





# Quantifying Fairness Disparities in Graph-Based Neural Network Recommender Systems for Protected Groups

Nikzad Chizari<sup>1</sup><sup>a</sup>, Keywan Tajfar<sup>2</sup><sup>b</sup>, Niloufar Shoeibi<sup>1</sup><sup>c</sup> and María N. Moreno-García<sup>1</sup><sup>d</sup>

<sup>1</sup>Department of Computer Science and Automation, Science Faculty, University of Salamanca,  
Plaza de los Caídos sn, 37008 Salamanca, Spain

<sup>2</sup>College of Science, School of Mathematics, Statistics, and Computer Science, Department of Statistics,  
University of Tehran, Tehran, Iran

**Keywords:** Recommender Systems, Bias, Fairness, Graph-Based Neural Networks.

**Abstract:** The wide acceptance of Recommender Systems (RS) among users for product and service suggestions has led to the proposal of multiple recommendation methods that have contributed to solving the problems presented by these systems. However, the focus on bias problems is much more limited. Some of the most successful and recent methods, such as Graph Neural Networks (GNNs), present problems of bias amplification and unfairness that need to be detected, measured, and addressed. In this study, an analysis of RS fairness is conducted, focusing on measuring unfairness toward protected groups, including gender and age. We quantify fairness disparities within these groups and evaluate recommendation quality for item lists using a metric based on Normalized Discounted Cumulative Gain (NDCG). Most bias assessment metrics in the literature are only valid for the rating prediction approach, but RS usually provide recommendations in the form of item lists. The metric for lists enhances the understanding of fairness dynamics in GNN-based RS, providing a more comprehensive perspective on the quality and equity of recommendations among different user groups.


## 1 INTRODUCTION


The abundance of information poses a challenge for users to find products that align with their preferences, and to address this, Recommender Systems (RS) have proven to be essential tools. These systems are now widely integrated into diverse applications like E-commerce platforms, entertainment platforms, social networks, and lifestyle apps (Ricci et al., 2022; Zheng and Wang, 2022; Pérez-Marcos et al., 2020; Lin et al., 2022; Chen et al., 2020). RS cannot only help lessen the problem of information overload but also lead to personalization based on users' interests (Rajeswari and Hariharan, 2016).


A great amount of research work in this area has been dedicated to enhancing the performance of RS and addressing their issues, among which bias mitigation is one of the most recent. Two of the most critical issues for RS are bias and fairness, which can lead to discrimination. A systematic and persistent departure


from a true value or an accurate portrayal of reality is referred to as bias, which occurs when a variety of elements that affect the decision-making or judgment process are present. Biases often come from underlying imbalances and inequalities in data, resulting in biased recommendations that can influence in users' choices of consumption (Boratto and Marras, 2021; Misztal-Radecka and Indurkha, 2021). Also, algorithm design can result in bias and discrimination in automated decisions (Misztal-Radecka and Indurkha, 2021; Gao et al., 2022b).

The widespread use of artificial intelligence and machine learning techniques in society has resulted in undesirable effects due to biased models, including economic, legal, ethical, and security issues that can harm companies (Di Noia et al., 2022; Fahse et al., 2021; Kordzadeh and Ghasemaghaei, 2022; Boratto et al., 2021; Boratto and Marras, 2021; Wang et al., 2023). Moreover, users may be dissatisfied with biased recommendations, further exacerbating the problem (Gao et al., 2022a). In addition, mitigation of bias is a concern of international organizations whose regulations include obligations related to this issue, especially in sensitive areas. (Di Noia et al., 2022).

<sup>a</sup> <https://orcid.org/0000-0002-7300-6126>

<sup>b</sup> <https://orcid.org/0000-0001-7624-5328>

<sup>c</sup> <https://orcid.org/0000-0003-4171-1653>

<sup>d</sup> <https://orcid.org/0000-0003-2809-3707>

The effects of decision making based on biased models can also be ethical and lead to decisions that discriminate against minority or marginalized groups.

Recent advances in deep learning, including Graph Neural Networks (GNNs), have improved performance of RS and addressed challenges, even with sparse data (Mu, 2018; Yu et al., 2023). GNNs excel at capturing relationships in graph data through message passing (Zhou et al., 2020) and have gained popularity for various graph-related tasks (Dong et al., 2022b; Zhang et al., 2021; Wu et al., 2020b). However, they raise concerns about bias and fairness, potentially discriminating against demographic subgroups defined by sensitive attributes like age, gender, or race. Addressing biases in GNNs remains relatively unexplored (Dong et al., 2022b; Dai and Wang, 2021; Dong et al., 2022a; Chen et al., 2022; Xu et al., 2021; Zeng et al., 2021; Chizari et al., 2022).

In RS, user-item interactions can be viewed as graphs, with the potential for improvement through additional data like social dynamics or context. While neural network-based methods, especially deep learning, have gained traction in RS, they excel at capturing complex user-item relationships. However, they are limited to Euclidean data, struggling with intricate high-order structures (Zhou et al., 2020; Gao et al., 2022b). Recent advancements in Graph Neural Networks (GNNs) have addressed these limitations by extending deep learning’s capabilities to handle non-Euclidean complexities (Bronstein et al., 2017; Li, 2023).

Several investigations have underscored the influence of graph structures and the underlying message-passing mechanisms within GNNs, shedding light on their propensity to accentuate both fairness concerns and broader social biases (Chizari et al., 2022; Dai and Wang, 2021; Chizari et al., 2023). Notably, within the landscape of social networks featuring graph architectures, nodes characterized by analogous sensitive attributes often exhibit a predilection for establishing connections with one another, distinguishing them from nodes marked by disparate attributes. This observable phenomenon engenders an environment wherein nodes sharing comparable sensitive traits become recipients of akin representations stemming from the amalgamation of neighboring features within the GNN framework. Conversely, nodes endowed with distinct sensitive attributes garner divergent representations. The ramifications of this dynamic are palpable, as it introduces a discernible bias into the decision-making trajectory (Dai and Wang, 2021).

Sensitive attributes in data, encompassing characteristics like race, gender, sexual orientation, reli-

gion, age, and disability status, are considered private and protected by privacy laws due to the potential for discrimination and harm (Oneto and Chiappa, 2020). Discrimination concerns socially significant categories associated with these attributes, legally protected in the United States (Barocas et al., 2017). Recognizing these sensitive attributes is essential in RS to ensure fairness and prevent biased recommendations that may be viewed as discriminatory under European or US laws.

In this study, the aim is to measure group unfairness and subgroup unfairness with sensitive attributes. We focus on the evaluation of item recommendation lists since there is hardly any work in the literature aimed at this type of output of RS, but most of it is focused on the rating prediction approach.

## 2 STATE OF THE ART

In this section, we present a comprehensive overview of prior research endeavors. This segment delves into the realm of bias and fairness challenges, exploring various evaluation approaches. The survey encompasses multiple layers, ranging from machine learning (ML) to GNN-based RS. We direct particular attention toward GNN-based RS models and the diverse array of fairness evaluation metrics employed in this context with respect to sensitive groups.

### 2.1 Bias and Fairness in Machine Learning (ML)

Machine learning (ML) models, which are trained on human-generated data, can inherit biases present in the data (Alelyani, 2021; Zeng et al., 2021). These biases can emerge due to various factors during data collection and sampling (Bruce et al., 2020). Unfortunately, such biases can persist in ML models, leading to unfair decisions and suboptimal outcomes (Fahse et al., 2021; Gao et al., 2022b; Mehrabi et al., 2021). The ML models themselves can even exacerbate these biases, impacting decision-making processes (Bernhardt et al., 2022). It’s evident that bias can manifest throughout the ML lifecycle, spanning data collection, pre-processing, algorithm design, model training, and result interpretation (Alelyani, 2021; Zeng et al., 2021). These biases can also originate externally from societal inequalities and discrimination (Bruce et al., 2020).

Numerous research endeavors focus on identifying and assessing biases and unfairness, especially concerning protected groups like gender, age, and

race. Various metrics rooted in statistical parameters are employed in these studies. Recent experiments emphasize the need to understand bias origins in specific contexts, pinpoint problems, and conduct accurate evaluations, providing a foundation for bias mitigation techniques (Caton and Haas, 2020; Verma and Rubin, 2018; Feldman et al., 2015; Hardt et al., 2016; Alelyani, 2021). Metrics can be categorized as individual-level or group-level \cite{caton2020fairness}. Individual-level metrics assess treatment equality for individuals with similar attributes, while group-level metrics evaluate disparate treatment among various groups.

## 2.2 Bias and Fairness in GNN-Based Models

Graph Neural Network (GNN)-based models have recently garnered attention due to their strong performance and applicability in various graph learning tasks (Dong et al., 2022b; Zhang et al., 2021; Wu et al., 2020b). However, despite their achievements, these algorithms are not immune to bias and fairness challenges. GNNs can inadvertently exhibit bias towards specific demographic subgroups defined by sensitive attributes such as age, gender, and race. Furthermore, research efforts towards understanding and measuring biases in GNNs have been relatively limited (Dong et al., 2022b; Dai and Wang, 2021; Dong et al., 2022a; Chen et al., 2022; Xu et al., 2021; Zeng et al., 2021).

Bias challenges within GNN algorithms stem from various factors, including biases embedded in the input network structure. While the message-passing mechanism is commonly associated with exacerbating bias, other aspects of the GNN's network structure are also influential. Understanding how structural biases manifest as biased predictions present challenges due to gaps in comprehension, such as the Fairness Notion Gap, Usability Gap, and Faithfulness Gap elucidated in (Dong et al., 2022b). The Fairness Notion Gap concerns instance-level bias evaluation, the Usability Gap pertains to fairness influenced by computational graph edges and their contributions, and the Faithfulness Gap addresses ensuring accurate bias explanations. The work in (Dong et al., 2022b) addresses these gaps by introducing a bias evaluation metric for node predictions and an explanatory framework. This metric quantifies node contributions to the divergence between output distributions of sensitive node subgroups based on attributes. While the literature explores various strategies to mitigate biases in GNN-based models, focused research on this aspect remains relatively limited.

In this realm, some studies including (Dai and Wang, 2021) and (Li et al., 2021) aim to combat discrimination and enhance fairness in GNNs with consideration to sensitive attribute information. In (Dai and Wang, 2021), a method is introduced that reduces bias while maintaining high accuracy in node classification. On the other hand, (Li et al., 2021) presents an approach for learning a fair adjacency matrix with strong graph structural constraints, aiming to achieve fair link prediction while minimizing the impact on accuracy. Additionally, (Loveland et al., 2022) proposes two model-agnostic algorithms for edge editing, leveraging gradient information from a fairness loss to identify edges that promote fairness enhancements.

## 2.3 Bias and Fairness in RS

Bias and fairness challenges within the environment of RS, encompass varied interpretations and can be categorized into distinct groups. Viewing this from a broader perspective, bias can be segmented into three classes akin to the divisions outlined for Machine Learning (ML) in (Mehrabi et al., 2021). These classes encompass bias in input data, signifying the data collection phase involving users; algorithmic bias in the model, manifesting during the learning phase of recommendation models based on the collected data; and bias in results, which impact subsequent user decisions and actions (Chen et al., 2020; Baeza-Yates, 2016). Expanding upon the intricacies of bias, these three classes can be further broken down into sub-classes, creating an interconnected circular framework.

Bias in data, stemming from disparities in test and training data distribution, manifests in various forms including selection bias, exposure bias, conformity bias, and position bias. Selection bias occurs when skewed rating distributions inadequately represent the entire rating spectrum. Exposure bias arises from users predominantly encountering specific items, leading to unobserved interactions that may not reflect their true preferences. Conformity bias emerges when users mimic the behavior of others due to skewed interaction labels. Position bias is seen when users favor items in higher positions over genuinely relevant ones (Chen et al., 2020; Sun et al., 2019). Algorithmic bias can occur throughout model creation, data pre-processing, training, and evaluation stages. Inductive bias, a constructive element, enhances model generalization by making assumptions that improve learning from training data and informed decision-making on unseen test data (Chen et al., 2020). Outcomes of bias fall into two

categories: popularity bias and unfairness. Popularity bias results from the long-tail effect in ratings, where a few popular items dominate user interactions, potentially leading to elevated scores for them at the expense of less popular items (Ahanger et al., 2022; Chen et al., 2020). Unfairness arises from systematic discrimination against specific groups (Chen et al., 2020).

These various forms of biases collectively contribute to a circular pattern, wherein biases in the data are propagated to the models, subsequently influencing the outcomes. This cycle is completed as biases from the outcomes find their way back to the data. Throughout each of these stages, new biases can be introduced, thus perpetuating the cycle (Fabri et al., 2022; Chen et al., 2020). This cyclical behavior adds complexity to the task of identifying and addressing biases, further emphasizing the challenge of bias recognition and mitigation (Mansoury et al., 2021; Chen et al., 2020).

Conversational Recommender Systems (CRS) are investigated in (Lin et al., 2022) to explore popularity bias systematically, introducing metrics from different angles such as exposure, success, and conversational utility. Similarly, (Abdollahpouri et al., 2019) addresses popularity bias and long-tail distribution in RS, proposing metrics like Average Recommendation Popularity (ARP), Average Percentage of Long Tail Items (APLT), and Average Coverage of Long Tail items (ACLT). However, the focus extends beyond popularity bias to the concern of unfairness towards protected groups due to biased recommendations. In the RS field, fairness has gained significance, being recognized as a resource allocation tool that shapes information exposure for users (Wang et al., 2021). This concept of fairness is categorized into process fairness (relating to the recommendation model) and output fairness (influencing users' information experiences).

- Process fairness pertains to equitable allocation within the models, features (e.g., race, gender), and learned representations.
- Outcome fairness, known as distributive justice, ensures fairness in recommendation results (Wang et al., 2021). Outcome fairness comprises two sub-categories: Grouped by Target and Grouped by Concept.
- Grouped by Target includes group-level and individual-level fairness. Group-level fairness involves fair outcomes across different groups, while individual-level fairness ensures fairness at the individual level (Wang et al., 2021).
- Grouped by Concept consists of multiple categorizations:

rizations:

- Consistent Fairness at the individual level emphasizes uniform treatment for similar individuals.
- Consistent Fairness at the group level strives for equitable treatment across different groups.
- Calibrated fairness, or merit-based fairness, relates an individual's merit to the outcome value.
- Counterfactual fairness mandates identical outcomes in both real and counterfactual scenarios.
- Envy-free fairness prevents individuals from envying others' outcomes.
- Rawlsian maximin fairness maximizes results for the weakest individual or group.
- Maximin-shared Fairness ensures outcomes surpass each individual's (or group's) maximin share (Wang et al., 2021).

The correlation between bias and fairness is very important. An in-depth examination is carried out in (Boratto et al., 2022) to address methods for alleviating consumer unfairness in the context of rating prediction using real-world datasets (LastFM and Movielens). The study entails a three-fold analysis. Firstly, the influence of bias mitigation on model accuracy, measured through metrics like NDCG/RMSE, is evaluated. Secondly, the impact of bias mitigation on unfairness is assessed. Lastly, the study explores whether disparate impact invariably harms minority groups, as Demographic Parity (DP) indicates. This investigation underscores the complexities involved in this domain and proposes potential solutions and optimization strategies. The selection of appropriate metrics for conducting such evaluations is also deemed crucial. This comprehensive study holds substantial relevance in the field.

## 2.4 Bias and Fairness in GNN-Based RS

The adoption of GNN-based RS has shown promise in enhancing result accuracy, as noted in previous studies (Steck et al., 2021; Khan et al., 2021; Mu, 2018). However, this improved performance often comes at the cost of introducing bias and fairness issues (Chizari et al., 2023; Dai and Wang, 2021). The inherent graph structure and the message-passing mechanism within GNNs can exacerbate bias problems, leading to inequitable outcomes. Furthermore, many RS applications are situated within social network contexts, where graph structures are prevalent. In such systems, nodes sharing similar sensitive attributes tend to establish connections with one another, distinguishing them from nodes with differing

sensitive attributes (e.g., the formation of connections among young individuals in social networks). This phenomenon creates an environment where nodes of comparable sensitive features receive akin representations through the aggregation of neighbor features within GNNs, while nodes with distinct sensitive features receive disparate representations. This dynamic results in a pronounced bias issue influencing decision-making processes (Dai and Wang, 2021).

In GNN-based RS, specific sensitive attributes can exacerbate existing biases within the network, prompting the need to quantify fairness in these contexts. To tackle this issue, relevant metrics should consider the distribution of positive classifications across distinct groups defined by various values of the sensitive attribute (Rahman et al., 2019; Wu et al., 2020a).

Recent research in GNN-based RS has addressed fairness issues and sensitive attributes. For example, (Rahman et al., 2019) focuses on quantifying and rectifying fairness problems, particularly group fairness and disparate impact, in graph embeddings. It introduces a concept called "equality of representation" to assess fairness in friendship-based RS. These methods are applied to real-world datasets, leading to the development of a fairness-aware graph embedding algorithm that effectively mitigates bias and improves key metrics.

Study of (Wu et al., 2021), the aim is to make fair recommendations by filtering sensitive information from representation learning. They use user and item embeddings, sensitive features, and a graph-based adversarial training process. Fairness is assessed with metrics like AUC for binary attributes and micro-averaged F1 for multivalued attributes, considering gender attribute imbalance. The model is tested on Lastfm-360K and MovieLens datasets.

In summary, despite the absence of sensitive features being a significant challenge in GNN-based RS, most research in this domain has focused on tackling discrimination against minorities or addressing information leakage issues. These types of unfairness and discrimination run contrary to existing regulations and anti-discriminatory laws. Additionally, understating the behavior of certain algorithms against bias and fairness is absolutely important. To do so, it is essential to use appropriate metrics that fit the domain and models to have more reliable results.

### 3 METHODOLOGY

In the present section, the methodology used in this research is explained along with information regard-

ing benchmark datasets and used metrics.

The primary objective of this research is to assess and quantify the degree of unfairness experienced by specific protected groups, namely gender and age, with a high degree of accuracy. In order to achieve this goal, the study focuses on the quantification of fairness disparities. These disparities serve as metrics to evaluate the quality and fairness of recommended items for these particular groups. In essence, the research aims to provide a robust and comprehensive assessment of the biases and inequities present in RS concerning gender and age attributes.

In addition, this study employs the NDCG (Normalized Discounted Cumulative Gain) evaluation metric as a specific measurement for assessing the recommendation quality within each of the protected groups. NDCG is a widely recognized metric used in RS to evaluate lists of recommended items.

The NDCG metric offers a more in-depth evaluation of recommendation quality by considering the position and relevance of items within recommendation lists. It takes into account both the order and importance of recommended items, making it particularly suitable for measuring the quality of recommendations in this context (Chia et al., 2022).

By utilizing NDCG as a specific evaluation metric for protected groups, this research aims to provide a comprehensive assessment of recommendation quality while ensuring fairness and equity for all users, regardless of their gender or age (age was discretized into two intervals, lower and higher 30 years old). This approach allows for a more nuanced understanding of the performance of recommender systems and their impact on different demographic groups.

#### 3.1 Benchmark Datasets

In this experiment, three real-world datasets are used to reach more accurate generalization. These datasets are well-known in the RS field and include certain characteristics that match bias and fairness assessment. Their selection was influenced by the inclusion of the specific sensitive attributes and biases being investigated. Therefore, the chosen datasets include users' gender and age as sensitive attributes, along with an uneven distribution of instances across various attribute values. The three real-world datasets used in this study are MovieLens 100K, LastFM 100K, and Book Recommendation. Detailed descriptions and Exploratory Data Analysis (EDA) for them are provided below.

- **MovieLens 100K.** MovieLens (gro, 2021) is a well-established resource frequently used in research within the field of RS. MovieLens is a non-

commercial online movie recommendation platform, and its dataset has been incrementally collected through random sampling from the website. This dataset comprises user ratings for movies, quantified on a star scale within the range of 1 to 5. Additionally, this dataset encompasses user information, including "Gender" and "Age" attributes, which have been identified as sensitive features according to capAI guidelines (Floridi et al., 2022).

- **LastFM 100K.** The LastFM dataset (Celma, 2010) is a widely recognized resource in the field of RS, particularly for music recommendations. This dataset encompasses user and artist information drawn from various regions around the world. Rather than utilizing a conventional rating system, this dataset quantifies user interactions based on the number of times each user has listened to individual artists, denoted as "weight." For the purposes of this research, we have utilized a pre-processed subset of the LastFM 360K dataset, which is well-suited for RS implementation. Within this subset, we have specifically chosen 100,000 interactions to form the basis of our study. In accordance with capAI guidelines (Floridi et al., 2022), gender and age are identified as sensitive attributes within this dataset. Notably, the dataset represents the frequency with which users have listened to specific music, which has been normalized to a scale ranging from 1 to 5 to enhance precision in the analysis.
- **Book Recommendation 100K.** The dataset used in a study by (Mobius, 2020) encompasses user ratings for a diverse array of books. For the purpose of our experiment, we have selected a representative 100,000-sample subset of this dataset. It's worth noting that this sample faithfully mirrors the distribution characteristics of the original dataset.

### 3.2 Recommendation Approaches

In this experiment, various types of models are utilized to achieve a better range of results, hence, providing superior comparison. Three distinct recommendation approaches are used in this research including Collaborative Filtering (CF), Matrix Factorization (MF), and GNN-based approaches. The goal was to choose the most representative algorithms within each category for comprehensive analysis. This diverse selection of methods allows us to expedite the evaluation of bias and fairness. In the upcoming section, we will provide an overview of the

methodologies corresponding to each approach utilized in this study.

### 3.3 Evaluation Metrics

In this section, the description and the categorization of used metrics are shown. In order to have a comprehensive understanding of model performance and bias and fairness aspects, two different types of metrics are used, for the assessment of reliability, as well as for bias and unfairness. As mentioned above, we have focused on the evaluation of item recommendation lists by means of rank metrics. In this context, various values of K have been employed to determine the top-K ranked items within the list, with K representing the list's size.

#### 3.3.1 Model Evaluation Metrics

The results presented complement the studies previously carried out (Chizari et al., 2023; Chizari et al., 2022) where various types of well-known performance metrics were used. These are Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (NDCG), Precision, Recall and item Hit Ratio (HR). This work has focused on the evaluation of NDCG for protected and unprotected groups for both the age and gender attributes.

#### 3.3.2 Bias and Fairness Metrics

In addition to the above assessment, we will delve into several bias and fairness evaluation metrics, with a particular emphasis on user-centric fairness measures. We have previously studied and provided a detailed exposition of the following bias and unfairness metrics (Chizari et al., 2023; Chizari et al., 2022):

- Average Popularity (Naghiaei et al., 2022)
- Gini Index (Sun et al., 2019; Lazovich et al., 2022)
- Item Coverage (Wang and Wang, 2022)
- Differential Fairness (DF) for sensitive attribute gender (Islam et al., 2021; Foulds et al., 2019)
- Value Unfairness (Aalam et al., 2022; Yao and Huang, 2017; Farnadi et al., 2018)
- Absolute Unfairness (Yao and Huang, 2017; Farnadi et al., 2018)

In this experimental analysis, with the aim of gaining a thorough insight into how models behave in terms of bias and fairness, a new metric of relative difference between groups was proposed. This metric is also implemented on list of top-K ranked items.

### 3.3.3 Proposed Metric

This research includes another metric proposed in order to measure the accuracy of recommendation for each protected and unprotected group (gender and age). To achieve this, initially,  $NDCG@k$ , which is designed to measure the effectiveness of a recommendation system by assessing the relevance and ranking of recommended items, is computed separately for mentioned groups. Subsequently, the relative difference between the  $NDCG@5k$  values for these two groups is calculated to assess their proximity or disparity. This is achieved by subtracting the  $NDCG@k$  for group 2 (e.g., males) from that of group 1 (e.g., females), then dividing the result by the average of the two values, and finally multiplying the outcome by 100 as it can be seen in below:

$$\frac{|NDCG@k_{group1} - NDCG@k_{group2}|}{((NDCG@k_{group1} + NDCG@k_{group2})/2)} * 100 \quad (1)$$

This particular metric serves as a dedicated and insightful tool for evaluating fairness among protected groups in the context of RS. It offers a unique perspective on fairness by focusing on how recommendations perform within these specific groups. Furthermore, it complements other fairness metrics used in the evaluation process, providing a more comprehensive and robust understanding of fairness outcomes. By comparing the results obtained from this metric with those derived from other fairness metrics, the research gains additional validation and a deeper insight into the fairness dynamics within the recommendation system. This approach enhances the credibility and completeness of the fairness assessment, ultimately contributing to a more thorough and meaningful analysis.

## 4 EXPERIMENTAL SETUP

### 4.1 Hardware Used

The research was carried out on a high-performance system featuring a Ryzen 7 5800H CPU, which offers 8 cores and 16 threads, operating at base and turbo frequencies of 3.2 GHz and 4.4 GHz, respectively. This AMD processor, based on the Zen 3 architecture, provided the computational power needed for our tasks. Additionally, the system was equipped with an RTX 3060 Mobile GPU, known for its 6GB VRAM, 3840 CUDA cores, and 120 Tensor Cores. This GPU, part of NVIDIA's Ampere architecture, proved essential for tasks such as machine learning model training. The system boasted a total of 16GB DDR4 RAM,

with approximately 15GB available for research purposes, ensuring efficient execution of complex computations and data handling.

### 4.2 Software and Libraries Used

Python, with CPython as the core interpreter, served as the primary programming language. The research was based on Recbole, an open-source library, and a modified fork named Recbole-FairRec. Custom metrics and models were integrated into this library, resulting in Recbole-Optimized. Key Python libraries included TensorFlow and PyTorch for machine learning and deep learning, NumPy for numerical computing, Pandas for data manipulation, and Scikit-learn for various machine learning tasks. Additional industry-standard libraries were used as needed for specific research requirements.

## 5 RESULTS

In the following sections, we delve into the results of our investigation into recommendation system fairness, with a specific focus on evaluating biases and unfairness concerning protected groups, including gender and age. Our study uses various evaluation metrics, including model evaluation metrics, bias and fairness metrics, and NDCG for quality of recommendation. Through experimentation, the aim is to shed light on the effectiveness of fairness-aware recommendation models and their impact on recommendation quality for different demographic segments.

### 5.1 Model Evaluation Results

In this section, a comprehensive and insightful comparative analysis of our evaluation results.

First, the different groups of sensitive attributes studied were evaluated separately using the NDCG metric for recommendation lists. The following figures 1 and 2 show the results obtained for the three datasets described in section 3.1 and for the eight recommendation methods tested.

### 5.2 Bias and Fairness Results

The results of the proposed metric that evaluates the relative difference on the performance between groups is provided secondly. Figures 3 and 4 show these results for three datasets on the previously mentioned models.

In Figure 1 NDCG performance for sensitive attribute gender is provided on two datasets. Higher

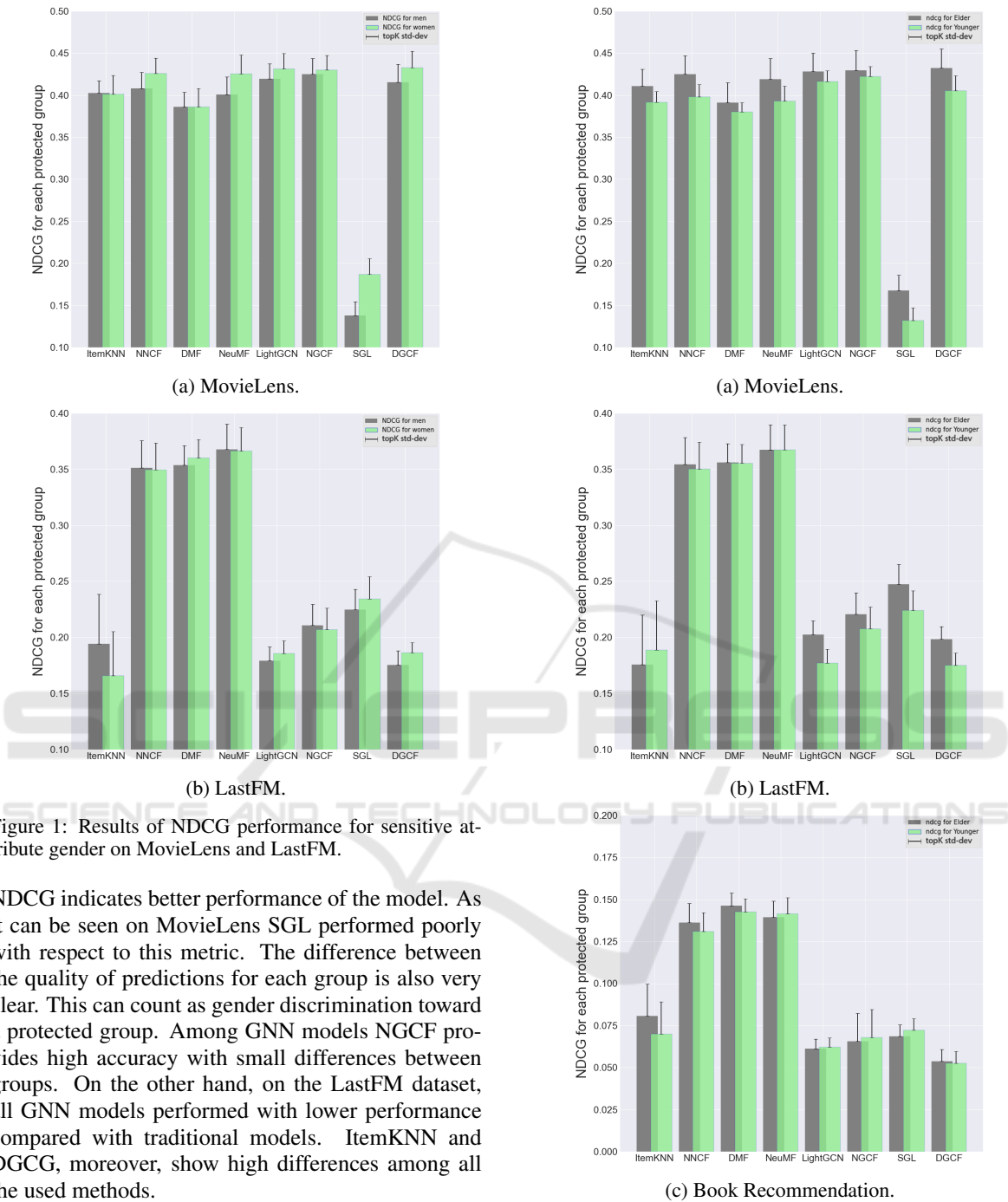


Figure 1: Results of NDCG performance for sensitive attribute gender on MovieLens and LastFM.

NDCG indicates better performance of the model. As it can be seen on MovieLens SGL performed poorly with respect to this metric. The difference between the quality of predictions for each group is also very clear. This can count as gender discrimination toward a protected group. Among GNN models NGCF provides high accuracy with small differences between groups. On the other hand, on the LastFM dataset, all GNN models performed with lower performance compared with traditional models. ItemKNN and DGCF, moreover, show high differences among all the used methods.

Figure 2 indicates the NDCG performance for sensitive attribute age for all datasets. Again on LastFM and Book Recommendation GNN models provide lower accuracy in comparison to conventional models. The quality differences between GNN models are higher in the LastFM dataset which indicates a higher rate of unfairness based on this dataset. SGL, also performs poorly on MovieLens.

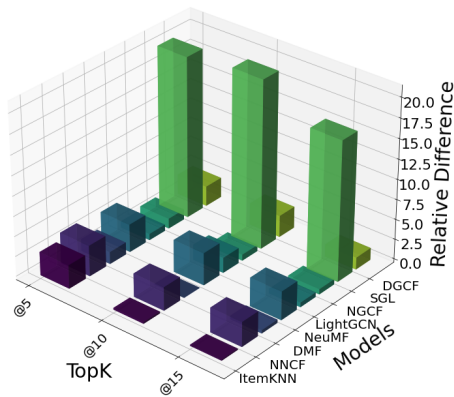
The following figures are provided to show the rel-

Figure 2: Results of NDCG performance for sensitive attribute age on MovieLens, LastFM, and Book Recommendation.

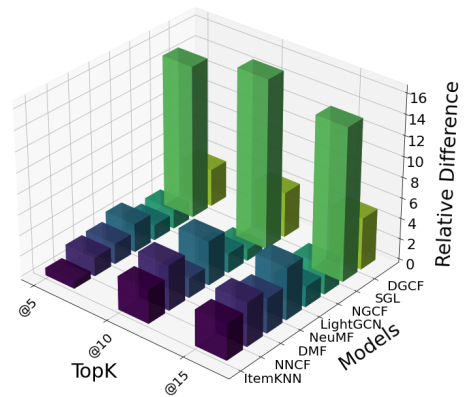
ative difference of the NDCG metric with respect to sensitive attributes.

Figure 3 shows the unfairness relative difference of sensitive attribute gender for MovieLens and LastFM. It can be seen that on the MovieLens dataset,

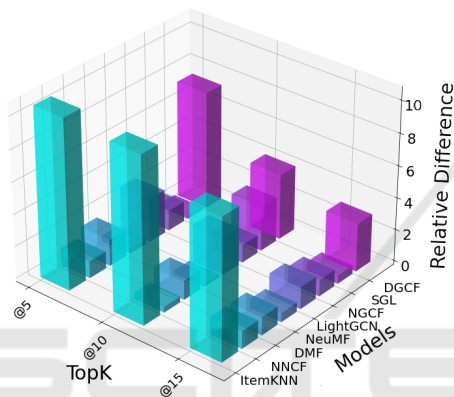




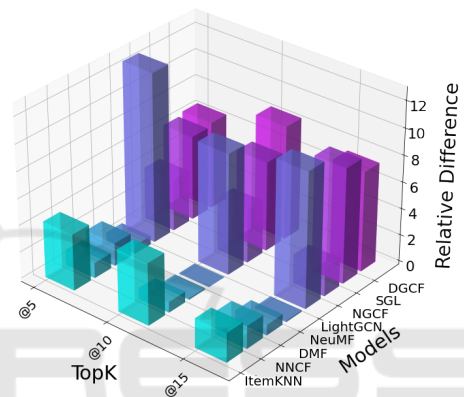
(a) MovieLens.



(a) MovieLens.



(b) LastFM.

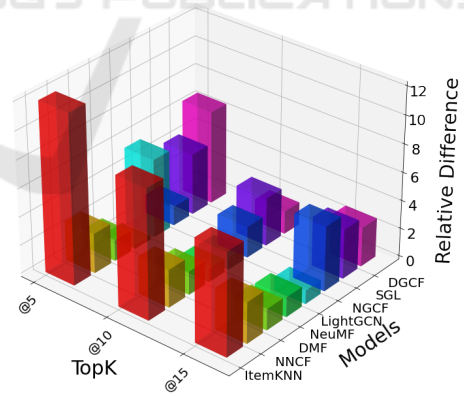


(b) LastFM.

Figure 3: Results of Unfairness Relative Difference of sensitive attribute gender for MovieLens and LastFM.

SGL has a significant relative difference which indicates high unfairness compared with other used models. Other used models performed moderately with respect to relative difference on MovieLens. On the other hand, on the LastFM dataset, DGCF shows higher unfairness among GNN methods and ItemKNN takes the first place regarding relative difference within all methods.

Figure 4 shows the unfairness relative difference of sensitive attribute gender for all datasets. SGL again, provides significant unfairness on MovieLens, although DGCF also shows a high unfairness compared with the rest. On the LastFM dataset, almost all GNN models show high unfairness except NGCF. LightGCN can be chosen as the most unfair one. For the Book Recommendation dataset, it can be witnessed that GNN models performed moderately and the most unfair method is ItemKNN.



(c) Book Recommendation.

Figure 4: Results of Unfairness Relative Difference of sensitive attribute age for MovieLens, LastFM, and Book Recommendation.

## 6 CONCLUSIONS AND FUTURE WORK

Fairness in RS holds great significance from both the user and service provider perspectives. Users rely on RS to receive personalized recommendations that align with their preferences and interests, while service providers aim to enhance user satisfaction and engagement. To assess and evaluate fairness in RS, a range of metrics have been developed in the state-of-the-art research. These metrics encompass various aspects, including individual and group fairness, providing valuable insights into recommendation quality for different user segments.

In this study, we provide a metric specifically designed to measure fairness disparities within RS recommendations, offering a fresh perspective on bias analysis. Unlike existing metrics, our new approach quantifies the differences in recommendation quality for protected groups, including gender and age. This metric allows us to evaluate how well the recommendations cater to the unique preferences and needs of these groups, shedding light on any potential biases or disparities in the system.

The introduction of this metric provides several benefits. Firstly, it enhances our understanding of fairness in RS by focusing on the quality of recommendations received by specific user groups, enabling a more granular assessment of bias. Secondly, it empowers service providers to tailor their recommendation algorithms to ensure fairness and inclusivity for all users. By having this information, RS platforms can make data-driven decisions to improve recommendation accuracy and user satisfaction, ultimately leading to a more equitable and effective RS ecosystem.

In our analysis of the three datasets (MovieLens, LastFM, and BookRec), we observed varying degrees of fairness and bias among different recommendation models across sensitive attributes, such as gender and age.

In the MovieLens dataset, models like DMF, LightGCN, NGCF, and DGCF demonstrated relatively fair recommendations for both male and female users, promoting fairness regardless of gender. They continued to exhibit fairness when considering the age-sensitive attribute, ensuring equitable recommendations for users across different age groups. In contrast, SGL did not provide fair recommendations in this dataset.

Turning our attention to the LastFM dataset, NCF, DMF, and NeuMF models displayed commendable fairness across protected groups, regardless of both sensitive attributes. These models maintained

minimal differences in NDCG accuracy between male and female users, indicating fairness in recommendations for both groups. The LightGCN model exhibited unique behavior, showing a higher relative NDCG difference in the age-sensitive attribute but a lower difference in the gender-sensitive attribute.

In the BookRec dataset, the relative difference in NDCG accuracy was generally low across various models. However, models exhibited some inconsistencies in their results, emphasizing the need for comprehensive fairness assessments.

For future work, the aim is to enhance the scalability of the used metrics to be capable of working on various features in different fields. These methods, moreover, can be applied to different sub-groups which can provide us with more detailed information regarding unfairness. Another type of accuracy method can also be used in order to measure the accuracy of recommended items in certain advantaged or disadvantaged groups.

## REFERENCES

- (2021). MovieLens.
- Aalam, S. W., Ahanger, A. B., Bhat, M. R., and Assad, A. (2022). Evaluation of fairness in recommender systems: A review. In *International Conference on Emerging Technologies in Computer Engineering*, pages 456–465. Springer.
- Abdollahpouri, H., Burke, R., and Mobasher, B. (2019). Managing popularity bias in recommender systems with personalized re-ranking. In *The thirty-second international flairs conference*.
- Ahanger, A. B., Aalam, S. W., Bhat, M. R., and Assad, A. (2022). Popularity bias in recommender systems—a review. In *International Conference on Emerging Technologies in Computer Engineering*, pages 431–444. Springer.
- Alelyani, S. (2021). Detection and evaluation of machine learning bias. *Applied Sciences*, 11(14):6271.
- Baeza-Yates, R. (2016). Data and algorithmic bias in the web. In *Proceedings of the 8th ACM Conference on Web Science*, pages 1–1.
- Barocas, S., Hardt, M., and Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, 1:2017.
- Bernhardt, M., Jones, C., and Glocker, B. (2022). Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. *Nature Medicine*, 28(6):1157–1158.
- Boratto, L., Fenu, G., and Marras, M. (2021). Connecting user and item perspectives in popularity debiasing for collaborative recommendation. *Information Processing & Management*, 58(1):102387.
- Boratto, L., Fenu, G., Marras, M., and Medda, G. (2022). Consumer fairness in recommender systems: Contextualizing definitions and mitigations. In *European*

- Conference on Information Retrieval*, pages 552–566. Springer.
- Boratto, L. and Marras, M. (2021). Advances in bias-aware recommendation on the web. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 1147–1149.
- Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42.
- Bruce, P., Bruce, A., and Gedeck, P. (2020). *Practical statistics for Data Scientists, 2nd edition*. O’Reilly Media, Inc.
- Caton, S. and Haas, C. (2020). Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*.
- Celma, O. (2010). *Music Recommendation and Discovery in the Long Tail*. Springer.
- Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., and He, X. (2020). Bias and debias in recommender system: A survey and future directions. *arXiv preprint arXiv:2010.03240*.
- Chen, Z., Xiao, T., and Kuang, K. (2022). Ba-gnn: On learning bias-aware graph neural network. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 3012–3024. IEEE.
- Chia, P. J., Tagliabue, J., Bianchi, F., He, C., and Ko, B. (2022). Beyond ndcg: behavioral testing of recommender systems with relict. In *Companion Proceedings of the Web Conference 2022*, pages 99–104.
- Chizari, N., Shoeibi, N., and Moreno-García, M. N. (2022). A comparative analysis of bias amplification in graph neural network approaches for recommender systems. *Electronics*, 11(20):3301.
- Chizari, N., Tajfar, K., and Moreno-García, M. N. (2023). Bias assessment approaches for addressing user-centered fairness in gnn-based recommender systems. *Information*, 14(2):131.
- Dai, E. and Wang, S. (2021). Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 680–688.
- Di Noia, T., Tintarev, N., Fatourou, P., and Schedl, M. (2022). Recommender systems under european ai regulations. *Communications of the ACM*, 65(4):69–73.
- Dong, Y., Liu, N., Jalaian, B., and Li, J. (2022a). Edits: Modeling and mitigating data bias for graph neural networks. In *Proceedings of the ACM Web Conference 2022*, pages 1259–1269.
- Dong, Y., Wang, S., Wang, Y., Derr, T., and Li, J. (2022b). On structural explanation of bias in graph neural networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 316–326.
- Fabbri, F., Croci, M. L., Bonchi, F., and Castillo, C. (2022). Exposure inequality in people recommender systems: The long-term effects. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 194–204.
- Fahse, T., Huber, V., and Giffen, B. v. (2021). Managing bias in machine learning projects. In *International Conference on Wirtschaftsinformatik*, pages 94–109. Springer.
- Farnadi, G., Kouki, P., Thompson, S. K., Srinivasan, S., and Getoor, L. (2018). A fairness-aware hybrid recommender system. *arXiv preprint arXiv:1809.09030*.
- Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.
- Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J., and Wen, Y. (2022). capai-a procedure for conducting conformity assessment of ai systems in line with the eu artificial intelligence act. Available at SSRN 4064091.
- Foulds, J. R., Islam, R., Keya, K. N., and Pan, S. (2019). Differential fairness. *UMBC Faculty Collection*.
- Gao, C., Lei, W., Chen, J., Wang, S., He, X., Li, S., Li, B., Zhang, Y., and Jiang, P. (2022a). Cirs: Bursting filter bubbles by counterfactual interactive recommender system. *arXiv preprint arXiv:2204.01266*.
- Gao, C., Wang, X., He, X., and Li, Y. (2022b). Graph neural networks for recommender system. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1623–1625.
- Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.
- 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*, pages 3779–3790.
- Khan, Z. Y., Niu, Z., Sandiwarno, S., and Prince, R. (2021). Deep learning techniques for rating prediction: a survey of the state-of-the-art. *Artificial Intelligence Review*, 54(1):95–135.
- Kordzadeh, N. and Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 31(3):388–409.
- Lazovich, T., Belli, L., Gonzales, A., Bower, A., Tantipongpipat, U., Lum, K., Huszar, F., and Chowdhury, R. (2022). Measuring disparate outcomes of content recommendation algorithms with distributional inequality metrics. *arXiv preprint arXiv:2202.01615*.
- Li, P., Wang, Y., Zhao, H., Hong, P., and Liu, H. (2021). On dyadic fairness: Exploring and mitigating bias in graph connections. In *International Conference on Learning Representations*.
- Li, X. (2023). *Graph Learning in Recommender Systems: Toward Structures and Causality*. PhD thesis, University of Illinois at Chicago.
- Lin, S., Wang, J., Zhu, Z., and Caverlee, J. (2022). Quantifying and mitigating popularity bias in conversational recommender systems. *arXiv preprint arXiv:2208.03298*.
- Loveland, D., Pan, J., Bhatena, A. F., and Lu, Y. (2022). Fairedit: Preserving fairness in graph neural net-

- works through greedy graph editing. *arXiv preprint arXiv:2201.03681*.
- Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., and Burke, R. (2021). A graph-based approach for mitigating multi-sided exposure bias in recommender systems. *ACM Transactions on Information Systems (TOIS)*, 40(2):1–31.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.
- Misztal-Radecka, J. and Indurkha, B. (2021). Bias-aware hierarchical clustering for detecting the discriminated groups of users in recommendation systems. *Information Processing & Management*, 58(3):102519.
- Mobius, A. (2020). Book recommendation dataset.
- Mu, R. (2018). A survey of recommender systems based on deep learning. *Ieee Access*, 6:69009–69022.
- Naghiaei, M., Rahmani, H. A., and Dehghan, M. (2022). The unfairness of popularity bias in book recommendation. *arXiv preprint arXiv:2202.13446*.
- Oneto, L. and Chiappa, S. (2020). Fairness in machine learning. In *Recent trends in learning from data: Tutorials from the inns big data and deep learning conference (innsbddl2019)*, pages 155–196. Springer.
- Pérez-Marcos, J., Martín-Gómez, L., Jiménez-Bravo, D. M., López, V. F., and Moreno-García, M. N. (2020). Hybrid system for video game recommendation based on implicit ratings and social networks. *Journal of Ambient Intelligence and Humanized Computing*, 11(11):4525–4535.
- Rahman, T., Surma, B., Backes, M., and Zhang, Y. (2019). Fairwalk: Towards fair graph embedding.
- Rajeswari, J. and Hariharan, S. (2016). Personalized search recommender system: State of art, experimental results and investigations. *International Journal of Education and Management Engineering*, 6(3):1–8.
- Ricci, F., Rokach, L., and Shapira, B. (2022). Recommender systems: Techniques, applications, and challenges. *Recommender Systems Handbook*, pages 1–35.
- Steck, H., Baltrunas, L., Elahi, E., Liang, D., Raimond, Y., and Basilico, J. (2021). Deep learning for recommender systems: A netflix case study. *AI Magazine*, 42(3):7–18.
- Sun, W., Khenissi, S., Nasraoui, O., and Shafto, P. (2019). Debiasing the human-recommender system feedback loop in collaborative filtering. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 645–651.
- Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (Fairware)*, pages 1–7. IEEE.
- Wang, S., Hu, L., Wang, Y., He, X., Sheng, Q. Z., Orgun, M. A., Cao, L., Ricci, F., and Yu, P. S. (2021). Graph learning based recommender systems: A review. *arXiv preprint arXiv:2105.06339*.
- Wang, X. and Wang, W. H. (2022). Providing item-side individual fairness for deep recommender systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 117–127.
- Wang, Y., Ma, W., Zhang, M., Liu, Y., and Ma, S. (2023). A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3):1–43.
- Wu, L., Chen, L., Shao, P., Hong, R., Wang, X., and Wang, M. (2021). Learning fair representations for recommendation: A graph-based perspective. In *Proceedings of the Web Conference 2021*, pages 2198–2208.
- Wu, S., Sun, F., Zhang, W., Xie, X., and Cui, B. (2020a). Graph neural networks in recommender systems: a survey. *ACM Computing Surveys (CSUR)*.
- Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. (2020b). A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24.
- Xu, B., Shen, H., Sun, B., An, R., Cao, Q., and Cheng, X. (2021). Towards consumer loan fraud detection: Graph neural networks with role-constrained conditional random field. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4537–4545.
- Yao, S. and Huang, B. (2017). Beyond parity: Fairness objectives for collaborative filtering. *Advances in neural information processing systems*, 30.
- Yu, J., Yin, H., Xia, X., Chen, T., Li, J., and Huang, Z. (2023). Self-supervised learning for recommender systems: A survey. *IEEE Transactions on Knowledge and Data Engineering*.
- Zeng, Z., Islam, R., Keya, K. N., Foulds, J., Song, Y., and Pan, S. (2021). Fair representation learning for heterogeneous information networks. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, volume 15.
- Zhang, Q., Wipf, D., Gan, Q., and Song, L. (2021). A biased graph neural network sampler with near-optimal regret. *Advances in Neural Information Processing Systems*, 34:8833–8844.
- Zheng, Y. and Wang, D. X. (2022). A survey of recommender systems with multi-objective optimization. *Neurocomputing*, 474:141–153.
- Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81.