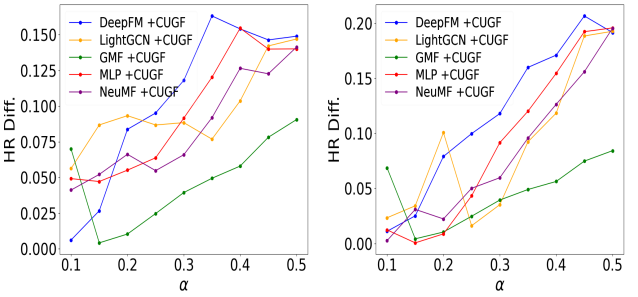
Implementation Details

All the methods are trained using the Adam optimizer. The batch size is set to 256, the learning rate to 0.001 and we train for 20 epochs. Detailed configurations for all base models can be found as follows: GMF uses an embedding size of 8; MLP employs layers of [64, 32, 16] with ReLU activation; NeuMF combines GMF and MLP with a GMF embedding size of 8 and MLP layers of [64, 32, 16], ReLU activation; DeepFM integrates 8 latent factors with deep layers of [50, 25, 10] with ReLU activation; LightGCN uses an embedding size of 8 and 3 layers with ReLU activation. When implementing the proposed framework to validate the performance, we set the coverage parameter $\alpha = 0.20$ via manual validation and the fairness threshold $\eta = 0.20$ via the bottom-line value that meet both fairness metrics (i.e., Difference (hit rate) and Difference (NDCG)). Code is publicly



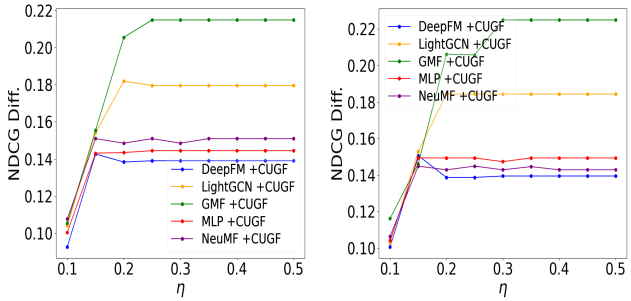
(a) $\eta = 0.15$ (b) $\eta = 0.20$

Figure 5: Fairness analysis of the base models after applying the CUGF framework in terms of **Diff. (HR)** with varying $\alpha = 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50$ on the **AmazonOffice** dataset grouped by **number of interactions** under different **fairness threshold** η.



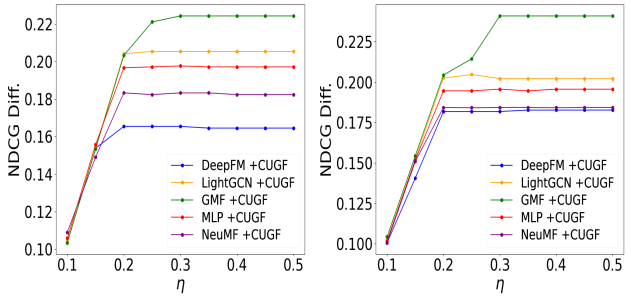
(a) $\eta = 0.15$ (b) $\eta = 0.20$

Figure 6: Fairness analysis of the base models after applying the CUGF framework in terms of **Diff. (HR)** with varying $\alpha = 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50$ on the **AmazonOffice** dataset grouped by **number of interactions & popular items interactions** under different **fairness threshold** η.



(a) $\alpha = 0.10$ (b) $\alpha = 0.20$

Figure 7: Fairness analysis of the base models after applying the CUGF framework in terms of **Diff. (NDCG)** with varying $\eta = 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50$ on the **MovieLens** dataset grouped by **number of interactions** under different **coverage parameter** α.



(a) $\alpha = 0.10$ (b) $\alpha = 0.20$

Figure 8: Fairness analysis of the base models after applying the CUGF framework in terms of **Diff. (NDCG)** with varying $\eta = 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50$ on the **MovieLens** dataset grouped by **number of interactions & popular items interactions** under different **coverage parameter** α.

available at https://github.com/kalpiree/CUGF-RS.

Performance and Fairness Evaluation

We implement the proposed CUGF framework on top of five base recommendation models and four datasets, i.e., AmazonOffice grouped by number of interactions, AmazonOffice grouped by number of interactions & popular items interactions, MovieLens grouped by number of interactions, and MovieLens grouped by number of interactions & popular items interactions, to evaluate the recommendation performance and fairness under pre-defined parameters α and η. Specifically, we evaluate the recommendation performance in terms of Average set size (Avg. Set Size) and evaluate the fairness in terms of Hit Rate Difference (Diff. (HR)) and NDCG Difference (Diff. (NDCG)), respectively. The results are shown in Tables 1 and 2. From these results, we can make the following observations:

- All base models, after applying the proposed CUGF framework, can produce valid prediction sets that meet both the coverage and fairness guarantee on all datasets.
- Specifically, LightGCN performs best w.r.t recommendation performance in terms of average set size on AmazonOffice Dataset grouped by number of interactions and MLP when grouped by number of interactions & popular items interactions respectively (Table 1). Similarly, LightGCN performs the best on MovieLens Datasets when users are grouped by number of interactions and NeuMF when grouped by the number of interactions & popular items interactions (Table 2).
- However, other models have larger set sizes compared with the above-mentioned models on the specified datasets. This is because the defined score function relies on the output of base recommendation models, and the models that can learn better relevance scores will also have better recommendation performance in our framework in terms of average set size.
- Meanwhile, all models meet the fairness threshold (i.e., 0.2) under both Diff. (HR) and Diff.(NDCG) metrics on all datasets. However, the best models in fairness under the two metrics vary a lot on different datasets. For

example, GMF performs the best on the AmazonOffice Dataset grouped by a number of interactions, while MLP performs the best on the AmazonOffice Dataset grouped by a number of interactions & popular items interactions under the Diff. (HR) metric. Besides, DeepFM performs the best on both grouped AmazonOffice Datasets under the Diff. (NDCG) metric. A similar phenomenon can also be observed in the remaining datasets. It is shown that fairness can be met in our framework on all base models, and different models have different fairness superiority on different datasets, which sheds light on the model choice in recommendation fairness.

- It can be concluded that the proposed CUGF framework is effective in generating prediction sets that meet the coverage guarantee while ensuring fairness of the recommended items for users across different groups for all recommendation models on all datasets. This also demonstrates that CUGF is data- and model-agnostic, which can be applied to any recommendation models and datasets.

Further Analysis

We empirically analyze the influence of the coverage parameter α and the fairness threshold η on the recommendation performance and fairness of the proposed framework CUGF.

Recommendation performance w.r.t varying α and η We select the AmazonOffice dataset grouped by number of interactions and grouped by number of interactions & popular items interactions to analyze the recommendation performance affected by varying coverage parameter α under specific fairness threshold η. Fig. 1 and Fig. 2 report the performance results with varying α from 0.10 to 0.50 (in increments of 0.05) under specific $\eta = 0.15, 0.20$ on the above two datasets in terms of average set size respectively.

We can observe that with increasing α, the average set size of all models after applying CUGF shows a decreasing trend under different fairness thresholds η in both figures. These results demonstrate that CUGF is an effective framework by producing valid prediction sets to accommodate to the pre-specified coverage α.

We select the MovieLens dataset grouped by number of interactions and grouped by number of interactions & popular items interactions to analyze the recommendation performance affected by varying fairness threshold η under specific coverage parameter α. Fig. 3 and Fig. 4 report the performance results with varying η varying from 0.10 to 0.50 (in increments of 0.05) under specific $\alpha = 0.10, 0.20$ on the above two datasets in terms of average set size respectively.

We can observe that with increasing η, the average set size of all models after applying CUGF decreases with a cascading effect in both figures. This is because increasing η means a loose constraint on fairness, when we put a loose constraint on fairness, the optimization of γ that dominates the prediction set will stop decreasing to a smaller value from time to time, which means CUGF stops adding new items to the prediction set, resulting in smaller set size in a cascading way. These results demonstrate that the proposed framework is effective in generating valid prediction sets by adapting to the user-pre-specified fairness threshold η.

Recommendation Fairness w.r.t varying α and η We select the AmazonOffice dataset grouped by the number of interactions and grouped by number of interactions & popular items interactions to analyze the fairness affected by varying coverage parameter α varying from 0.10 to 0.50 (in increments of 0.05) under specific fairness threshold $\eta = 0.15, 0.20$ in terms of Diff.(HR) in Fig. 5 and Fig. 6 respectively.

We can observe that with increasing α, the Diff. (HR) of all models after applying CUGF increases in both figures. This may be because relaxed constraints on coverage will result in more uncertainty to recommend items for users, which will lead to a larger extent of unfairness (e.g., larger Diff. (HR)), but still under the user-specified fairness threshold $\eta = 0.15, 0.20$ in according figures. These results demonstrate that the proposed framework is effective in guaranteeing and evaluating fairness by accommodating the user-specified coverage α.

We use the MovieLens dataset grouped by the number of interactions & popular items interactions to analyze the fairness affected by varying η varying from 0.10 to 0.50 (in increments of 0.05) under specific $\alpha = 0.10, 0.20$ in terms of Diff. (NDCG) in Fig. 7 and Fig. 8 respectively.

We can see that the Diff. (NDCG) of all models after applying CUGF generally remains steady after an initial and quick climb to a certain unfairness level with increasing η in both figures. These results demonstrate that the fairness of the proposed framework is not so sensitive to parameter η after a certain fairness level, which provides guidance on the parameter choice of fairness threshold for all models. For example, in Fig. 7(a) and Fig. 8(a), the fairness threshold can be set to 0.25 while set to 0.30 in Fig. 7(b) and Fig. 8(b).

It can be concluded that CUGF effectively guarantees both recommendation performance and fairness, which also sheds light on model and parameter choice when deploying the CUGF framework in real-world applications.

Conclusion

To address the credibility and fairness problems in RS, this paper introduces conformal prediction as a measure of uncertainty quantification for RS and proposes a reliable and fair recommendation framework called CUGF to generate optimal prediction sets on top of recommendation models that meet both the coverage/credibility and fairness guarantees. A novel optimization algorithm is designed to optimize the parameter that dominates the set and fairness based on a greedy strategy. Our empirical studies validate the effectiveness of the proposed framework with respect to the recommendation performance and fairness. This work bridges the gap between recommender systems' lack of credibility and unfairness as a whole and will surely inspire future research in trustworthy recommendation systems.

Acknowledgments

This work is partially supported by the Australian Research Council (ARC) Under Grants DP220103717 and LE220100078, and the National Natural Science Foundation of China under Grants No.62072257.

References

Abdollahpouri, H.; Mansoury, M.; Burke, R.; and Mobasher, B. 2019. The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286*.

Adomavicius, G.; and Tuzhilin, A. 2005. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6): 734–749.

Aggarwal, C. C. 2016. *Recommender Systems: The Textbook*. Springer Publishing Company, Incorporated, 1st edition. ISBN 3319296574.

Albora, G.; Rossi Mori, L.; and Zaccaria, A. 2023. Sapling Similarity: A performing and interpretable memory-based tool for recommendation. *Know.-Based Syst.*, 275.

Angelopoulos, A.; and Bates, S. 2023. *Conformal Prediction: A Gentle Introduction*. Now Publishers.

Angelopoulos, A. N.; and Bates, S. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Angelopoulos, A. N.; Bates, S.; Jordan, M.; and Malik, J. 2021. Uncertainty Sets for Image Classifiers using Conformal Prediction. In *International Conference on Learning Representations*.

Chang, S.; Zhang, Y.; Tang, J.; Yin, D.; Chang, Y.; Hasegawa-Johnson, M. A.; and Huang, T. S. 2017. Streaming Recommender Systems. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, 381–389. International World Wide Web Conferences Steering Committee.

Chen, C.; Ma, W.; Zhang, M.; Wang, C.; Liu, Y.; and Ma, S. 2023. Revisiting Negative Sampling vs. Non-sampling in Implicit Recommendation. *ACM Trans. Inf. Syst.*, 41(1).

Cui, F.; Yu, S.; Chai, Y.; Qian, Y.; Jiang, Y.; Liu, Y.; Liu, X.; and Li, J. 2024. A Bayesian Deep Recommender System for Uncertainty-Aware Online Physician Recommendation. *Information & Management*, 104027.

Ekstrand, M. D.; Tian, M.; Kazi, M. R. I.; Mehrpouyan, H.; and Kluver, D. 2018. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, 242–250. New York, NY, USA: Association for Computing Machinery.

Fisch, A.; Schuster, T.; Jaakkola, T. S.; and Barzilay, R. 2021. Efficient Conformal Prediction via Cascaded Inference with Expanded Admission. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Fu, Z.; Xian, Y.; Gao, R.; Zhao, J.; Huang, Q.; Ge, Y.; Xu, S.; Geng, S.; Shah, C.; Zhang, Y.; and de Melo, G. 2020. Fairness-Aware Explainable Recommendation over Knowledge Graphs. arXiv:2006.02046.

Gao, C.; Zheng, Y.; Li, N.; Li, Y.; Qin, Y.; Piao, J.; Quan, Y.; Chang, J.; Jin, D.; He, X.; et al. 2023. A survey of graph neural networks for recommender systems: Challenges, methods, and directions. *ACM Transactions on Recommender Systems*, 1(1): 1–51.

Geyik, S. C.; Ambler, S.; and Kenthapadi, K. 2019. 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. ACM.

Guo, H.; Tang, R.; Ye, Y.; Li, Z.; and He, X. 2017. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. arXiv:1703.04247.

Han, S. C.; Lim, T.; Long, S.; Burgstaller, B.; and Poon, J. 2021. GLocal-K: Global and Local Kernels for Recommender Systems. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21. ACM.

Harper, F. M.; and Konstan, J. A. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.*, 5(4).

He, J.; and Chu, W. 2010. *A social network-based recommender system (SNRS)*, volume 12, 47–74. Data Mining for Social Network Data.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, 639–648. Association for Computing Machinery.

He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; and Chua, T.-S. 2017. Neural Collaborative Filtering. arXiv:1708.05031.

Islam, R.; Keya, K. N.; Zeng, Z.; Pan, S.; and Foulds, J. 2021. Debiasing Career Recommendations with Neural Fair Collaborative Filtering. In *Proceedings of the Web Conference 2021*, WWW '21. Association for Computing Machinery.

Jannach, D.; Zanker, M.; Felfernig, A.; and Friedrich, G. 2010. *Recommender systems: an introduction*. Cambridge University Press.

Kim, D.; and Suh, B. 2019. Enhancing VAEs for collaborative filtering: flexible priors & gating mechanisms. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, 403–407. New York, NY, USA: Association for Computing Machinery.

Knyazev, N.; and Oosterhuis, H. 2023. A Lightweight Method for Modeling Confidence in Recommendations with Learned Beta Distributions. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23. ACM.

Ko, H.; Lee, S.; Park, Y.; and Choi, A. 2022. A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1): 141.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8): 30–37.

KWEON, W.; Kang, S.; Jang, S.; and Yu, H. 2024. Top-Personalized-K Recommendation. In *The Web Conference 2024*.

Li, Y.; Chen, H.; Fu, Z.; Ge, Y.; and Zhang, Y. 2021. User-oriented fairness in recommendation. In *Proceedings of the web conference 2021*, 624–632.

Liu, W.; Tsang, I. W.; and Muller, K. R. 2017. An Easy-to-hard Learning Paradigm for Multiple Classes and Multiple Labels. *Journal of Machine Learning Research*, 18: 94:1–94:38.

Lu, J.; Wu, D.; Mao, M.; Wang, W.; and Zhang, G. 2015. Recommender system application developments: a survey. *Decision support systems*, 74: 12–32.

Mao, Y.; Wang, Z.; Liu, W.; Lin, X.; and Xie, P. 2022. MetaWeighting: Learning to Weight Tasks in Multi-Task Learning. In *ACL*, 3436–3448. Association for Computational Linguistics.

Mao, Y.; Yun, S.; Liu, W.; and Du, B. 2020. Tchebycheff Procedure for Multi-task Text Classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 4217–4226.

McAuley, J.; Targett, C.; Shi, Q.; and van den Hengel, A. 2015. Image-based Recommendations on Styles and Substitutes. arXiv:1506.04757.

Naghiaei, M.; Rahmani, H. A.; Aliannejadi, M.; and Sonboli, N. 2022. Towards Confidence-aware Calibrated Recommendation. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*.

Quach, V.; Fisch, A.; Schuster, T.; Yala, A.; Sohn, J. H.; Jaakkola, T. S.; and Barzilay, R. 2023. Conformal language modeling. *arXiv preprint arXiv:2306.10193*.

Rahmani, H. A.; Naghiaei, M.; Dehghan, M.; and Aliannejadi, M. 2022. Experiments on Generalizability of User-Oriented Fairness in Recommender Systems. arXiv:2205.08289.

Rendle, S. 2010. Factorization Machines. In *2010 IEEE International Conference on Data Mining*, 995–1000.

Ricci, F.; Rokach, L.; and Shapira, B. 2021. Recommender systems: Techniques, applications, and challenges. *Recommender systems handbook*, 1–35.

Ricci, F.; Rokach, L.; Shapira, B.; and Kantor, P. B. 2010. *Recommender Systems Handbook*. Berlin, Heidelberg: Springer-Verlag.

Romano, Y.; Patterson, E.; and Candes, E. 2019. Conformalized quantile regression. *Advances in neural information processing systems*, 32.

Romano, Y.; Sesia, M.; and Candes, E. 2020. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33: 3581–3591.

Schafer, B.; Konstan, J.; and Riedl, J. 1999. Recommender Systems in E-Commerce. *1st ACM Conference on Electronic Commerce, Denver, Colorado, United States*.

Stratigi, M.; Kondylakis, H.; and Stefanidis, K. 2017. Fairness in Group Recommendations in the Health Domain. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic Learning in a Random World*. Berlin, Heidelberg: Springer-Verlag. ISBN 0387001522.

Yao, S.; and Huang, B. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. arXiv:1705.08804.

Zhang, S.; Yao, L.; Sun, A.; and Tay, Y. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1): 1–38.

Zhu, Z.; Wang, J.; Zhang, Y.; and Caverlee, J. 2018. Fairness-aware recommendation of information curators. *arXiv preprint arXiv:1809.03040*.