Minds and Machines (2024) 34:33
https://doi.org/10.1007/s11023-024-09685-x

Toward a Responsible Fairness Analysis: From Binary to Multiclass and Multigroup Assessment in Graph Neural Network-Based User Modeling Tasks

Erasmo Purificato[1,3] · **Ludovico Boratto**[2] · **Ernesto William De Luca**[1,3]

Received: 31 May 2023 / Accepted: 11 June 2024 / Published online: 17 July 2024

Abstract
User modeling is a key topic in many applications, mainly social networks and information retrieval systems. To assess the effectiveness of a user modeling approach, its capability to classify personal characteristics (e.g., the gender, age, or consumption grade of the users) is evaluated. Due to the fact that some of the attributes to predict are multiclass (e.g., age usually encompasses multiple ranges), assessing *fairness* in user modeling becomes a challenge since most of the related metrics work with binary attributes. As a workaround, the original multiclass attributes are usually binarized to meet standard fairness metrics definitions where both the target class and sensitive attribute (such as gender or age) are binary. However, this alters the original conditions, and fairness is evaluated on classes that differ from those used in the classification. In this article, we extend the definitions of four existing fairness metrics (related to disparate impact and disparate mistreatment) from binary to multiclass scenarios, considering different settings where either the target class or the sensitive attribute includes more than two groups. Our work endeavors to bridge the gap between formal definitions and real use cases in bias detection. The results of the experiments, conducted on four real-world datasets by leveraging two state-of-the-art graph neural network-based models for user modeling, show that the proposed generalization of fairness metrics can lead to a more effective and fine-grained comprehension of disadvantaged sensitive groups and, in some cases, to a better analysis of machine learning models originally deemed to be fair. The source code and the preprocessed datasets are available at the following link: https://github.com/erasmopurif/toward-responsible-fairness-analysis.

Keywords User modeling · User profiling · Algorithmic fairness · Bias detection · Graph neural networks · Responsible artificial intelligence

Extended author information available on the last page of the article

1 Introduction

Living in the current digital era, the interaction with artificial intelligence (AI) systems has become, consciously or not, an integral part of everyone's life. In particular, among the most widespread and used tools, information retrieval (IR) systems and recommender systems (RSs) deal with providing relevant information to the end-users, according to their information needs, personality traits, and context, in an effective and efficient manner. In a scenario where the interplay with such systems produces a massive amount of personal data on a daily basis, the given need of deducting individuals' interests, characteristics, and behaviors is met by **user modeling** (in literature, used interchangeably with *user profiling*) techniques (Eke et al., 2019), which primarily aim to build a faithful user representation (i.e., a *user model*) (Purificato et al., 2024), starting from generated data. The initial *explicit* profiling approaches mostly take into account data derived from online surveys or forms and are mainly based on static user characteristics (Poo et al., 2003). Due to the familiar distrust and concern about providing personal information in a direct way by people, *implicit* strategies are usually considered in modern user modeling approaches. These techniques are also referred to as *behavioral user profiling* or *user behavior modeling* (Purificato et al., 2024). In the existing literature, the effectiveness and the performance of user modeling methods are commonly evaluated by assessing the related machine learning (ML) or deep learning (DL) model's accuracy in classifying a specific attribute (Chen et al., 2019) (e.g. a user's consumption level for an e-commerce platform). Recently, as automated decision-making systems have become ubiquitous in all areas, there has been an increasing realization that the development of such models and their results should adhere to a set of ethical principles (European-Commission, 2019). This has led to a push for research on topics such as transparency (Wang et al., 2019), privacy (Purificato et al., 2021), sustainability (Nilashi et al., 2019), and social equity (Gómez et al., 2021). With regard to the latter aspect, **algorithmic fairness** (Kleinberg et al., 2018; Mitchell et al., 2021) has received significant attention in both academic research and industry projects, mainly due to the increased awareness of the potential risks that unfair AI systems could pose to certain social groups. On the one hand, several studies have been carried out to investigate the possible sources of unfairness in automated systems (Loveland et al., 2022; Pessach & Shmueli, 2020). These sources are typically classified into two primary categories: (1) biased *data*, and (2) *algorithms* that are susceptible to the biases already inside the training datasets. On the other hand, a significant amount of procedures has been delivered to detect and mitigate bias in ML and DL (Barocas et al., 2019; Caton & Haas, 2020; Verma & Rubin, 2018), especially in user-related scenarios (Purificato et al., 2023), IR systems (Ekstrand et al., 2022; Gao & Shah, 2019; Ramos & Boratto, 2020) and RSs (Leonhardt et al., 2018; Ramos et al., 2020). Over the last few years, numerous fairness metrics have been defined in the literature (Barocas & Selbst, 2016; Berk et al., 2021; Dwork et al., 2012; Feldman et al., 2015; Hardt et al., 2016; Zafar et al., 2017), each with a unique focus on a particular aspect

of what could be deemed as "*fair*". Even though a universal connotation of the concept of fairness has not been defined yet, as illustrated by Chierichetti et al. (2019), the majority of the metrics that prioritize *classification parity* (meaning that predictive performance scores, i.e., true positive, true negative, false positive and false negative rates, should be equal across groups defined by the selected sensitive attributes) share a common feature, that is identifying and mitigating bias and inequity in *binary* problems. The reasons why this practice is widespread lie in two motivations that stand out above the others, as described by Caton and Haas (2020): (1) many applications involving ML models are originally binary (e.g. hiring vs. not hiring, granting vs. not granting a loan); (2) quantifying fairness on a binary dependent variable is mathematically more suitable. Although these two grounds are technically unexceptionable, what we want to deepen and analyze in this work concerns the implications that the application of such binary metrics may have in the context of user modeling, especially when applied in real-world scenarios, from an ethical and responsible perspective. Our arguments align with a similar criticism moved by Barocas et al. (2019): "*Most proposed fairness interventions start by assuming such a (binary) categorization. But when building real systems, enforcing rigid categories of people can be ethically questionable.*".

In this article, we aim to propose novel fairness metrics for a responsible evaluation in real and multi-valued contexts after analyzing different controversial aspects of binary fairness assessment in user modeling models and discussing the related issues from an ethical point of view. The term "*responsible*" refers, in this case, to the principles of Responsible AI (Dignum, 2019), which emphasizes accountability and transparency in the development and deployment of AI systems, often correlated with the concept of Human-Centered AI (Shneiderman, 2022).

In particular, we highlight the cases reported below:

- Fairness metrics are usually applied in user modeling scenarios where the classification techniques consider both the target class (e.g., consumption grade for e-commerce, salary) and the sensitive attribute (e.g., gender, age, race) as *binary*, often unnaturally.
- When evaluating a model's ability to produce fair results, it is commonly assessed based on the *absolute difference* between the scores of the two sensitive groups under consideration, and this can be perilous for both a system and user perspective. Adopting this approach, it is not feasible to identify disadvantaged groups for every possible combination of model, dataset, and fairness metrics. Consequently, it is not possible to implement targeted interventions to mitigate these issues in a real-world setting.
- In situations where there is not a clear binary separation within an actual attribute distribution, arranging a target class and/or a sensitive group into a binary representation by revising the initial data conditions might produce an imprecise assessment of a model's fairness.

From our point of view, there are two significant reasons why making fairness assessments of the actual distribution of classes and groups is paramount. On the

one hand, if user modeling is less effective for specific groups, they would inevitably receive less effective services (e.g., ads or recommendations). On the other hand, arranging different classes and groups into a binary representation can produce an inaccurate appraisal of models' fairness by even altering the original data conditions.

As the main contribution of the presented article, we extend the definition of four fairness metrics related to the concepts of *disparate impact* (Barocas & Selbst, 2016; Wan et al., 2021) and *disparate mistreatment* (Zafar et al., 2017), from *binary* to multi-*class* and multi-*group* scenarios.

Our work aims to bridge the gap between formal definitions and real use cases in bias detection by responding to the following research questions:

- **RQ1**: To what extent can multigroup fairness metrics impact a model's fairness evaluation with respect to the related binary metrics?
- **RQ2**: To what extent can multiclass and multigroup fairness metrics improve bias detection and future mitigation in real-world cases?

In the specific domain of behavioral user modeling, a recently published work (Purificato et al., 2022) presented an accurate assessment of fairness in binary scenarios employing effective **Graph Neural Network** (GNN)-based architectures in the field. This work is the first of its kind in the literature and reveals that different user modeling paradigms in GNNs have an impact on fairness results.

GNNs (Hamilton et al., 2017; Kipf & Welling, 2017; Veličković et al., 2017; Zhang et al., 2019) are currently considered as the state-of-the-art technologies for graph data, which constitutes one of the most suitable structure for modeling user behaviors, where nodes and edges depict, respectively, users and interactions among them. Besides user modeling (Chen et al., 2023, 2019; Rahimi et al., 2018; Yan et al., 2021), GNNs have been proven to be successful in several domains, such as IR (Cui et al., 2022), RSs (He et al., 2020; Ying et al., 2018) and natural language processing (Yao et al., 2019).

Given the aforementioned motivation, to address the challenges posed by the research questions, we also focus on *graph neural network-based models* for *behavioral* (i.e., implicit) *user modeling* as the case study of the presented work.

To provide a concrete example, in our experimental settings, we consider an input graph either heterogeneous (i.e., composed of different kinds of nodes, such as items and product) or homogeneous (i.e., composed of nodes of the same kind, such as all users in a social network). The node attributes constitute the user characteristics, where one attribute is selected as the target class for the classification predictions (i.e., the user modeling task), and a second attribute, specifically a personal trait, is chosen as the sensitive attribute for fairness assessment (e.g., gender or age). The graph's nodes are linked by connections that depend on the specific dataset (e.g., a "buy" relationship in an e-commerce or "follow" relationship in a social network).

By leveraging two of the most-performing state-of-the-art GNNs for user profiling tasks (i.e., **CatGCN** (Chen et al., 2023), and **RHGN** (Yan et al., 2021), the same used in the pioneering work in binary scenarios by Purificato et al. (2022)), we perform a beyond-accuracy analysis considering three settings of target classes and

sensitive attribute groups: (1) binary class and binary group; (2) binary class and multi-group; (3) multi-class and multi-group. We decided to keep out of our analysis the fourth possible combination, namely multi-class and binary group, because our primary focus is on the multiplicity of sensitive attributes, and that situation has already been investigated (Denis et al., 2021). The experiments are conducted on four real-world datasets (i.e., **Alibaba**, **JD**, **Pokec**, and **NBA**). The analysis of the results shows that the proposed generalization of fairness metrics to multi-class and multi-group cases can lead to a more effective and fine-grained comprehension of disadvantaged sensitive groups and a better assessment of ML and DL models (specifically GNNs in our work) improperly deemed as fair. Although this article's experiments focus on user modeling, it is worth noticing that, given the nature of the problem, the approach and resulting insights are not limited to this scenario. We envision these extensions being used to assess fairness in numerous ML tasks.

Our contributions can be summarized as follows:

- After reviewing the existing literature about user modeling on graph data (Sect. 2.1) and algorithmic fairness (Sect. 2.2), we provide a preliminary overview of the adopted (binary) fairness metrics (Sect. 3.1) and GNN-based user modeling models (Sect. 3.2) and datasets (Sect. 3.3).
- Based on these foundations, we discuss the implications of applying binary fairness metrics for user modeling tasks under an ethical and human-centered perspective (Sect. 4), starting from the analysis of the fairness assessment of GNN-based models for user modeling in a binary scenario published by Purificato et al. (2022).
- After setting the stage with the preliminary analysis discussed in the previous points, in the core part of this article, we extend classification fairness metrics definitions to cover scenarios in which both the target classes and the sensitive attributes are multiclass (Sect. 5).
- For the first time in the literature, we perform a comprehensive analysis to assess the effects of adopting the proposed generalized metrics over their binary version and evaluate them on four real-world datasets to assess (un)fairness in both binary and multiclass/multigroup settings (Sect. 6).
- We provide observations from the lessons learned and show that a binarization of the attributes can create the false perception that a user modeling approach is fair towards the users it models (Sect. 6.2). However, unfairness treatments can be uncovered when assessing fairness under the same conditions in which the classification was performed.
- We conclude the article by drawing potential future research directions for the addressed topic (Sect. 7).

2 Related Work

This section presents some relevant research work related to our context of interest. Because of our scenario's complexity and heterogeneity, we separately discuss the literature about *user modeling*, considering graph structures and GNN-based

models, and *algorithmic fairness*, with a specific focus on papers that introduced multi-class approaches, procedures or metrics.

2.1 User Modeling

The topic of **user modeling** on graph data was deeply investigated for the first time by Li et al. (2012), who leveraged a heterogeneous graph built upon two interaction types, namely "following" and "tweeting", to infer users' location. Progressively, the scene has been taken by GNN-based models, such as the work proposed by Rahimi et al. (2018) and Chen et al. (2019). In the former, users' location is detected through a geolocation model based on graph convolutional networks (GCNs) exploiting text and network information. The latter presented a user-representation learning approach with a heterogeneous graph attention network (HGAT), taking into account the graph structure and the attention mechanism to discern the importance of each node's neighbor. The most recent and promising models in this field were presented recently, and they are the two GNNs analyzed in our paper. A GCN-based model showing the advantages of boosting the node representation before executing the user profiling task has been proposed by Chen et al. (2023). Yan et al. (2021) presented a heterogeneous graph network (HEN) instead of intending to improve the model's performance by exploiting multiple types of relations and entities for user profiling.

As briefly discussed in the previous section, the current common practice in literature is to rank user modeling models and approaches based entirely on their capabilities to provide accurate predictions of a specific individual's characteristics. The aspect of bias detection has been poorly investigated for GNN models so far, with only a few works published in the area (e.g., Dai and Wang (2021); Dong et al. (2021)), and most of them mainly focusing on novel debiasing methods rather than analyzing potential fairness metrics limitations.

2.2 Algorithmic Fairness

In recent years, *bias* and *fairness* in ML have become the focus of growing attention. While the benefits of algorithmic decision-making can be compelling for large organizations and academic research, there is a potential for the output of these algorithms to be unfair (Mehrabi et al., 2021). If unfairness does occur, it can have significant perceptual and legal implications for organizations that opt to rely on machines to make important decisions (Caton & Haas, 2020). Therefore, it is essential to establish quantitative measures for bias and fairness in machine learning. By "bias", we mean that an ML model exhibits a preference for one characterization over another. In other words, the model has a lower error rate for one class than another. In particular, in this paper, we refer to *group fairness*, which is primarily focused on the outcomes of privileged and unprivileged groups (Binns, 2020). A group that characterizes an instance from the training data is a protected feature. In the broadest sense, group fairness separates a population into groups defined by protected attributes and aims to achieve equity across these groups. However, regardless

of the specific notion of fairness adopted, there exists an evident gap in the existing literature between the strategies and methods for binary and multi-class scenarios, as already described in Sect. 1. Only a few articles have been published in the last couple of years tackling this topic.

Blakeney et al. (2022) proposed two novel metrics, i.e. *Combined Error Variance* (CEV) and *Symmetric Distance Error* (SDE), to quantitatively assess the biases of each correspondent class in the comparison between two different models. CEV measures the inclination of a deep neural network to drop performance on one class in favor of others, while SDE computes the differences among the classes to be selected more or less frequently depending on the numerousness of their training examples.

Putzel and Lee (2022) took into account the problem of transforming the outcome of a black-box classifier by expanding the "post-processing" approach proposed by Hardt et al. (2016) to produce adjusted fair predictions for the analyzed model.

Denis et al. (2021) extended the known definition of *demographic parity* (Feldman et al., 2015) to the multi-class classification context for exact and approximate fairness cases and also provided optimal solutions for the classifier under both conditions.

Alghamdi et al. (2022) focus on creating fair probabilistic classifiers for multi-class classification tasks. The proposed approach involves "projecting" a pre-trained classifier, which may be biased, onto the set of models that fulfill the target group fairness criteria. The resulting projected model is determined by post-processing the pre-trained classifier's outputs with a multiplicative factor. Moreover, the authors introduced an iterative algorithm that can be parallelized to calculate the projected classifier and provide guarantees for both sample complexity and convergence.

The main limitation of these (and other existing) works that we aim to overcome in the presented paper is the lack of an in-depth analysis of the effect of the application of binary fairness metrics in real-world scenarios, with particular reference to user modeling. Indeed, in most cases, novel proposed procedures and methods have the goal of only solving "mathematically" the issue of binary categorization for bias detection and mitigation.

3 Preliminaries

In this section, we first illustrate the definition of the standard binary fairness metrics (Sect. 3.1), which constitute the basis of our novel extended metrics. Then, we provide an overview of the GNN-based models (Sect. 3.2) and the datasets (Sect. 3.3) considered in our case study.

3.1 Standard Fairness Metrics Definition

As discussed in Sect. 1, several fairness metrics have been proposed in the literature in the last decade (Barocas & Selbst, 2016; Berk et al., 2021; Dwork et al., 2012; Feldman et al., 2015; Hardt et al., 2016; Zafar et al., 2017). Our article

considers the two families of fairness metrics covering the perspectives of *disparate impact* and *disparate mistreatment*.

Disparate impact, also known as *adverse impact*, occurs when people are apparently treated similarly by a procedure or a system, but they are subject to indirect and often unintentional discrimination (Hajian et al., 2016). This usually occurs when certain groups are systematically discriminated against, even though no sensitive attribute is considered when making the predictions. Hence, disparities arise due to some proxy attributes (Wan et al., 2021). This family of metrics was chosen since the GNNs considered by our reference models only aggregate information from neighbors. In turn, no sensitive attribute is explicitly considered during the classification process. Disparate impact can be employed under scenarios in which no explicit link between the predicted label and the sensitive attribute exists, meaning that it is hard to define the validity of a decision for a group member based on the historical training data (Zafar et al., 2017).

Disparate mistreatment is related to scenarios in which it becomes challenging to evaluate the correctness of a prediction for users associated with a specific sensitive attribute value. Rather than considering the corrected predictions, it is assessed by measuring the *misclassification rates* for groups of users characterized by different values of a sensitive attribute (Zafar et al., 2017). Finally, assessing significance in contexts where the misclassification costs depend on the group affected by the error is particularly useful.

Our evaluation of the disparate impact of the analyzed models will be done through metrics assessing perspectives such as *statistical parity*, *equal opportunity*, and *overall accuracy equality*. We select the *treatment equality* metric for disparate mistreatment. For each of these metrics, we consider $y \in \{0, 1\}$ as the binary target label and $\hat{y} \in \{0, 1\}$ as the prediction of the user modeling model $f : x \rightarrow y$. The sensitive attribute is denoted with $s \in \{0, 1\}$. In the metrics' descriptions, we also exploit the following notation, which relates to classification properties: TP, FP, TN, and FN, denoting *true positives*, *false positives*, *true negatives* and *false negatives*, respectively.

Statistical parity (SP) (or *demographic parity*) (Dwork et al., 2012; Feldman et al., 2015) defines fairness as an equal probability for each group of being assigned to the positive class, i.e. predictions independent are from the sensitive attributes.

$$P(\hat{y} = 1 \mid s = 0) = P(\hat{y} = 1 \mid s = 1) \tag{1}$$

Equal opportunity (EO) (Hardt et al., 2016) requires the probability of a subject in a positive class to be classified with the positive outcome to be equal for each group, i.e. TP should be the same across groups.

$$P(\hat{y} = 1 \mid y = 1, s = 0) = P(\hat{y} = 1 \mid y = 1, s = 1) \tag{2}$$

Overall accuracy equality (OAE) (Berk et al., 2021) defines fairness as the equal probability of a subject from either the positive or the negative class to be assigned to its respective class, i.e. each group should have the same prediction accuracy.

$$P(\hat{y} = 0 \mid y = 0, s = 0) + P(\hat{y} = 1 \mid y = 1, s = 0) = \\ = P(\hat{y} = 0 \mid y = 0, s = 1) + P(\hat{y} = 1 \mid y = 1, s = 1) \tag{3}$$

Treatment equality (TE) (Berk et al., 2021) requires the ratio of errors made by the classifier to be equal across different groups, i.e. each group should have the same FN and FP ratio.

$$\frac{P(\hat{y} = 1 \mid y = 0, s = 0)}{P(\hat{y} = 0 \mid y = 1, s = 0)} = \frac{P(\hat{y} = 1 \mid y = 0, s = 1)}{P(\hat{y} = 0 \mid y = 1, s = 1)} \tag{4}$$

3.2 Analysed GNN Models

The foundation of our fairness assessment is represented by two recent GNN-based models, which currently represent the most effective advances in user profiling, i.e. *CatGCN* and *RHGN*.

The first, **CatGCN** (Chen et al., 2023), is a graph convolutional network (GCN) model that performs graph learning on categorical node features. Instead of considering the original node representation, the model integrates two additional forms of interaction into the learning process. The first of them, a local interaction, is multiplication-based and performed on each pair of node features. The second is an addition-based interaction, in which the model builds on an artificial feature graph. The introduction of these forms of interaction before the graph convolution can improve the effectiveness of user modeling.

The second approach, named Relation-aware Heterogeneous Graph Network (**RHGN**) (Yan et al., 2021), models different forms of interactions happening between entities on a heterogeneous graph. Node importance and meta-relation significance on the graph are learned via transformer-like multi-relation attention, while information from multiple sources is collected via a heterogeneous graph propagation network. A comparison with other GNN-based models on user modeling tasks shows the approach's effectiveness.

3.3 Datasets

In this section, we describe the four datasets used in the presented work: *Alibaba*, *JD*, *Pokec*, and *NBA*.

Alibaba dataset[1] consists of click-through rate data related to the ads of Alibaba's Taobao platform, provided by the Alibaba Group's Tianchi Lab in 2018. Both CatGCN and RHGN models performed their original evaluation with this dataset. The heterogeneous graph generated as the models' input is composed of two types of nodes, users and items (i.e., products). A user node includes attributes related to gender, age, consumption grade, student status, and region of living. An item node has only one attribute, thus the category to which the product belongs. User and

[1] https://tianchi.aliyun.com/dataset/dataDetail?dataId=56

item nodes are connected with the *click* relationship, and the edges are not weighted. Following the experimental setup of CatGCN, we select the product types as the categorical features associated with the users for the same model. Thus, only the items clicked by at least two users have been taken into account to create the *co-click* relationship adopted as the model's local interaction. In order to make the study coherent between the analyzed GNNs, the same filtering procedure has been applied to RHGN before creating the heterogeneous graph. As the target class for the user modeling task and the sensitive attribute, we choose the user's *consumption grade* (denoted as *buy*) and *age*, respectively. Considering the binary scenario, we generate the *bin-buy* variable from the original 3-level *buy* attribute by merging *mid* ($y = 1$) and *high* ($y = 2$) levels, and the *bin-age* variable from the 7-level *age* attribute by merging labels as follows: $s_A = \{s_0, s_1, s_2, s_3\}$ and $s_B = \{s_4, s_5, s_6\}$. In the Alibaba dataset, the age range of each class is not specified and is only characterized by a label. Both binarisations have been made to define a clear separation between the two groups.

JD dataset[2] consists of 100 000 users randomly sampled from JD.com, one of the largest e-commerce sites in the world, and collected by Chen et al. (2019). It includes users' profiles, information about items (i.e., products), click and order logs ranging from February 2018 to February 2019, and has been used in the original RHGN paper for its experimental evaluation. User profile (i.e., gender and age) and product data (i.e., category information, brand, and price) are leveraged for generating the user and item nodes for the heterogeneous input graph. Given the massive size of this dataset, and since the scope of the presented work is not the evaluation of models' performance on user modeling tasks, we make a sample of the dataset taking the 15% of the items and consider only one relationship type, i.e., *click*, as the graph edges, to create comparable experimental setups. As for the Alibaba dataset, a *co-click* relationship is used as CatGCN's local interaction. To make the different experiments as consistent as possible, we generate a variable named *expense level* and use it as the profiling task target class. We exploit the existing *purchase* relationship between user and item nodes, and the *count* of bought items and each single *price* to compute a user's total expense. After removing duplicate values, we divided the list of expenses into four quartiles to extract the boundaries for creating a 4-level variable. The binary variable *bin-exp* has been constructed by isolating the *low* level ($y = 0$) and merging the others, following the practice adopted on the Alibaba dataset. The 5-level *age* variable is the sensitive attribute, and in this case, we binarised it (*bin-age*) by considering users under and over 35 years old. The resulting binary sensitive attribute groups are composed as follows: $s_A = \{s_0, s_1\}$ and $s_B = \{s_2, s_3, s_4\}$.

Pokec is the most popular social network in Slovakia, which is very similar to Facebook and X (former Twitter). This dataset[3] has already been used in other relevant works, such as in Dai and Wang (2021). It contains anonymized data of the whole social network in 2012 and has been published by Takac and Zabovsky

[2] https://github.com/guyulongcs/IJCAI2019_HGAT

[3] https://snap.stanford.edu/data/soc-pokec.html

Table 1 Characteristics of the used datasets

Dataset	Users	Items	Edges	Features
Alibaba	166,958	64,553	427,464	2820
JD	38,322	49,634	315,970	2056
Pokec	13,504	–	882,765	70
NBA	403	–	16,570	178

Table 2 Distribution of the *original* target classes and sensitive attribute groups

Dataset	Label	% Class/Group						
		0	*1*	*2*	*3*	*4*	*5*	*6*
Alibaba	buy	32.48%	60.30%	7.22%	–	–	–	–
	age	21.74%	1.61%	17.56%	23.72%	30.83%	4.53%	0.01%
JD	expense	40.99%	15.68%	23.97%	19.36%	–	–	–
	age	23.59%	7.53%	50.17%	16.95%	1.76%	–	–
Pokec	work-field	47.67%	21.10%	13.12%	12.41%	5.70%	–	–
	age	38.85%	30.10%	13.64%	9.98%	7.43%	–	–
NBA	salary	22.33%	38.21%	39.45%	–	–	–	–
	age	39.95%	37.37%	22.58%	–	–	–	–

(2012). The nodes of the input graph are homogeneous and represent users of the platform, having specific attributes (e.g., gender, age, hobbies, interests, and working field). The edges are the connections between the users, represented by a *follow* relationship, and they are not weighted. For the user modeling task, we employ the *working field* as the target class. The categories of this attribute are identified solely by labels and lack detailed documentation. For this reason, the binarization process, which generates the *bin-work-field* variable, is performed by isolating the most numerous class ($y = 0$) to establish a distinct division between the two groups, ensuring clarity in group delineation. The *age* is used as the sensitive attribute. For this dataset, each node includes the exact age of the users. To generate a meaningful and almost balanced set of levels, we consider the following ranges to create five groups: under 18, 18–23, 24–28, 28–35, and over 35. Given that it is a social network, to create the *bin-age* attribute for the binary scenario, we decided to consider the users under and over 18 years old. Specifically, the groups created are: $s_A = \{s_0\}$ and $s_B = \{s_1, s_2, s_3, s_4\}$.

NBA dataset[4] is an extension of a Kaggle dataset,[5] used by Dai and Wang (2021) and containing the info about around 400 NBA basketball players. The performance statistics of players in the 2016–2017 season and other various information (e.g., nationality, age, and salary) are provided, and constitute the attributes of our

[4] https://github.com/EnyanDai/FairGNN/tree/main/dataset/NBA

[5] https://www.kaggle.com/noahgift/social-power-nba

Table 3 Distribution of the *binarized* target classes and sensitive attribute groups

Dataset	Label	% Class/Group	
		0	*1*
Alibaba	bin-buy	32.48%	67.52%
	bin-age	64.63%	35.37%
JD	bin-exp	40.99%	59.01%
	bin-age	67.12%	32.88%
Pokec	bin-work-field	47.67%	52.33%
	bin-age	38.85%	61.15%
NBA	bin-salary	60.55%	39.45%
	bin-age	60.05%	39.95%

homogeneous input graph nodes. The graph edges, not weighted, are represented by the relations between players on a social network (i.e., Twitter, retrieved from the official crawling API by Dai and Wang (2021)). For this dataset, the three-level *salary* attribute is adopted as the target class of the performed user modeling task, which is binarized by isolating the top-level class ($y = 2$) for the binary scenario (*bin-salary*). As the sensitive attribute, we employed the *age* attribute. From the individual values, we first created three ranges considering meaningful groups for basketball players (i.e., under 25, 25–30, and over 30), and then we merged the two highest groups to create the *bin-age* variable with the following split: $s_A = \{s_0\}$ and $s_B = \{s_1, s_2\}$.

Table 1 shows information about the four datasets, where *features* refers to the dimension of CatGCN's input categorical feature array. Tables 2 and 3 display, respectively, the distribution within the datasets of the target classes and sensitive attribute groups in the original and binarized scenarios.

In our study, we binarized the target classes and sensitive attributes to allow for a comprehensive analysis of class distributions. This approach enabled us to examine both balanced and unbalanced distributions of positive and negative classes. By implementing this binarization strategy, we ensured that our analysis could robustly capture the effects of varying class distributions on the outcomes, thereby enhancing the reliability and generalizability of our findings.

4 Ethical Implications of Fairness Analysis in a Binary Scenario

The first fairness assessment of GNN-based models for user modeling shown in the literature has been recently published by Purificato et al. (2022), which constitutes the starting point of the contributions presented in our article. In their work, the authors analyzed two state-of-the-art GNNs (i.e. *CatGCN* and *RHGN*, described in Sect. 3.2) by evaluating their performance in two binary user modeling scenarios (exploiting *Alibaba* and *JD* datasets, illustrated in Sect. 3.3) and assessing *disparate impact* and *disparate mistreatment* values (Sect. 3.1) for both models in both user modeling tasks.

In particular, for the mentioned assessment, they quantitatively evaluate the disparate impact and disparate mistreatment of the analyzed models by operationalizing the metrics defined by Eqs. (1)–(4) as follows:

Table 4 Fairness metrics computation without absolute value for Purificato et al. (2022)

Dataset	Model	Δ^*_{SP}	Δ^*_{EO}
Alibaba	CatGCN	− 0.045 ±0.021	0.139 ±0.074
	RHGN	0.019 ±0.012	− 0.133 ±0.086
JD	CatGCN	0.033 ±0.013	− 0.052 ±0.016
	RHGN	0.009 ±0.007	− 0.042 ±0.017

$$\Delta_{SP} = |P(\hat{y} = 1 \mid s = 0) - P(\hat{y} = 1 \mid s = 1)|, \tag{5}$$

$$\Delta_{EO} = |P(\hat{y} = 1 \mid y = 1, s = 0) - P(\hat{y} = 1 \mid y = 1, s = 1)|, \tag{6}$$

$$\begin{aligned} \Delta_{OAE} = |&P(\hat{y} = 0 \mid y = 0, s = 0) + P(\hat{y} = 1 \mid y = 1, s = 0) - \\ &- P(\hat{y} = 0 \mid y = 0, s = 1) - P(\hat{y} = 1 \mid y = 1, s = 1)|, \end{aligned} \tag{7}$$

$$\Delta_{TE} = \left| \frac{P(\hat{y} = 1 \mid y = 0, s = 0)}{P(\hat{y} = 0 \mid y = 1, s = 0)} - \frac{P(\hat{y} = 1 \mid y = 0, s = 1)}{P(\hat{y} = 0 \mid y = 1, s = 1)} \right| \tag{8}$$

Through an extensive set of experiments, the authors derived several observations about the analyzed models, correlating their different user modeling paradigms with the fairness metrics scores to create a baseline for future assessments:

1. The ability of *RHGN* to represent users through multiple interaction modeling gains better values in terms of fairness than a model only relying on binary associations between users and items, as *CatGCN*, which also amplifies discrimination by modeling users' local interactions.
2. Even though *RHGN* demonstrates to be a fairer model than *CatGCN*, a debiasing process is equally needed in order to exploit the user models produced by both GNNs while deeming them as fair.
3. In scenarios where the correctness of a decision on the target label w.r.t. the sensitive attributes are not well defined or where there is a high cost for misclassified instances, a complete fairness assessment should always take into account disparate mistreatment evaluation since disparate impact results could be misleading for these specific contexts.

As discussed in Sect. 1, in the algorithmic fairness literature, many researchers disagree with the choices of binarising target class and using absolute value scores for computing fairness. To further demonstrate the limitations of these practices and set the path to our multiclass and multigroup metrics proposal, which constitutes the core contribution of this article, we conducted two types of experiments based on the computation made by Purificato et al. (2022).[6] In the first one, we focused

[6] Source code of Purificato et al. (2022) at https://github.com/erasmopurif/do_gnns_build_fair_models.

Table 5 Statistical parity scores for binary and multiclass sensitive attribute groups for Purificato et al. (2022) (RHGN model and Alibaba dataset)

Binary group	SP	Multiclass group	SP
s_A	0.887 ± 0.015	s_0	0.81 ± 0.02
		s_1	0.91 ± 0.02
		s_2	0.91 ± 0.01
		s_3	0.92 ± 0.01
s_B	0.797 ± 0.055	s_4	0.89 ± 0.01
		s_5	0.72 ± 0.03
		s_6	0.78 ± 0.07

on the use of the *absolute difference* of the computed fairness metrics. The setting is straightforward: we removed the absolute value from the fairness computation of the analyzed models and executed the same experiments presented in the original papers with the default parameters. Table 4 displays the results of the computation of Δ^*_{SP} and Δ^*_{EO} (i.e., Δ_{SP} and Δ_{EO} without absolute value), showing the evident alternation of positive and negative scores. The resulting trend means that the unfairness (regardless of the specific value) might be directed towards one sensitive group or the other for a given combination of model and dataset.

Ethical consideration 1 Considering the absolute difference score in the fairness analysis can be hazardous. In particular, from both a system and user perspective, with this practice, we cannot clearly figure out the disadvantaged groups for every specific combination of model, dataset, and fairness metrics, and thus unable to make in place just tailored interventions to mitigate the issue in a real-world scenario.

Concerning the issue related to fairness analysis in binary scenarios, we conducted experiments to understand the influence of binarization on fairness scores. Not being the core of this article, we discuss below only the results for a specific combination of model and dataset. The derived implication can be easily extended to the other cases.

In particular, we focused on RHGN model and Alibaba dataset from Purificato et al. (2022) work, also adopting the original binary classification task but with the following setting for the sensitive attribute. On the one hand, we considered its original multiclass distribution (seven groups, named as s_0-s_6) and calculated every single *statistical parity* (*SP*) probability; on the other hand, we binarized the attribute, as done in the original paper, and again computed the single probabilities for the binary groups. The resulting binary sensitive attribute groups are composed as follows: $s_A = \{s_0, s_1, s_2, s_3\}$, $s_B = \{s_4, s_5, s_6\}$. The results are shown in Table 5.

The observation derived from these results is that binarization can lead to misleading evaluation of a specific subgroup. In this specific experiment, the group s_0 should be treated as a disadvantaged group if considered in the fine-grained assessment, but it would be treated as an advantaged group when included in the binary group s_A. The opposite applies to group s_4.

Ethical consideration 2 In many of the current works about fairness evaluation of automated systems, the sensitive attributes (that are usually natively multiclass) are made binary to meet the standard fairness metrics definitions. From our point of view, there are two crucial reasons why it is essential to evaluate fairness by examining the actual distribution of sensitive groups. Firstly, if the system at hand is not as effective for certain groups, they will end up receiving less effective services, such as targeted advertisements or recommendations. Secondly, reducing the different classes and groups into a binary representation can lead to an incorrect evaluation of the fairness of models, potentially distorting the original data conditions.

5 Multiclass and Multigroup Fairness Metrics

In the following section, we describe the context, the motivations, and the steps that led to the definition of **multigroup fairness metrics** first and then finally to the general **multiclass and multigroup metrics**.

As reported in Sect. 2, one of the primary reasons behind the standardized adoption of binary fairness metrics is that many ethically questionable applications involving AI systems are binary by definition (e.g., hiring vs. not hiring). The main issue with this motivation is that it cannot be true when considering sensitive attributes. This is basically due to the common understanding that almost no human traits should be viewed as binary, neither gender nor, even more so, age.

For **multigroup fairness metrics**, we take the same class variables as the binary case (Sect. 3.1), that is $y \in \{0, 1\}$ as the binary target label and $\hat{y} \in \{0, 1\}$ as the model prediction. Let N be the number of sensitive attribute groups s. We define the following equations, where the resulting score should be equal across the groups to satisfy the specific fairness metrics:

- **Multigroup statistical parity**

$$P(\hat{y} = 1 \mid s = n), \forall n \in \{0, ..., N-1\} \tag{9}$$

- **Multigroup equal opportunity**

$$P(\hat{y} = 1 \mid y = 1, s = n), \forall n \in \{0, ..., N-1\} \tag{10}$$

- **Multigroup overall accuracy equality**

$$P(\hat{y} = 0 \mid y = 0, s = n) + P(\hat{y} = 1 \mid y = 1, s = n), \forall n \in \{0, ..., N-1\} \tag{11}$$

- **Multigroup treatment equality**

$$\frac{P(\hat{y} = 1 \mid y = 0, s = n)}{P(\hat{y} = 0 \mid y = 1, s = n)}, \forall n \in \{0, ..., N-1\} \tag{12}$$

The second rationale in favor of the binary fairness metrics definition reported by Caton and Haas (2020) relates to the mathematical ease of quantifying a binary variable instead of a multivalue one. Following the above multigroup definitions, we propose a further extension to **multiclass and multigroup fairness metrics** without adding any mathematical complexity, with the goal of proving that an apparently simple generalization can lead to a much better and deeper fairness analysis. Let M and N be the number of classes y,$\hat{y}$ and groups s, respectively. The score of each metric displayed below should be equal across every class and group:

- **Multiclass and multigroup statistical parity**

$$P(\hat{y} = m \mid s = n), \forall m \in \{0, ..., M-1\} \wedge \forall n \in \{0, ..., N-1\} \tag{13}$$

- **Multiclass and multigroup equal opportunity**

$$P(\hat{y} = m \mid y = m, s = n), \forall m \in \{0, ..., M-1\} \wedge \forall n \in \{0, ..., N-1\} \tag{14}$$

- **Multiclass and multigroup overall accuracy equality**

$$\sum_{m=0}^{M-1} P(\hat{y} = m \mid y = m, s = n), \forall n \in \{0, ..., N-1\} \tag{15}$$

- **Multiclass and multigroup treatment equality**

$$\frac{P(\hat{y} = m \mid y \neq m, s = n)}{P(\hat{y} \neq m \mid y = m, s = n)}, \forall m \in \{0, ..., M-1\} \wedge \forall n \in \{0, ..., N-1\} \tag{16}$$

It is worth noticing that when considering the multiclass and multigroup scenario, the definition of *equal opportunity* in Eq. (14) would also apply to the extension of the *equalized odds* metric (Hardt et al., 2016) in the same context.

6 Experimental Fairness Assessment

In this section, we present the empirical study conducted to assess the effects of the proposed multiclass and multigroup fairness metrics compared to the standard binary metrics to answer the following research questions, already introduced in Sect. 1:

- **RQ1**: To what extent can multigroup fairness metrics impact a model's fairness evaluation with respect to the related binary metrics?
- **RQ2**: To what extent can multiclass and multigroup fairness metrics improve bias detection and future mitigation in real-world cases?

Table 6 Experiment results of the user modeling tasks, for each combination of dataset, model, and setting (binary or multiclass)

Dataset	Model	Performance (binary)		Performance (multiclass)	
		Accuracy	*F1-score*	*Accuracy*	*F1-score*
Alibaba	CatGCN	0.776 ±0.021	0.718 ±0.005	0.535 ±0.031	0.501 ±0.012
	RHGN	0.803 ±0.006	0.711 ±0.016	0.618 ±0.002	0.587 ±0.018
JD	CatGCN	0.732 ±0.008	0.706 ±0.006	0.502 ±0.002	0.498 ±0.013
	RHGN	0.738 ±0.004	0.702 ±0.007	0.575 ±0.010	0.525 ±0.017
Pokec	CatGCN	0.808 ±0.002	0.797 ±0.002	0.445 ±0.004	0.398 ±0.006
	RHGN	0.799 ±0.022	0.779 ±0.013	0.455 ±0.004	0.404 ±0.003
NBA	CatGCN	0.743 ±0.074	0.709 ±0.052	0.593 ±0.067	0.541 ±0.072
	RHGN	0.768 ±0.043	0.721 ±0.071	0.581 ±0.051	0.527 ±0.035

Below, we describe the experiments carried out to investigate the research questions set and the chosen parameters. The experimental results conclude the section.

6.1 Experimental setting

To enable the analyzed GNN-based models to perform properly on the user modeling tasks with the provided dataset and the selected target classes and sensitive attributes, we execute a hyper-parameters selection as described in the following. For CatGCN, the *learning rate* is searched in {0.001, 0.01, 0.1}, the L_2 regularization coefficient and the *dropout ratio* are tuned among {1e-5, 1e-4} and {0.1, 0.3, 0.5, 0.7}, respectively, and the aggregation parameter α is searched within {0.1, 0.3, 0.5, 0.7, 0.9}. For RHGN, the *learning rate* and the L_2 regularization coefficient are searched in {0.01, 0.1} and {1e-5, 1e-4}, respectively; the hidden dimension of the two layers of the entity-level aggregation network is searched in {32, 64}, while the number of heads in multi-head attention is tuned among {1, 2}. All other parameters are set according to the original papers. After the grid search, we ran the experiments 40 times for each fairness metric. We execute the illustrated operations on a GPU Nvidia Quadro RTX 8000 48GB.

6.2 Experimental results

In this section, we discuss the results and findings of the empirical studies designed for each research question. Before diving into the fairness assessments, the experiment results of the user modeling tasks, for each combination of dataset, model, and setting (binary or multiclass) are shown in Table 6, reporting the performance scores in the form of *accuracy* and *F1-score*. Displaying this table aims to amplify the understanding of the chosen models' effectiveness and enhance the relevance of the presented fairness metrics in all their variations.

Table 7 Qualitative analysis of the comparative results between *binary* and *multigroup* scenarios leading to the considerations for RQ1. The *multigroup* column includes the differences from the binary case

Dataset	Metric	Model /setting (Ref. Figures)					
		CatGCN			RHGN		
		Binary	Multigroup		Binary	Multigroup	
Alibaba	SP	s_A adv.	$s_{0.A}$ dis., $s_{4.B}$, $s_{6.B}$ adv.	(1a-2a)	s_A adv.	$s_{0.A}$ dis., $s_{4.B}$ adv.	(9a-10a)
	EO	Fair	$s_{0.A}$, $s_{5.B}$ dis.	(1b-2b)	s_A adv.	$s_{0.A}$ dis., $s_{4.B}$ adv.	(9b-10b)
	OAE	Fair	$s_{1.A}$, $s_{2.A}$ adv.	(1c-2c)	Fair	$s_{3.A}$, $s_{4.B}$ dis., $s_{5.B}$, $s_{6.B}$ adv.	(9c-10c)
	TE	s_A adv.	$s_{0.A}$ dis.	(1d-2d)	s_B adv.	$s_{4.B}$ dis.	(9d-10d)
JD	SP	Fair	$s_{3.B}$, $s_{4.B}$ dis.	(3a-4a)	s_B adv.	$s_{3.B}$ dis.	(11a-12a)
	EO	Fair	$s_{3.B}$ dis.	(3b-4b)	Fair	$s_{0.A}$ dis.	(11b-12b)
	OAE	Fair	$s_{2.B}$, $s_{3.B}$ dis.	(3c-4c)	Fair	$s_{2.B}$ adv.	(11c-12c)
	TE	s_B adv.	$s_{2.B}$ dis.	(3d-4d)	Fair	$s_{0.A}$ adv., $s_{1.A}$, $s_{2.B}$ dis.	(11d-12d)
Pokec	SP	Fair	$s_{3.B}$, $s_{4.B}$ adv.	(5a-6a)	Fair	$s_{2.B}$ dis.	(13a-14a)
	EO	Fair	$s_{3.B}$ adv.	(5b-6b)	Fair	*no diff.*	(13b-14b)
	OAE	s_B adv.	$s_{1.B}$, $s_{2.B}$, $s_{4.B}$ dis.	(5c-6c)	Fair	$s_{3.B}$ adv.	(13c-14c)
	TE	Fair	$s_{3.B}$, $s_{4.B}$ dis.	(5d-6d)	s_A adv.	$s_{1.B}$, $s_{2.B}$ adv.	(13d-14d)
NBA	SP	s_A adv.	$s_{2.B}$ adv.	(7a-8a)	s_A adv.	$s_{2.B}$ adv.	(15a-16a)
	EO	s_A adv.	$s_{2.B}$ adv.	(7b-8b)	s_A adv.	$s_{2.B}$ adv.	(15a-16a)
	OAE	s_A adv.	*no diff.*	(7c-8c)	s_B adv.	$s_{1.B}$ dis.	(15a-16a)
	TE	s_A adv.	$s_{1.B}$ adv.	(7d-8d)	s_A adv.	*No diff.*	(15a-16a)

6.2.1 Comparing multigroup and binary fairness evaluation (RQ1)

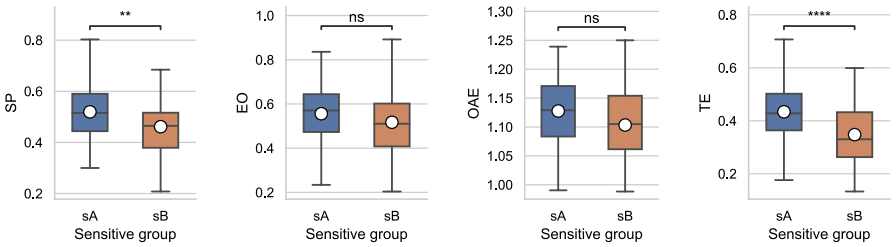
This experiment analyses the potential benefit of adopting multigroup metrics over binary metrics for a proper fairness assessment. In this scenario, the target class is binary, and we evaluate the differences between analyzing binary or multigroup sensitive attributes. For each of the two models, CatGCN and RHGN, we first run the user modeling task (i.e., classification of *bin-buy* class for the Alibaba dataset, *bin-exp* for the JD dataset, *bin-work-field* for the Pokec dataset, and *bin-salary* for the NBA dataset), then we computed the scores of the binary fairness metrics, defined by Eqs. (1)–(4), and multigroup fairness metrics, defined by Eqs. (9)–(12).

To fortify the credibility of our findings and ensure that the differences observed are not products of random chance, we employed a *Mann–Whitney-Wilcoxon*[7] (Mann & Whitney, 1947; Wilcoxon, 1992) test on every couple of groups. We implemented

[7] The Mann–Whitney-Wilcoxon test is a non-parametric test used to assess whether two independent samples come from the same distribution. Unlike the t-test, it does not require the assumption of normal distribution, making it a more flexible and reliable choice for our data's distribution characteristics.

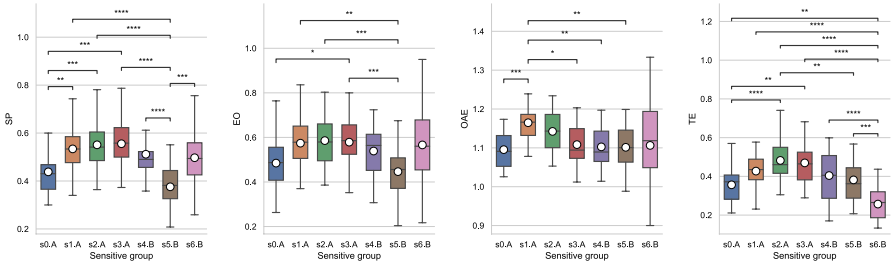
Table 8 Description of the cases derived from the assessment of the comparative results between *binary* and *multigroup* scenarios

#	Binary scenario	Multigroup scenario
1	Binarized group advantaged	Related fine-grained original groups (all or some) disadvantaged
2	Binarized group advantaged	Opposite (i.e., belonging to the other binarized group) fine-grained original groups (all or some) advantaged
3	Fair result	Some fine-grained groups particularly disadvantaged (or advantaged to the detriment of others)



(a) SP (b) EO (c) OAE (d) TE

Fig. 1 Fairness assessment of **CatGCN** model on **Alibaba** dataset in the binary class (positive output) and binary group scenario

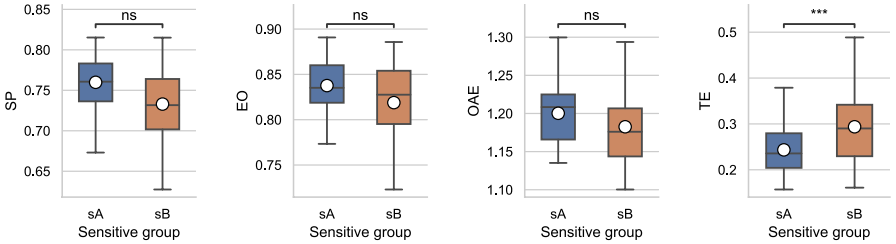


(a) SP (b) EO (c) OAE (d) TE

Fig. 2 Fairness assessment of **CatGCN** model on **Alibaba** dataset in the binary class (positive output) and multigroup scenario

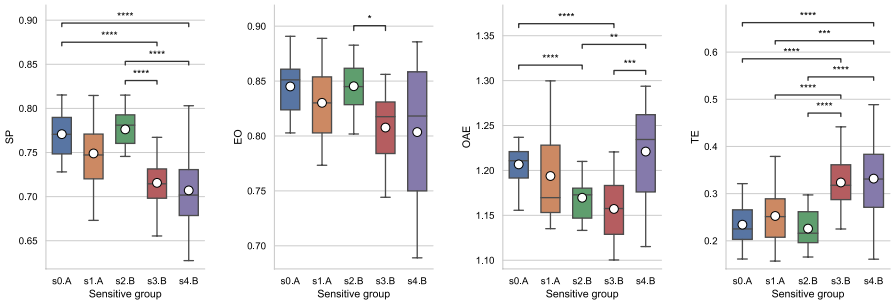
the statistical test with 1000 repetitions, ensuring the stability and replicability of our results. Moreover, we applied the *Bonferroni* correction[8] (Haynes, 2013), a

[8] The Bonferroni correction adjusts the significance thresholds to account for the increased chance of observing significant results purely by chance when multiple tests are performed simultaneously. By dividing the desired significance level by the number of comparisons made, the Bonferroni correction safeguards against the risk of false positives, thus ensuring that the differences we report are indeed statistically significant and not a result of random variation or the sheer number of tests conducted.



(a) SP (b) EO (c) OAE (d) TE

Fig. 3 Fairness assessment of **CatGCN** model on **JD** dataset in the binary class (positive output) and binary group scenario



(a) SP (b) EO (c) OAE (d) TE

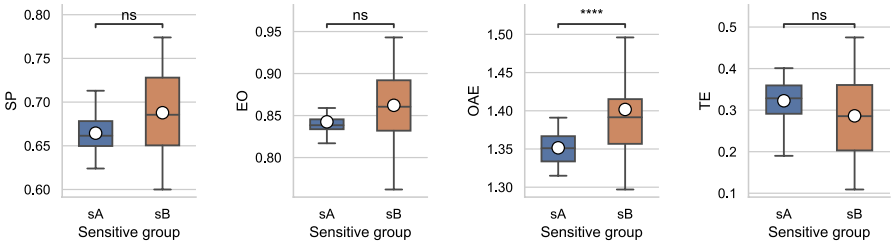
Fig. 4 Fairness assessment of **CatGCN** model on **JD** dataset in the binary class (positive output) and multigroup scenario

conservative statistical approach designed to counteract the problem of multiple comparisons.

Table 7 displays a comprehensive qualitative analysis of the comparative results between binary and multigroup scenarios. In particular, for each combination of dataset, metric, model, and setting (binary or multigroup), we reported, for the multigroup case, the differences from the related binary case.

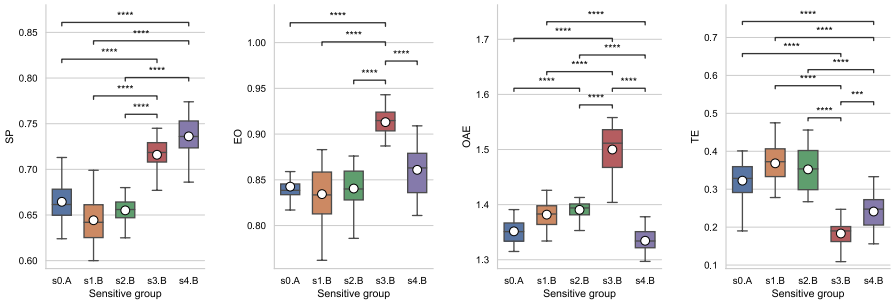
The experimental results for each combination of the earlier-mentioned model, dataset, and scenario are shown in Figs. 1–16. In the presented charts, each pair of box plots is annotated with a statistical symbol, reflecting the statistical significance of the difference between the two groups under comparison based on the p-value, a measure of the strength of the evidence against the null hypothesis. Specifically:

- A notation of [ns] indicates a non-significant difference, suggesting the evidence is not strong enough to reject the null hypothesis for the difference between the groups, implying that the observed difference could be due to random chance rather than systematic unfairness.
- Symbols ranging from [$*$] to [$****$] denote increasing levels of statistical significance, associated with decreasing p-values. A statistically significant difference



(a) SP (b) EO (c) OAE (d) TE

Fig. 5 Fairness assessment of **CatGCN** model on **Pokec** dataset in the binary class (positive output) and binary group scenario



(a) SP (b) EO (c) OAE (d) TE

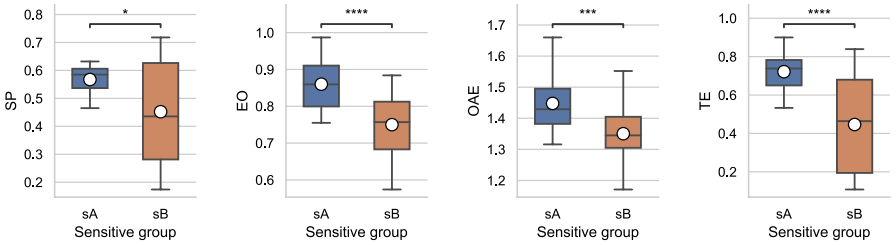
Fig. 6 Fairness assessment of **CatGCN** model on **Pokec** dataset in the binary class (positive output) and multigroup scenario

implies that the likelihood of the observed data occurring under the null hypothesis is low. This significant difference is indicative of real, consistent disparities between the groups, and in the context of our study, it points toward the presence of unfairness. The specific levels of significance are:

[*]: Significant difference with a p-value less than 0.05 but greater than 0.01.
[**]: More significant difference with a p-value less than 0.01 but greater than 0.001.
[***]: Highly significant difference with a p-value less than 0.001 but greater than 0.0001.
[****]: Extremely significant difference with a p-value less than 0.0001.

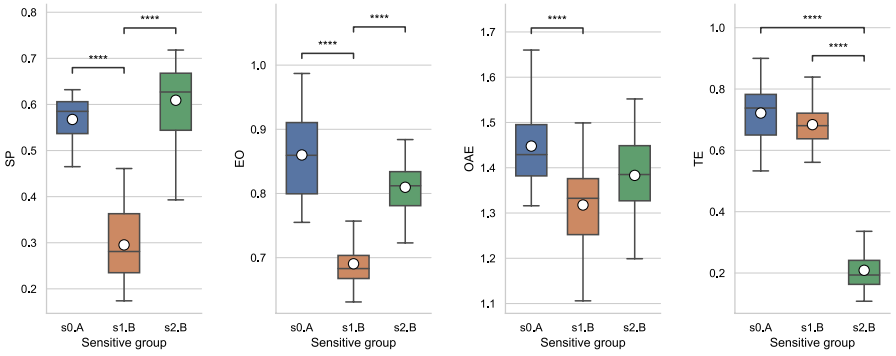
From the assessment, we identified three specific crucial cases from the analysis of the experiment results in this context, which are shown and described in Table 8.

To give some practical evidence of what exactly these results can tell us, we can consider the experiments involving CatGCN model on JD dataset (i.e., expense level as the target class for classification and age as the sensitive



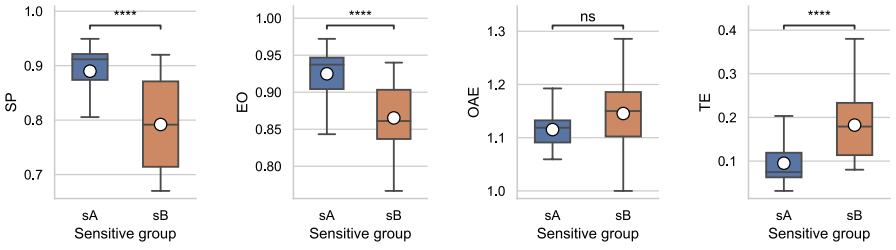
(a) SP (b) EO (c) OAE (d) TE

Fig. 7 Fairness assessment of **CatGCN** model on **NBA** dataset in the binary class (positive output) and binary group scenario



(a) SP (b) EO (c) OAE (d) TE

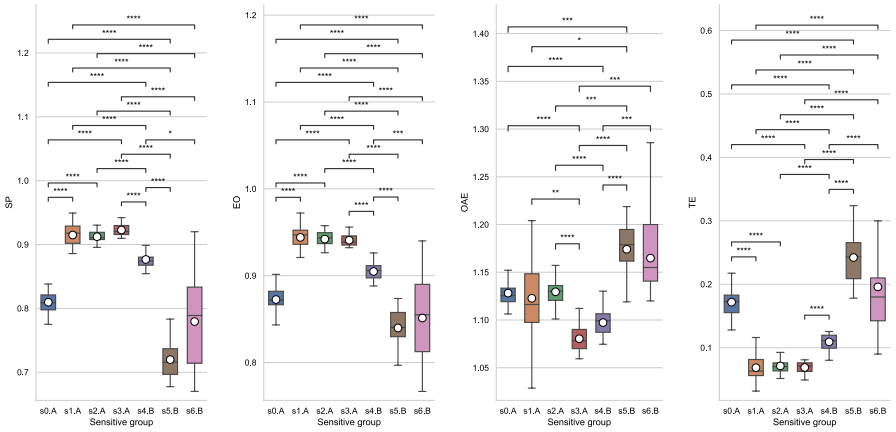
Fig. 8 Fairness assessment of **CatGCN** model on **NBA** dataset in the binary class (positive output) and multigroup scenario



(a) SP (b) EO (c) OAE (d) TE

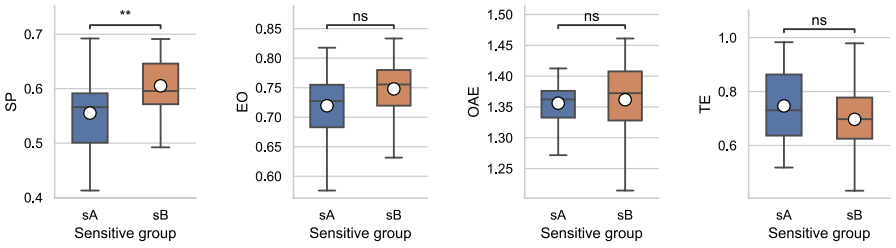
Fig. 9 Fairness assessment of **RHGN** model on **Alibaba** dataset in the binary class (positive output) and binary group scenario

attribute) and related to Figs. 3a–4a and Figs. 3d–4d. In the former, under *statistical parity* constraints, the binary age groups are fair, and no mitigation is needed; with a fine-grained multigroup analysis, instead, two age subgroups are exposed as disadvantaged, and an intervention can be planned to address inequities. In the latter instance, given the *treatment equality* scores, the results show a



(a) SP (b) EO (c) OAE (d) TE

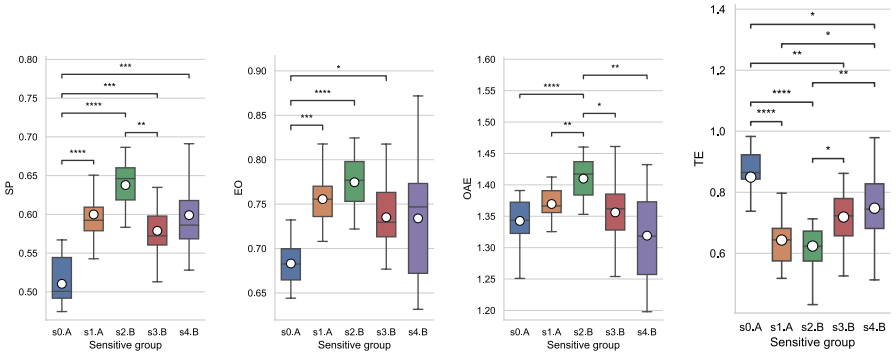
Fig. 10 Fairness assessment of **RHGN** model on **Alibaba** dataset in the binary class (positive output) and multigroup scenario



(a) SP (b) EO (c) OAE (d) TE

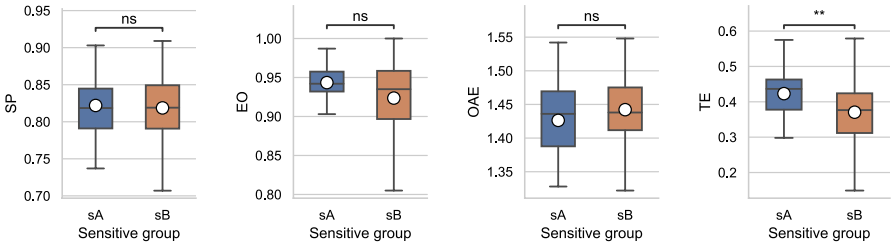
Fig. 11 Fairness assessment of **RHGN** model on **JD** dataset in the binary class (positive output) and binary group scenario

binary age group as advantaged, but the detailed multigroup assessment reveals that within that group, there is a specific subgroup that is disadvantaged; in this case, applying a bias mitigation procedure in the binary scenario would even worsen the discrimination toward this age subgroup.



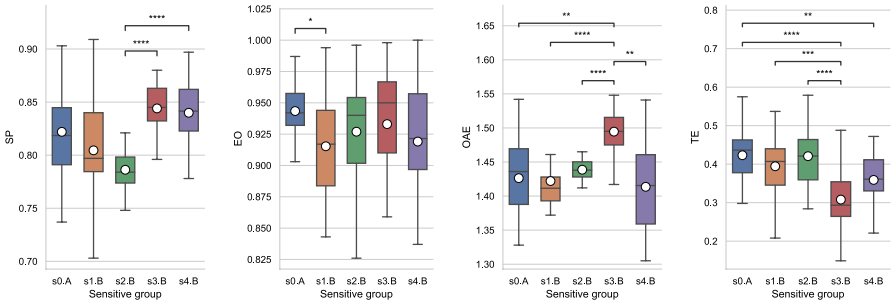
(a) SP (b) EO (c) OAE (d) TE

Fig. 12 Fairness assessment of **RHGN** model on **JD** dataset in the binary class (positive output) and multigroup scenario



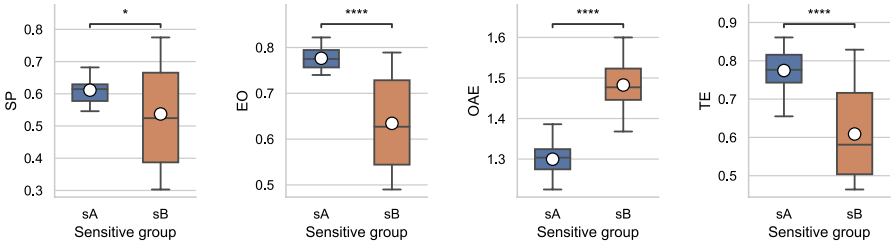
(a) SP (b) EO (c) OAE (d) TE

Fig. 13 Fairness assessment of **RHGN** model on **Pokec** dataset in the binary class (positive output) and binary group scenario



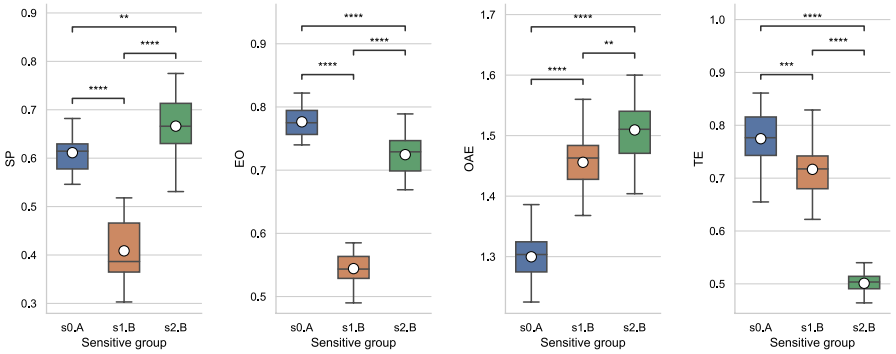
(a) SP (b) EO (c) OAE (d) TE

Fig. 14 Fairness assessment of **RHGN** model on **Pokec** dataset in the binary class (positive output) and multigroup scenario



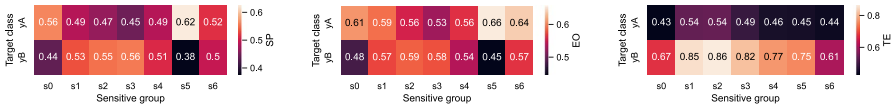
(a) SP (b) EO (c) OAE (d) TE

Fig. 15 Fairness assessment of **RHGN** model on **NBA** dataset in the binary class (positive output) and binary group scenario



(a) SP (b) EO (c) OAE (d) TE

Fig. 16 Fairness assessment of **RHGN** model on **NBA** dataset in the binary class (positive output) and multigroup scenario

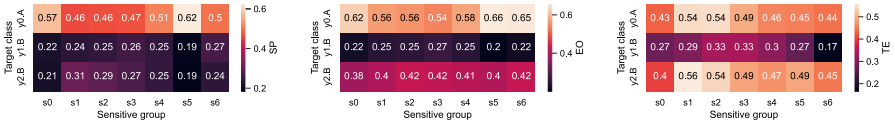


(a) SP (b) EO (c) TE

Fig. 17 Fairness assessment of **CatGCN** model on **Alibaba** dataset in the binary class (both outputs) and multigroup scenario

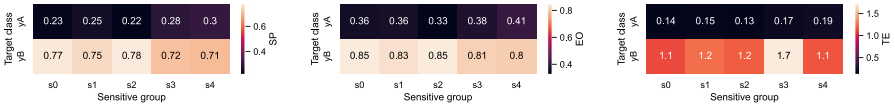
Observation 1

A fine-grained fairness analysis, leveraging multigroup metrics, allows the discovery of actual discrimination among sensitive groups hidden by a binary assessment, either in situations where biases do not seem to exist or where a disadvantaged group is wrongly deemed as advantaged.



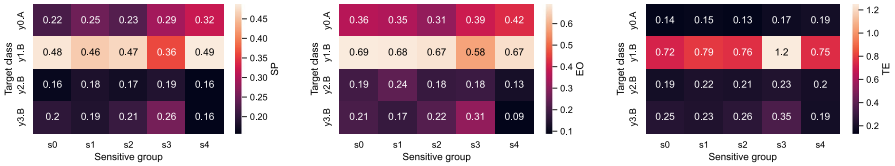
(a) SP (b) EO (c) TE

Fig. 18 Fairness assessment of **CatGCN** model on **Alibaba** dataset in the multiclass and multigroup scenario



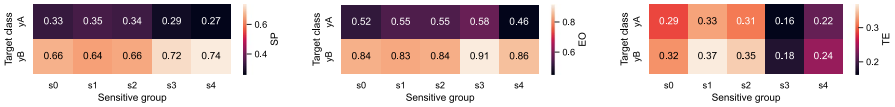
(a) SP (b) EO (c) TE

Fig. 19 Fairness assessment of **CatGCN** model on **JD** dataset in the binary class (both outputs) and multigroup scenario



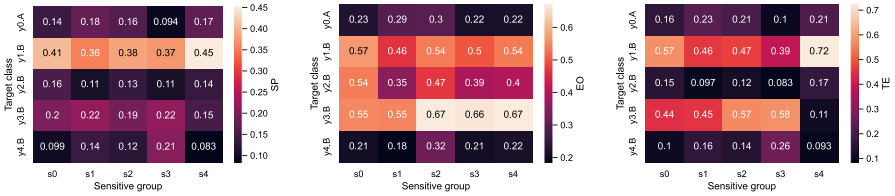
(a) SP (b) EO (c) TE

Fig. 20 Fairness assessment of **CatGCN** model on **JD** dataset in the multiclass and multigroup scenario



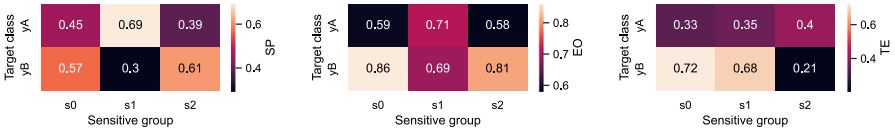
(a) SP (b) EO (c) TE

Fig. 21 Fairness assessment of **CatGCN** model on **Pokec** dataset in the binary class (both outputs) and multigroup scenario



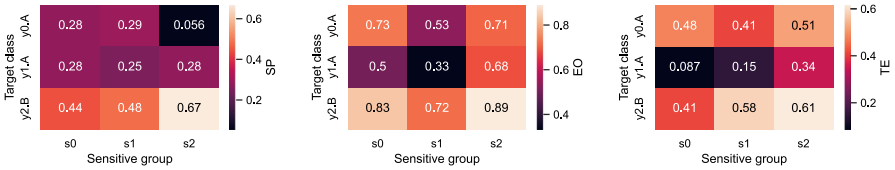
(a) SP (b) EO (c) TE

Fig. 22 Fairness assessment of **CatGCN** model on **Pokec** dataset in the multiclass and multigroup scenario



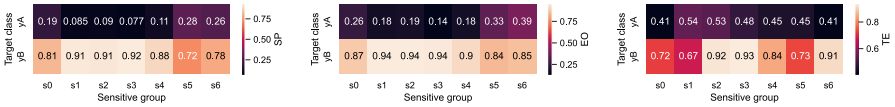
(a) SP (b) EO (c) TE

Fig. 23 Fairness assessment of **CatGCN** model on **NBA** dataset in the binary class (both outputs) and multigroup scenario



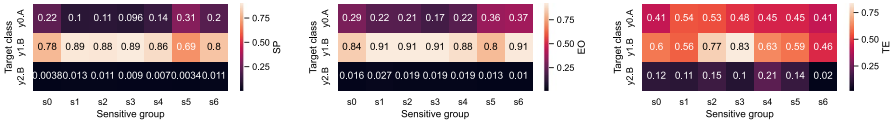
(a) SP (b) EO (c) TE

Fig. 24 Fairness assessment of **CatGCN** model on **NBA** dataset in the multiclass and multigroup scenario



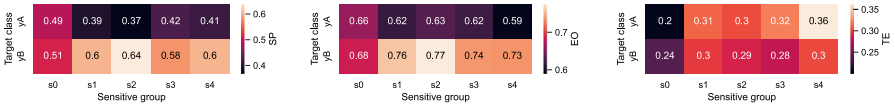
(a) SP (b) EO (c) TE

Fig. 25 Fairness assessment of **RHGN** model on **Alibaba** dataset in the binary class (both outputs) and multigroup scenario



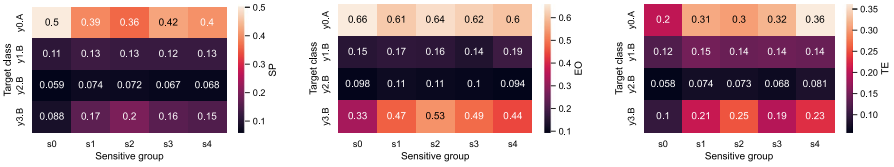
(a) SP (b) EO (c) TE

Fig. 26 Fairness assessment of **RHGN** model on **Alibaba** dataset in the multiclass and multigroup scenario



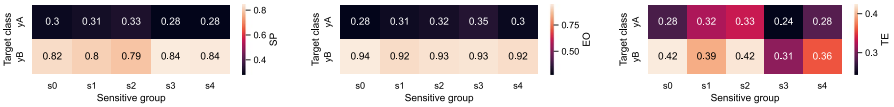
(a) SP (b) EO (c) TE

Fig. 27 Fairness assessment of **RHGN** model on **JD** dataset in the binary class (both outputs) and multigroup scenario



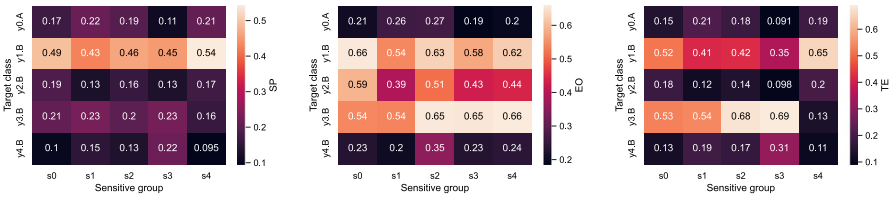
(a) SP (b) EO (c) TE

Fig. 28 Fairness assessment of **RHGN** model on **JD** dataset in the multiclass and multigroup scenario



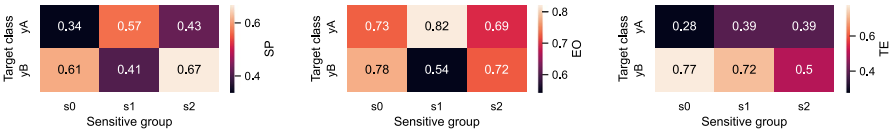
(a) SP (b) EO (c) TE

Fig. 29 Fairness assessment of **RHGN** model on **Pokec** dataset in the binary class (both outputs) and multigroup scenario



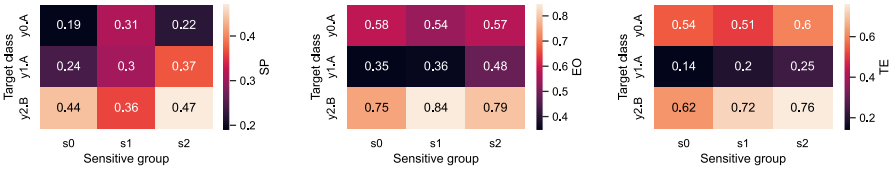
(a) SP (b) EO (c) TE

Fig. 30 Fairness assessment of **RHGN** model on **Pokec** dataset in the multiclass and multigroup scenario



(a) SP (b) EO (c) TE

Fig. 31 Fairness assessment of **RHGN** model on **NBA** dataset in the binary class (both outputs) and multigroup scenario



(a) SP (b) EO (c) TE

Fig. 32 Fairness assessment of **RHGN** model on **NBA** dataset in the multiclass and multigroup scenario

Table 9 Qualitative analysis of the comparative results between *multigroup* and *multiclass* scenarios leading to the considerations for RQ2

Dataset	Metric	Model			
		CatGCN	Ref. Figures	RHGN	Ref. Figures
Alibaba	SP	†, ⋄	(17a, 18a)	∗	(25a, 26a)
	EO	⋄	(17b, 18b)	∗	(25b, 26b)
	TE	∗	(17c, 18c)	∗	(25c, 26c)
JD	SP	∗	(19a, 20a)	†, ⋄	(27a, 28a)
	EO	∗	(19b, 20b)	†, ⋄	(27b, 28b)
	TE	∗	(19c, 20c)	⋄	(27c, 28c)
Pokec	SP	∗,⩓	(21a, 22a)	∗	(29a, 30a)
	EO	∗,⩓	(21b, 22b)	∗	(29b, 30b)
	TE	⊙	(21c, 22c)	∗,⩓	(29c, 30c)
NBA	SP	†	(23a, 24a)	☆	(31a, 32a)
	EO	⋄	(23b, 24b)	†, ⋄	(31b, 32b)
	TE	†	(23c, 24c)	–	(31c, 32c)

The symbols refer to the derived cases described in Table 10

6.2.2 Assessing Multiclass and Multigroup Fairness Metrics in Real-World Cases (RQ2)

Most of the standard (binary) fairness metrics rely on the selection of a *positive* class for their computation. As already discussed in this article, when a binarization of an originally multiclass target variable is applied, like in Dai and Wang (2021), such a selection is made almost randomly because neither of the two classes can be really considered the positive one. In this experiment, our goal is to examine how a multiclass and multigroup fairness assessment should always be preferred to have a clear picture of all possible discrimination created by the models. We perform the same

Table 10 Description of the cases derived from the assessment of the comparative results between *multigroup* and *multiclass* scenarios

#	Symbol	Multigroup scenario	Multigroup and multiclass scenario
1	†	Binarized class advantaged	All related fine-grained classes significantly disadvantaged
2	⋄	Fair result	All fine-grained classes, belonging to the same binarized class, disadvantaged
3	∗	Binarized class advantaged	Only one or a few of the related fine-grained classes significantly disadvantaged
4	⊙	Fair result	Only one or a few of the related fine-grained classes, belonging to the same binarized class, disadvantaged
5	☆	Unfair result	Fair result
6	⩓	Binarized class disadvantaged	Even greater unfairness against that class (and related fine-grained classes)

profiling task described in the previous section and compute the metrics defined in Eqs. (13)-(16).

In Figs. 17–32, the results for each combination of model, dataset, and metrics are displayed. Concerning the metrics, the multiclass and multigroup *OAE* (Eq. (15)) are not taken into account in this evaluation because the results would have been identical to those in the previous experiments, due to its definition, which sums up the probabilities of the target classes. For the sake of clarity, we only show the mean values without other stats in the result charts. As for the previous evaluation, how the single classes are binarized is illustrated in Sect. 3.3.

Similarly to the binary-multigroup scenario (Sect. 6.2.1), in Table 9, we provided an effective way to visualize the outcomes of the qualitative analysis of the comparative results between multigroup and multiclass scenarios. For each combination of dataset, metric, and model, the associated findings are represented within the table as symbols, each of them referring to one of the six specific cases we derived from the analysis of the experiment results in this context, and illustrated in Table 10.

In order to demonstrate the practical implications of these findings, we can examine the experiments conducted with the CatGCN model on the Alibaba dataset (i.e., consumption grade as the target class for classification and age as sensitive attribute) and corresponding to Figs. 17a and 18a. Computing the fairness scores with the *statistical parity* metric, we face two different situations. Given that in this evaluation the sensitive attribute is always considered as multi-valued, for a particular age group, we have, on one side, the mid and high consumption levels deemed as advantaged when taken together in the binary evaluation, while disadvantaged when individually considered for the multiclass assessment; on the other side, for another age group, the binary results are fair, but the detailed analysis reveals all that the mid and high consumption levels are again disadvantaged when considered separately.

Observation 2 In contexts where there is not an actual "positive" class, the exploitation of multiclass and multigroup fairness metrics to evaluate both binary and multiclass scenarios provides the actual picture of models' discrimination, providing a clear understanding of which are the discriminated groups across all the classes and what is the fairness difference between them in order to design a proper bias mitigation strategy.

7 Conclusion and Future Work

In this article, we presented a novel responsible and ethical approach for algorithmic fairness assessment, including analyzing real-world scenarios, introducing *multigroup* and *multiclass* metrics, and evaluating them in real-world user modeling tasks leveraging state-of-the-art GNN-based models. Starting from ethical implications derived from the common practice of assessing fairness in binary scenarios and, in particular, from a recent fairness analysis of GNNs designed for user modeling (Purificato et al., 2022), we extended the definition of four different existing fairness binary metrics related to disparate treatment and disparate mistreatment

notion (i.e. statistical parity, equal opportunity, overall accuracy equality, and treatment equality) to a multigroup and multiclass scenario, as the principal contribution of this article. The presented evaluation considered user modeling tasks on four real-world datasets and the exploitation of two state-of-the-art GNN-based models, namely CatGCN and RHGN. In this study, we aimed to evaluate the impact of utilizing a finer level of granularity in assessing fairness within multiple sub-populations on the detection of undetected or incorrectly assumed discrimination across distinct minority groups. Our findings demonstrate that employing multigroup measures in evaluating fairness facilitates the identification of unfair treatment among vulnerable populations despite initial impressions of equity or instances of mistaken advantageous outcomes for certain underprivileged cohorts. Ultimately, our results highlight the importance of considering nuanced perspectives when examining bias in order to ensure the most accurate representation of systemic issues affecting marginalized communities. Moreover, when dealing with circumstances involving no true beneficial outcome category, our research uses multiparty and multiplex metrics to appraise binary and multiclass situations. This approach presents a comprehensive view of any model's prejudice, delineating the affected groups throughout the various classes while measuring their diversity gaps. With these findings, informed strategies can then be developed to alleviate and eliminate unwarranted biases within such complex systems. We also acknowledged the necessity of a focused investigation into the relationships between model-dataset combinations and fairness outcomes. In future studies, along with exploring additional models not limited to GNNs, we will also extend our research to unravel these interactions, particularly examining the impact of dataset characteristics and the construction of binary groups/classes on fairness scores. This will involve a rigorous analysis to understand the nuances and potential correlations, further enriching our comprehension of fairness in automated decision-making systems.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Alghamdi, W., Hsu, H., Jeong, H., Wang, H., Michalak, P. W., Asoodeh, S., & Calmon, F. P. (2022). Beyond adult and compas: Fairness in multi-class prediction. arXiv preprint arXiv:2206.07801

Barocas, S., Hardt, M., & Narayanan. A. (2019). *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*, 671.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research, 50*(1), 3–44.

Binns, R. (2020). On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 514–524).

Blakeney, C., Atkinson, G., Huish, N., Yan, Y., Metsis, V., & Zong, Z. (2022). Measuring bias and fairness in multiclass classification. In *2022 IEEE International Conference on Networking, Architecture and Storage (NAS)* (pp. 1–6). IEEE.

Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys (CSUR), 56(7)*, 1–38.

Chen, W., Feng, F., Wang, Q., He, X., Song, C., Ling, G., & Zhang, Y. (2023). apr. Catgcn: Graph convolutional networks with categorical node features. *IEEE Transactions on Knowledge and Data Engineering, 35*(4), 3500–3511. https://doi.org/10.1109/TKDE.2021.3133013

Chen, W., Gu, Y., Ren, Z., He, X., Xie, H., Guo, T., Yin, D., & Zhang, Y. (2019). Semi-supervised user profiling with heterogeneous graph attention networks. In *Proceedings of the 28th international joint conference on artificial intelligence* (pp. 2116–2122).

Chierichetti, F., Kumar, R., Lattanzi, S., & Vassilvtiskii, S. (2019). Matroids, matchings, and fairness. In *The 22nd international conference on artificial intelligence and statistics* (pp. 2212–2220). PMLR.

Cui, H., Lu, J., Ge, Y., & Yang, C. (2022). How can graph neural networks help document retrieval: A case study on cord19 with concept map generation. In *European conference on information retrieval* (pp. 75–83). Springer.

Dai, E., & 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* (pp. 680–688).

Denis, C., Elie, R., Hebiri, M., & Hu, F. (2021). Fairness guarantee in multi-class classification. arXiv preprint arXiv:2109.13642.

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way* (Vol. 1). Springer.

Dong, Y., Kang, J., Tong, H., & Li, J. (2021). Individual fairness for graph neural networks: A ranking based approach. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining* (pp. 300–310).

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214–226).

Eke, C. I., Norman, A. A., Shuib, L., & Nweke, H. F. (2019). A survey of user profiling: State-of-the-art, challenges, and solutions. *IEEE Access, 7*, 144907–144924.

Ekstrand, M. D., Das, A., Burke, R., & Diaz, F. (2022). Fairness in information access systems. *Foundations and Trends® in Information Retrieval, 16*(1–2), 1–177.

European-Commission. (2019). *Ethics guidelines for trustworthy AI*. Publications Office.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* ((pp. 259–268).

Gao, R., & Shah, C. (2019). How fair can we go: Detecting the boundaries of fairness optimization in information retrieval. In *Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval* (pp. 229–236).

Gómez, E., Shui Zhang, C., Boratto, L., Salamó, M., & Marras, M. (2021). The winner takes it all: Geographic imbalance and provider (un) fairness in educational recommender systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (pp. 1808–1812).

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2125–2126).

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems, 30*, 1024–1034.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems, 29*, 3315–3323.

Haynes, W. (2013). *Bonferroni correction*. Springer.

He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., & 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* (pp. 639–648).

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *5th international conference on learning representations, ICLR 2017, conference track proceedings*.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. *In AEA Papers and Proceedings, 108*, 22–27.

Leonhardt, J., Anand, A., & Khosla, M. (2018). User fairness in recommender systems. In *Companion Proceedings of the Web Conference 2018* (pp. 101–102).

Li, R., Wang, S., Deng, H., Wang, R., & Chang, K. C. C. (2012). Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1023–1031).

Loveland, D., Pan, J., Bhathena, A. F., & Lu, Y. (2022). Fairedit: Preserving fairness in graph neural networks through greedy graph editing. arXiv preprint arXiv:2201.03681.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics, 18(1)*, 50–60.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR), 54*(6), 1–35.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application, 8*, 141–163.

Nilashi, M., Rupani, P. F., Rupani, M. M., Kamyab, H., Shao, W., Ahmadi, H., Rashid, T. A., & Aljojo, N. (2019). Measuring sustainability through ecological sustainability and human sustainability: A machine learning approach. *Journal of Cleaner Production, 240*, 118162.

Pessach, D., & Shmueli, E. (2020). Algorithmic fairness. arXiv preprint arXiv:2001.09784.

Poo, D., Chng, B., & Goh, J. M. (2003). A hybrid approach for user profiling. In *Proceedings of the 36th annual Hawaii international conference on system sciences* (pp. 9–13). IEEE.

Purificato, E., Boratto, L., & De Luca, E. W. (2022). Do graph neural networks build fair user models? assessing disparate impact and mistreatment in behavioural user profiling. In *Proceedings of the 31st ACM international conference on information & knowledge management* (pp. 4399–4403).

Purificato, E., Boratto, L., & De Luca, E. W. (2024). User Modeling and User Profiling: A Comprehensive Survey. arXiv preprint arXiv:2402.09660.

Purificato, E., Lorenzo, F., Fallucchi, F., & Luca, E. W. D. (2023, Apr). The use of responsible artificial intelligence techniques in the context of loan approval processes. *International Journal of Human-Computer Interaction*, 1543–1562. https://doi.org/10.1080/10447318.2022.2081284 .

Purificato, E., Wehnert, S., & De Luca, E. W. (2021). Dynamic privacy-preserving recommendations on academic graph data. *Computers, 10*(9), 107.

Putzel, P., & Lee, S. (2022). Blackbox post-processing for multiclass fairness. In *Proceedings of the workshop on artificial intelligence safety 2022 (SafeAI 2022) co-located with the thirty-sixth AAAI conference on artificial intelligence (AAAI 2022)*, CEUR-WS (Vol. 3087).

Rahimi, A., Cohn, T., & Baldwin, T. (2018). Semi-supervised user geolocation via graph convolutional networks. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (Vol. 1: Long Papers, pp. 2009–2019).

Ramos, G., & Boratto, L. (2020). Reputation (in)dependence in ranking systems: Demographics influence over output disparities. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, SIGIR 2020* (pp. 2061–2064). ACM.

Ramos, G., Boratto, L., & Caleiro, C. (2020). On the negative impact of social influence in recommender systems: A study of bribery in collaborative hybrid algorithms. *Information Processing and Management, 57*(2), 10205. https://doi.org/10.1016/j.ipm.2019.102058

Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.

Takac, L., & Zabovsky, M. (2012). Data analysis in public social networks. In *International scientific conference and international workshop present day trends of innovations, 1(6)*, 1–6.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. arXiv preprint arXiv:1710.10903.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *IEEE/ACM international workshop on software fairness (FairWare 2018)* (pp. 1–7). IEEE.

Wan, M., Zha, D., Liu, N., & Zou, N. (2021). Modeling techniques for machine learning fairness: A survey. arXiv preprint arXiv:2111.03015.

Wang, Q., Ming, Y., Jin, Z., Shen, Q., Liu, D., Smith, M.J., Veeramachaneni, K., & Qu, H. (2019). Atmseer: Increasing transparency and controllability in automated machine learning. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–12).

Wilcoxon, F. (1992). *Individual comparisons by ranking methods, breakthroughs in statistics: Methodology and distribution, 196–202*. Springer.

Yan, Q., Zhang, Y., Liu, Q., Wu, S., & Wang, L. (2021, Oct). Relation-aware heterogeneous graph for user profiling. In *Proceedings of the 30th ACM international conference on information & knowledge management* (pp. 3573–3577). Association for Computing Machinery.

Yao, L., Mao, C., & Luo, Y. (2019). Graph convolutional networks for text classification. *In Proceedings of the AAAI Conference on Artificial Intelligence, 33*, 7370–7377.

Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W. L., & Leskovec, J. (2018). Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 974–983).

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on World Wide Web* (pp. 1171–1180).

Zhang, C., Song, D., Huang, C., Swami, A., & Chawla, N. V. (2019). Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 793–803).

Authors and Affiliations

Erasmo Purificato[1,3] · Ludovico Boratto[2] · Ernesto William De Luca[1,3]

✉ Erasmo Purificato
erasmo.purificato@acm.org

Ludovico Boratto
ludovico.boratto@acm.org

Ernesto William De Luca
ernesto.deluca@ovgu.de

[1] Otto von Guericke University Magdeburg, Magdeburg, Germany

[2] University of Cagliari, Cagliari, Italy

[3] Leibniz Institute for Educational Media | Georg Eckert Institute, Brunswick, Germany