Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

A Probabilistic Greedy Attempt to be Fair in Neural Team Recommendation

Hamed Loghmani | **Mahdis Saeedi** | **Gabriel Rueda** | **Edwin Paul** | **Hossein Fani**

[1] School of Computer Science, University of Windsor, Ontario, Canada

Correspondence
Corresponding author Hossein Fani,
Email: hfani@uwindsor.ca

Abstract

Neural team recommendation has brought state-of-the-art efficacy while enhancing efficiency at forming teams of experts whose success in completing complex tasks is *almost surely* guaranteed; albeit they overlook fairness, that is, predicted teams are heavily biased toward popular and male experts, falling short of recommending female or *non*popular experts. In this work, we introduce and formalize the *fair team recommendation* problem in view of group-based notions of fairness. We further propose a probabilistic greedy reranking algorithm to achieve fairness with respect to popularity or gender biases in neural models with respect to demographic parity and equality of opportunity. Specifically, we aim to ensure a minimum representation of experts from the disadvantaged nonpopular or female groups by reranking the neural model's ranked list of recommended experts. Our experiments on three large-scale benchmark datasets demonstrate: (1) neural team recommenders heavily suffer from biases toward popular and male experts; (2) our reranking method can substantially mitigate such biases while maintaining teams' efficacy; (3) in the presence of extreme biases, post-processing reranking methods alone fall short, urging further tandem integration of pre-process and in-process debiasing techniques. The code to reproduce the experiments reported in this paper is available at `https://github.com/fani-lab/Adila`.

KEYWORDS
Fair Team Recommendation, Neural Team Recommendation, Social Information Retrieval.

1 | INTRODUCTION

As modern tasks have surpassed the capacity of individuals, forming teams of experts whose collaboration for a common goal yields success has been a surge of research interest in many disciplines, including psychology[1,2], the science of team science (SciTS)[3], and industrial engineering[4]. Forming teams can be seen as social information retrieval (Social IR) where the right group of experts are searched and hired to solve the task at hand. Traditionally, teams were formed manually by relying on human experience and instinct; a tedious, error-prone, and suboptimal process for an overwhelming number of experts, a multitude of objectives to optimize (e.g., budget, time and team size constraints), and hidden personal and societal biases, among other reasons. As a result, a rich body of various computational methods, from operations research[5,6,7,8,9,10,11,12,13,14,15] , social network analysis[16,17,18,19], and recently, machine learning [20,21,22,23,24,25,26,27,28] have been proposed. Specifically, neural models learn the distributions of experts and their skill sets in the context of successful and unsuccessful teams from training datasets to recommend future teams that are *almost surely* successful. Such models have brought state-of-the-art efficacy while enhancing efficiency, taking the stage and becoming canonical in team recommendation literature.

The primary focus of existing team recommenders is, however, the maximization of the models' accuracy (utility), largely ignoring the fairness in their ranked list of recommended experts, leading to discrimination, reduced visibility for already disadvantaged experts, and gender disparities[29,30,31]. These unfair biases, far from being random, originate mainly from training datasets. As seen in Figure 1, such datasets are segmented toward male experts, and females are heavily underrepresented, like in the `dblp` dataset of computer research articles with %86 male vs. %14 female researchers. Also, from Figure 2, datasets in team recommendation suffer from popularity bias; that is, the majority of *non*popular experts have scarcely participated in the

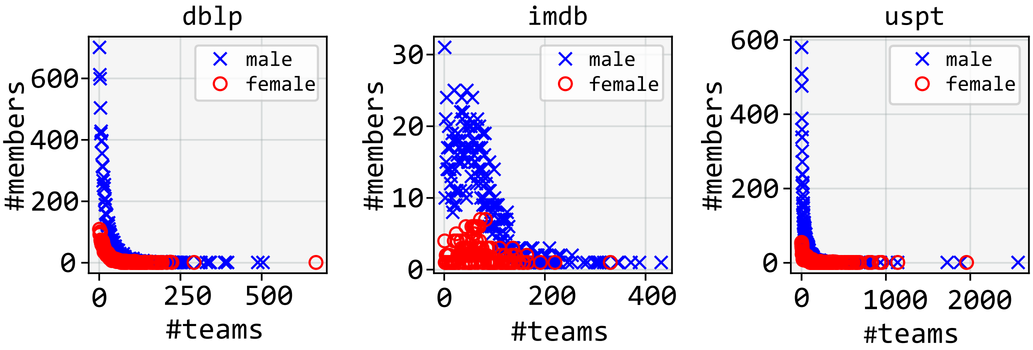

FIGURE 1 Distribution of experts in terms of gender in `dblp`, `imdb`, and `uspt` datasets.

(successful) teams, whereas few popular experts are in many teams. Therefore, popular or male experts would receive more attention and are more frequently recommended by a machine learning model, leading to systematic discrimination against already disadvantaged nonpopular or female experts. To the best of our knowledge, there is no fairness-aware approach in neural team recommendation methods except that of Loghmani et al.[32], who applied deterministic greedy reranking algorithms to mitigate popularity bias and showed such deterministic methods can mitigate bias but at the cost of a substantial drop in the models' accuracy.

In this paper, foremost, we introduce and formalize the *fair team recommendation* problem to foster standards and conventions, which the literature on the team recommendation problem lacks. We set forth a unified set of notations to define the problem in view of group-based notions of fairness, including demographic (statistical) parity[33], equalized odds[34], and equality of opportunity[34]. We further incorporate the notions of fairness in tandem with experts' skills in team recommendations to facilitate recommending merit-based teams while equal opportunity is also maximized. Specifically, we propose a probabilistic fairness-aware *re*ranking method to adjust the ordering of experts in the final ranked list of recommended experts to address potential biases and promote fairness concerning gender or popularity biases. As opposed to pre-processing-based methods, which modify data or its labels before model training, or in-processing techniques, which focus on balancing model accuracy with fairness considerations during training, our method is model-agnostic and belongs to *post-processing* category of methods, which seek to improve the fairness of model's outputs after training, without adjustments to the data, training procedure, or the model's architecture. Moreover, being probabilistic, our approach holds advantages over deterministic methods for managing real-world uncertainties. Instead of providing rigid decisions, our approach offers distributions over possible outcomes, resulting in more adaptive solutions. To illustrate the effectiveness of our proposed approach, we perform experiments on three large-scale benchmark datasets of computer science articles (`dblp`)[35,18], moving pictures (`imdb`)[36,16], and US patents (`uspt`)[37]. Our results show that our proposed approach substantially mitigates popularity bias while maintaining the success rate of the recommended teams. With respect to gender bias, however, our approach's impact has been marginal due to the highly sparse distribution of female experts in the training datasets (`%14`, `%12`, and `%14` in `dblp`, `imdb` and `uspt`, respectively), urging further future studies on integration of pre-process and in-process debiasing techniques.

In summary, our key contributions are as follows:

1. We defined the problem of fair team recommendation in view of group-based notions of fairness including demographic (statistical) parity, equalized odds, and equality of opportunity.
2. We proposed a model-agnostic post-processing and probabilistic reranking method to mitigate unfair biases in the recommended teams of experts by neural team recommendation models.
3. We demonstrated the performance of our proposed method in the presence of gender or popularity biases with respect to demographic parity and equality of opportunity on three large-scale datasets from different domains.
4. We developed an open-source reproducible framework hosting canonical neural models as the cutting-edge class of approaches, along with large-scale training datasets from varying domains that integrated our proposed and baseline debiasing reranking algorithms.

Our work addresses the ever-growing need to identify and facilitate successful yet diverse teamwork based on merit while fairness is also maximized, which is one of the pillars of growth in scientific and industrial communities. Employers will be

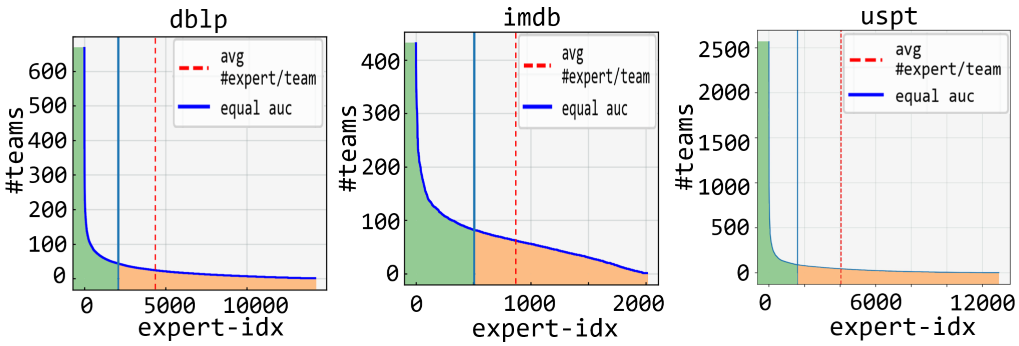

FIGURE 2 Distribution of experts in terms of popularity in dblp, imdb, and uspt datasets. While defining popularity can be controversial, recommender system literature follows *sociometric* popularity[38], where items (herein, experts) of the *head* in the participation distribution are labeled as popular items. As seen, there are two alternatives to split the distribution into *head* and *tail* parts: *i*) the average number of teams per expert over the entire dataset, or *ii*) equal area under the curve (auc). In this paper, we opt for the former. For the dblp dataset, the average stands at 23.02 teams, 62.45 teams for the imdb dataset, and 44.69 teams for the uspt dataset. Therefore, in dblp, the proportion of popular to nonpopular experts becomes 0.313 to 0.687, in imdb, it is 0.426 to 0.574 and for uspt it stands at 0.314 to 0.686.

able to identify highly-skilled, diverse workers to fill labour gaps and increase innovation. As AI-based solutions are making notable impacts on how job opportunities are allocated to various groups in society, systematic consideration of fairness in this process is key. The rest of the paper is organized as follows: we first present the related works in Section 2, then we continue with the problem definition, where we elaborate basic foundations and formalize fairness objectives based on which a fair team is defined. We propose our approach in Section 3. The experimental setup and evaluation are described in Section 4, followed by concluding remarks in Section 5.

2 | RELATED WORK

The works related to this paper are largely around *i*) neural team recommendation methods and *ii*) fairness-aware recommendation methods.

2.1 | Neural Team Recommendation

Among the proposed team recommendation methods, we focus on neural models as the cutting-edge computational methods which offer efficiency and effectiveness due to the inherently iterative and online learning procedure. Proposed neural team recommendation models include non-variational feedforward[27,39], variational Bayesian network[22,27,20,39], and graph neural network[28,26,40]. Initially, Rad et al.[27] defined team recommendation as a multilabel classification task and, as a naive baseline for a minimum level of comparison, developed a simple feedforward network with one hidden layer to map the required subset of skills in the input layer onto a subset of experts in the output layer using the standard cross-entropy loss. Rad et al.[27,20] then proposed a variational Bayesian network to mitigate the popularity bias through uncertainty in neural model weights in the form of Gaussian distributions. In this line, Dashti et al[21] further proposed negative sampling heuristics assuming groups of experts who have little or no collaborative experience for the required subset of skills have a low chance for a successful collaboration and can be considered as *virtually un*successful teams. Given that popular experts were dominant in the training datasets, Dashti et al. presume that groups of popular experts are more likely to be selected as negative samples of teams. Successfully as they are, the primary focus of Dashti et al. and Rad et al. was the maximization of the efficacy by tailoring the recommended experts for a team to the required skills only, overlooking to substantiate whether the higher efficacy comes with mitigation of popularity bias.

Sapienza et al.[28] were the first to use a graph neural network in the form of an autoencoder for team recommendation in online multiplayer games. Later, Rad et al.[26] proposed to transfer dense vector representations of skills for the input of variational Bayesian neural network from a heterogeneous graph whose nodes are teams, experts, skills, and locations and edges connect experts who have collaborated in a team residing in a location using Dong et al.'s `metapath2vec`[41] and obtained the state-of-the-art performance. More recently, Kaw et al.[40] employed deep graph infomax[42], a graph convolution network with attention layer as an encoder, to learn more effective vector representations of skills in less training epochs owing to the convolutional architecture and contrastive learning procedure.

Nonetheless and despite a few efforts[27,20,21], existing neural team recommendation models still withhold extreme biases. Meanwhile, accounting for fairness in neural models has gained significant importance for their widespread applications in everyday lives, like in healthcare[43,44,45,46], information retrieval[47,48], computer vision[49,50,51,52,53], and recommendation systems[30,54,55,56,57,58,59,60]. To this end, in this paper, we are among the first to formalize the fair team recommendation problem with respect to group-based notions of fairness and undertake an empirical investigation to bridge the fairness gap through a probabilistic post-processing reranking method in favor of recommending more female or nonpopular experts while controlling the accuracy of the recommended teams.

2.2 | Fairness-aware Recommendation

Theoretically, fairness guarantees in machine learning algorithms have been defined at an individual level[61] where an individual should be treated consistently[62] or based on a group of individuals where a disadvantaged group, also known as a protected group, should be treated similarly to the advantaged group as a whole[63,64]. Different fairness-aware methods have been proposed to either discover and measure unfair biases[65], or to mitigate them via debiasing algorithms[30,54,66,57,59,67] at individual or group levels.

Debiasing algorithms can further be categorized based on their placement in the machine learning pipeline: *i*) pre-processing[68] methods modify data or its labels by re-sampling heuristics before model training, *ii*) in-processing[56,58,69] techniques modify models' optimization process to trade-off accuracy with fairness considerations, and *iii*) post-processing[30,70,66,69,57,71] methods modify models' outputs during inference, which may involve modifying thresholds, scoring rules, or reranking of the recommended list of items[72,34]. The latter category is of particular interest for it can be model-agnostic and plugged into a model with little to no modification to the model's architecture or negative impact on its predictive power. Related to this paper, we explain seminal reranking debiasing methods that achieve group-based fairness in recommendation tasks. Geyik et al.[57] propose greedy reranking algorithms to ensure prior desired distributions for disadvantaged protected group within the top-k items. At each iteration i; $1 \leqslant i \leqslant \mathrm{k}$, lower and upper bounds are calculated for protected group members to guarantee the desired distribution within top-i. To measure bias in original rankings and rerankings, they use `skew`[57] and normalized discounted cumulative KL-divergence (`ndkl`)[65]. Geyik et al.'s algorithms are, however, deterministic and fall short in the presence of real-world uncertainties. In contrast, Zehlike et al.[30] have proposed a probabilistic method to produce a top-k ranking while maintaining fairness towards multiple protected groups. They rely on statistical tests and aim for a minimum proportion of protected items in each subset of the ranked items. Utilizing cumulative distribution functions, they calculate the minimum number of required protected items at a given position to hold the fairness criteria with a pre-defined confidence level. Instead of providing rigid decisions, they offer distributions over possible outcomes, ensuring that items with small probability are not completely disregarded.

To the best of our knowledge, there is yet to be a neural team recommendation method that specifically takes fairness into account except Loghmani et al.[32], wherein the application of deterministic reranking algorithms[57] to mitigate popularity bias in neural team recommenders[20,27,21] were shown futile due to the substantial compromise to the models' accuracy.

3 | FAIR NEURAL TEAM RECOMMENDATION

In this section, we introduce the necessary notations and definitions for neural team recommendation, on the one hand, and group-based fairness, on the other hand. Then, we provide a formal problem statement to recommend a fair team of experts.

3.1 | Preliminaries

Given a set of skills $\mathcal{S} = \{s\}$ and a set of experts $\mathcal{E} = \{e\}$, a team of experts $\mathrm{E} \subseteq \mathcal{E}; \mathrm{E} \neq \varnothing$ that collectively cover a skill set $\mathrm{S} \subseteq \mathcal{S}; \mathrm{S} \neq \varnothing$ is shown by (S, E) along with its success status y where $y \in \{0, 1\}$ which is known *a priori*. Further, $\mathcal{T} = \{(\mathrm{S}, \mathrm{E})_y : y \in \{0, 1\}\}$ indexes all teams, successful and unsuccessful. In team recommendation literature[28,73,10,74,75,76,77,78], an expert's set

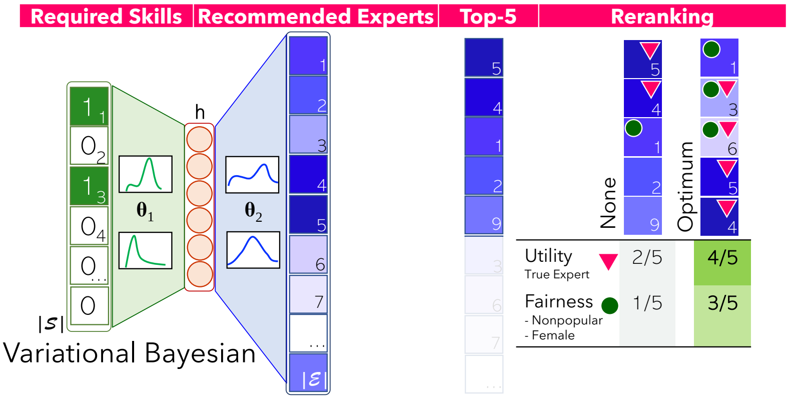

FIGURE 3 Post-processing fairness-aware reranking for fair neural team recommendation. ▽ indicates the correctly predicted expert (accuracy) and ○ shows members of the protected group (fairness). The goal is to maximize fairness while maintaining the model's accuracy.

of skills has been *estimated* through her participation in successful teams denoted by $\mathrm{S}_e = \{s : (s \in \mathrm{S}, e \in \mathrm{E})_{y=1} \in \mathcal{T}\}$, i.e., an expert member of a successful team inherits *all* the required skills of the team, as the expert obtains knowledge about all required skills through collaboration with other expert members of the team.

For a given subset of required skills S, the goal of the team recommendation problem is to recommend an optimal subset of experts E whose collaboration as a team leads to success, i.e., $(\mathrm{S}, \mathrm{E})_{y=1}$, while avoiding a potentially unsuccessful subset of experts E′, i.e., $(\mathrm{S}, \mathrm{E}')_{y=0}$. More concretely, the team recommendation problem is to find a mapping function f of parameters θ from the power set of skills to the powerset of experts such that $f_\theta : \mathcal{P}(\mathcal{S}) \rightarrow \mathcal{P}(\mathcal{E}), f_\theta(\mathrm{S}) = \mathrm{E}$.

As shown in Figure 3, the output of a neural team recommendation method for a required set of skills S is a ranked list of *all* experts where each expert $e \in \mathcal{E}$ is assigned a probability of her membership in the final recommended team. The recommended team is a subset of experts $\mathrm{E} \subseteq \mathcal{E}$ with the top-k highest probabilities. Given a subset of skills S and all teams $\mathcal{T}$ as the training set, neural team recommendation estimates $f_{\boldsymbol{\theta}}(\mathrm{S})$ using a multilayer neural network that learns, from $\mathcal{T}$, to map a vector representation of subset of skills S, referred to as v_{S}, to a vector representation of subset of experts E, referred to as v_{E}, by maximizing the posterior probability of $\boldsymbol{\theta}$ in $f_{\boldsymbol{\theta}}$ over $\mathcal{T}$, that is, $argmax_{\boldsymbol{\theta}} \;\; p(\boldsymbol{\theta}|\mathcal{T})$. For the vector representation of subset of skills v_{S}, neural team recommendation methods adopt either *i*) the *occurrence* vector representation for S, which is a Boolean vector of size $|\mathcal{S}|$, i.e., $v_{\mathrm{S}} \in \{0, 1\}^{|\mathcal{S}|}$ where $v_{\mathrm{S}}[s] = 1$ if $s \in \mathrm{S}$, and 0 otherwise, or *ii*) a dense low d-dimensional vector representation of S, $d << |\mathcal{S}|$, pretrained by e.g., a graph neural network method[40,26]. In the output layer for vector representation of subset of experts v_{E}, neural team recommendation methods frame the problem as a multilabel Boolean classification task and used occurrence vector representation for E, that is, $v_{\mathrm{E}} \in [0, 1]^{|\mathcal{E}|}$ where $v_{\mathrm{E}}[e] = 1$ if $e \in \mathrm{E}$, and 0 otherwise. Using a neural model of one hidden layer $\mathbf{h}$ of size d, without loss of generality to multiple hidden layers, with the input layer v_{S} and output layer v_{E}, a neural team recommendation method can be formalized as[39,21,23,26,27]:

$$\mathbf{h} = \pi(\boldsymbol{\theta}_1 v_{\mathrm{S}} + \mathbf{b}_1) \tag{1}$$

$$logits \rightarrow \mathbf{z} = \boldsymbol{\theta}_2 \mathbf{h} + \mathbf{b}_2 \tag{2}$$

$$v_{\mathrm{E}} = \sigma(\mathbf{z}) \tag{3}$$

where π is a nonlinear activation function, σ is the `sigmoid` function, and $\boldsymbol{\theta} = \boldsymbol{\theta}_1 \cup \boldsymbol{\theta}_2 \cup \mathbf{b}_1 \cup \mathbf{b}_2$ are learnable parameters for the mapping function f.

3.2 | Fair Team Recommendation

To eschew varied interpretations and to provide actionable criteria to design and evaluate fairness-aware algorithms, fairness has been mathematically formalized, with a level of abstraction from an underlying real-world scenario, based on well-known notions of justice and equity at an individual level, or at a group level like females vs. males. In this paper, we focus on group-based notions of fairness.

Given a protected attribute $a = \{a_1, ..., a_n\}$, e.g., gender=$\{\mathtt{0}$: male, $\mathtt{1}$: female$\}$ or popularity=$\{\mathtt{0}$: popular, $\mathtt{1}$: nonpopular$\}$, we divide experts into groups per their attribute values, each referred to as a protected group $\mathcal{G}_{a_i}$, e.g., female $\mathcal{G}_1$ vs. male $\mathcal{G}_0$ experts, or nonpopular $\mathcal{G}_1$ vs. popular $\mathcal{G}_0$ experts, such that experts of a protected group share the same value for a protected attribute. We then define the notions of group fairness for team recommendation as follows.

3.2.1 | Demographic Parity

Demographic parity, also called statistical parity, is to provide equal treatment to protected groups, i.e., the proportion of individuals receiving a favorable outcome should be consistent across all protected groups regardless of protected attributes. Given $\mathcal{D}$ the set of decisions, demographic parity requires a decision $d \in \mathcal{D}$ for members of protected groups to be oblivious to the value of the protected attribute. Formally,

$$\forall d \in \mathcal{D}, i \neq j \in a; p(\hat{d} \mid e_{a_i}) = p(\hat{d}|e'_{a_j}) \tag{4}$$

where $\hat{d}$ is the predicted decision for the correct decision d and $e_{a_i} \in \mathcal{G}_{a_i}$ is an expert whose value of protected attribute a is a_i. In fair team recommendation, we assume decisions are about the Boolean membership status of experts in the team of experts E, i.e., $e \in \mathrm{E}$ or $e \notin \mathrm{E}$, and protected attributes are either gender or popularity. Hence, Equation 4 becomes:

$$\forall e_0, e'_1 \in \mathcal{E}; [p(e_0 \in \ \mathrm{E}) = p(e'_1 \in \mathrm{E})] \wedge [p(e_0 \notin \mathrm{E}) = p(e'_1 \notin \mathrm{E})] \tag{5}$$

Intuitively, demographic parity enforces the membership in a team to be independent of values of a protected attribute for team members, i.e., no regard to their popularity, gender, ethnicity, or any other protected characteristic. However, demographic parity overlooks experts' qualifications; no criteria for experts' memberships has been defined in Equation 4 and Equation 5.

3.2.2 | Equalized Odds

Equalized odds is a stronger notion of fairness. While demographic parity emphasizes equal treatment by ensuring a similar proportion of positive outcomes across protected groups, equalized odds go further by ensuring that the ranking is equitable across groups for both qualified and non-qualified cases. In other words, it applies demographic parity on subsets of protected groups whose members are qualified (not qualified) to receive the correct decision in a Boolean decision set and Boolean protected attribute:

$$\forall d \in \mathcal{D} = \{\mathtt{0}, \mathtt{1}\}; p(\hat{d} \mid e_0, d) = p(\hat{d} \mid e_1, d) \tag{6}$$

In fair team recommendation, equalized odds would ensure that among skilled experts, individuals from different protected groups have equal chances of appearing in top positions, while among unqualified members, the likelihood of being incorrectly ranked highly remains consistent across protected groups. In other words it guarantees that the protected groups have equal true positive rates and false positive rates simultaneously. A *qualified* expert e for a team with a required subset of skills S can be simply defined as a Boolean measure based on whether the expert's skill set has a nonempty intersection with S, that is, $\mathrm{S}_e \cap \mathrm{S} \neq \varnothing$. Alternatively, as future work, we can consider a non-Boolean measure, e.g., the size of the intersection, to show more (less) qualified experts.

3.2.3 | Equality of Opportunity

Equality of opportunity is a relaxed version that only requires equal true positive rates across protected groups, without constraining the false positive rates. The intuition is to ensure that among individuals who are qualified for a positive outcome, the probability of receiving a positive prediction should be equal regardless of their protected attributes. This is less restrictive than equalized odds but can be more practical to implement while still preventing discrimination against qualified individuals from protected groups.

TABLE 1 The minimum number of required disadvantaged experts in the top-$k \in \{1, \cdots, 10\}$; equivalently, the minimum number of successes with statistical confidence of 0.9 in k Bernoulli trials each with winning probability of $p = 0.6$.

α	p	k	$\mathbb{F}^{-1}(k)$
0.1	0.6	1	0
		2	0
		3	1
		4	1
		5	2
		6	2
		7	3
		8	3
		9	4
		10	4

If we prioritize fairness for the decision $d = 1$, that is, recommending an expert to be in a team as an advantaged outcome, and ignore the possible discrimination caused by the *'not recommended'* decision $d = 0$, we have a less strict variant of equalized odds, referred to as equality of opportunity, as follows:

$$p(\hat{d} \mid e_0, d = 1) = p(\hat{d} \mid e_1, d = 1) \tag{7}$$

Hence, from Equation 6 and Equation 7:

$$p(e_0 \in \mathrm{E} \mid e_0, (\mathrm{S}_{e_0} \cap \mathrm{S} \neq \varnothing)) = p(e_1 \in \mathrm{E} \mid e_1, (\mathrm{S}_{e_1} \cap \mathrm{S} \neq \varnothing)) \tag{8}$$

It is worth reminding that, based on demographic parity, an expert having *no* required skills might be recommended. In contrast, based on the equalized dds and equality of opportunity, members must be qualified for the required skills, that is, the intersection of their skills and the required skills must be non-empty (Equation 6 and Equation 7).

3.2.4 | A Fair Team

Once we define the notions of fairness for the team recommendation problem, we can formalize a fair team by the fair *identity* function $\mathbb{I}$ as follows:

$$\mathbb{I}(\mathrm{E}) = \begin{cases} \text{Demographic Parity} & \begin{cases} 1 & \textit{iff} \text{ Equation 5} \\ 0 & \textit{not} \text{ a fair team} \end{cases} \\ \text{Equality of Opportunity} & \begin{cases} 1 & \textit{iff} \text{ Equation 8} \\ 0 & \textit{not} \text{ a fair team} \end{cases} \end{cases} \tag{9}$$

where E is a subset of experts in a team (S, E), which is fair with respect to demographic parity *iff* Equation 5. Alternatively, it is fair with respect to the notion of equality of opportunity *iff* Equation 8. It is worth noting that merely recommending a fair team while neglecting its success measures is also undesirable, e.g., a team of nonpopular experts who fall short of accomplishing tasks. Hence, metrics of accuracy (utility) based on the team's true label of success (y) should also be measured for a team recommender on top of fairness.

3.3 | Proposed Probabilistic Reranking Method

Let S be a subset of skills and $f_\theta(\mathrm{S})$ is the team recommendation method estimated by a neural model that recommends an optimum subset of experts, E who collectively cover the required subset of skills S and are *almost surely* successful, i.e.,$f_\theta(\mathrm{S}) = \mathrm{E}$ such that $(\mathrm{S}, \mathrm{E})_{y=1}$. If E is *not* a fair team, i.e., $\mathbb{I}(f_\theta(\mathrm{S})) = \mathbb{I}(\mathrm{E}) = 0$, our goal is to estimate a function $g : \mathcal{E} \rightarrow \mathcal{E}; g(\mathrm{E}) = \mathrm{E}^*$ such that:

$$\mathbb{I}(\mathrm{E}^*) = \mathbb{I}(g(\mathrm{E})) = \mathbb{I}(g(f_\theta(\mathrm{S}))) = 1 \tag{10}$$

We frame our reranking method based on *independent* Bernoulli trials of win/lose to find a fair team in the top-k ranked list of a model's prediction in Equation 1. Given p as the desired proportion of protected experts in a fair team and a significance level α, which is selected based on the underlying domain, we test, at each position, top-$\{1, \cdots, k\}$, if the ranked list statistically

TABLE 2 Statistics for the benchmark datasets utilized in our experiments.

	`dblp`	`imdb`	`uspt`
#teams	4,877,383	507,034	7,068,508
#experts	5,022,955	876,981	3,508,807
#skills	89,504	28	241,961
Avg #experts in teams	3.06	1.88	2.51
%popular experts (`avg`)	31.30%	42.60%	31.40%
%female experts	14.20%	12.30%	13.80%

significantly follows Bernoulli distribution with winning probability p. Let $\mathrm{E}_{r,k}$ be the first k experts of E ranked by a ranker r based on experts' probabilities produced by a team recommendation method that estimates f_θ. Further, let $|\mathrm{E}_{r,k}|_\eta$ be the current number of protected experts in the $\mathrm{E}_{r,k}$, $\mathbb{F}(|\mathrm{E}|_\eta; |\mathrm{E}|, p)$ be the cumulative distribution function for a binomial probability of having $|\mathrm{E}|_\eta$ experts of a protected group in the predicted team $|\mathrm{E}|$ independent Bernoulli trials with a winning rate p. We calculate the $\mathbb{F}$'s inverse function $\mathbb{F}^{-1}(\alpha; k, p)$ to determine the minimum number of required protected experts in E from top-$\{\mathtt{1}, \cdots, k\}$. Table 1 illustrates the values of $\mathbb{F}^{-1}(\alpha = \mathtt{0.1}; k, p = \mathtt{0.6})$ for top-k =10. We denote $\mathcal{G}_{a_1} = \{e_{\mathtt{1}}^{(1)}, \cdots, e_{\mathtt{1}}^{(l)}\}$ as the set of l experts in the disadvantaged protected group (e.g., female or nonpopular experts). Then, as for reranking function $g(\mathrm{E})$, we propose a new ranking r' based on the following:

$$\begin{cases} \hat{g}(e_i^{(k)}) = e_i^{(k)} & \text{if } \mathbb{F}^{-1}(\alpha; k, p) \leqslant |\mathrm{E}_{r,k}|_\eta \\ \\ \hat{g}(e_i^{(k)}) = e_{\mathtt{1}}^{(1)};\ \hat{g}(e_i^{(k+1)}) = e_{\mathtt{1}}^{(2)};\ \dots;\ \hat{g}(e_i^{(k+m-1)}) = e_{\mathtt{1}}^{(m)} & \text{if } \mathbb{F}^{-1}(\alpha; k, p) > |\mathrm{E}_{r,k}|_\eta \end{cases} \tag{11}$$

where $m = \mathbb{F}^{-1}(\alpha; k, p) - |\mathrm{E}_{r,k}|_\eta$ is the least number of protected experts to be added to make a team fair. Our method *inserts* the experts by shifting the experts down the list. For example, should a male expert in the fourth position ($e_{\mathtt{1}}^{(4)}$) be replaced with the most qualified female expert in the seventh position ($e_{\mathtt{0}}^{(7)}$), the female expert would be moved up to the fourth position ($\hat{g}(e_{\mathtt{i}}^{(4)}) = e_{\mathtt{0}}^{(7)}$) and shift the list down such that the male expert in the fourth position moves to the fifth $e_{\mathtt{1}}^{(5)}$, and so on. Therefore, as we rerank, all experts are still ranked based on their qualifications. If the number of disadvantaged experts up until the fourth position, i.e., $|\mathrm{E}_{r,k=4}|_\eta$, is greater than or equal to $\mathbb{F}^{-1}(\alpha; k = 4, p)$, sufficient disadvantaged experts have been inserted up to this position. Otherwise, m disadvantaged experts should be inserted depending on the given notion of fairness. For equality of opportunity, we select from *qualified* disadvantaged experts. However, for demographic parity, we overlook the qualification criteria and simply select from disadvantaged experts.

To form a fair ranking of k experts, we assume there exist at least k members from each protected group. From datasets in real-world scenarios, as seen in Table 2, the average number of members in a team (team size) is `3.06`, `1.88`, `2.51` in `dblp`, `imdb`, and `uspt`, respectively, which is almost surely less than the number of disadvantage experts, i.e., female or nonpopular experts, in the entire datasets. We empirically evaluate our reranking method on three large-scale datasets using three fairness metrics, namely `ndkl`[65], `skew`[57] and `expo`[59]. Meanwhile, we evaluate the models' accuracy by information retrieval metrics, including `map` and `ndcg`.

4 | EXPERIMENTS

In this section, we lay out the details of our experiments and findings to answer the following research questions:

RQ1: Does our proposed probabilistic reranking method mitigate unfair biases, including popularity bias and gender bias individually, in the recommended team of experts based on demographic parity and equality of opportunity while maintaining the team's likelihood of success?

RQ2: Does our proposed probabilistic reranking method outperform deterministic reranking methods in mitigating popularity and gender biases in view of demographic parity and equality of opportunity?

RQ3: Does our proposed probabilistic reranking method effectively reduce bias while enhancing the exposure to success ratio, as measured by utility-aware exposure (`expo`), for disadvantaged groups, e.g., nonpopular and female experts?

RQ4: Is the effect of our proposed reranking method consistent across datasets from different domains?

4.1 | Datasets

We evaluate our proposed method on three well-known large-scale benchmark datasets in team recommendation literature including `dblp`[20,21,27,35], `imdb`[21,36,16], and `uspt`[39,37]. In `dblp`, each team is a publication in computer science consisting of authors as the experts and the fields of study (fos) as the skills. In `imdb`, each instance is a movie. We consider each movie as a team whose members are the cast and crew, and the movies' genres are the teams' skills. The choice of `imdb` in team formation literature is not to be confused with its use cases in review analysis research; herein, the goal is to form a team of casts and crews for a *movie production* as opposed to a movie recommendation[36,16]. In `uspt`, each instance is a patent issued by the United States Patents and Trademarks consisting of inventors (experts) and subcategories (skills). Table 2 reports statistics on the datasets. From Figure 1, male experts are dominating teams while female experts have participated sparingly in all datasets. Also, from Figure 2, all datasets suffer from the long tail problem in the distributions of teams over experts, i.e., many experts (researchers in `dblp`, cast and crew in `imdb`, and inventors in `uspt`) have participated in very few teams (papers in `dblp`, movies in `imdb`, and inventions in `uspt`).

4.1.1 | Dataset Labeling Criteria

Success Labels: Benchmark datasets in team recommendation literature consist of successful teams only, missing unsuccessful ones. For instance, the `dblp` lacks unsuccessful submissions. Further, what it means for a team to be successful has remained controversial. For instance, in the movie industry, it is debatable whether a movie's success should be measured based on its immediate reception by the people (box office) or critical reviews (ratings) within a long span of time. In the absence of explicit labels for unsuccessful teams, neural team recommendation methods presume *all* instances of teams in the training dataset as successful (positive samples) and proceed with the training procedure[26,27,20,40].

Popularity Labels: Defining popularity could be controversial. To avoid varied interpretations, we followed social science[38] and recommender system literatures[79,80], where the popularity status of an expert can be *objectively* measured based on the number of teams the expert has participated in, referred to as *sociometric* popularity[38]. Further, although an expert's participation in many teams, e.g., research papers in `dblp` or movies in `imdb`, may *not* necessarily indicate popularity from the people's perspectives, repetition of the expert in many training samples of teams from the neural model's perspective does.

Accordingly, we can adopt two alternatives, as shown in Figure 2: *i*) an expert is popular if the expert participated in more than the average number of teams per expert over the entire dataset, and nonpopular otherwise (`avg`), or *ii*) an expert is popular if she belongs to the *short head* in the 2-d curve of the distribution of experts in teams, and nonpopular otherwise. We split the curve into *short head* and *long tail* based on equal area under the curve (`auc`). In this paper, we opt for the former setting. For the `dblp` dataset, the average stands at `23.02` teams, `62.45` teams for the `imdb` dataset, and `44.69` teams for the `uspt` dataset. Therefore, in `dblp`, the proportion of popular to nonpopular experts becomes `0.313` to `0.687`, in `imdb`, it is `0.426` to `0.574` and for `uspt` it stands at `0.314` to `0.686`.

Gender Labels: Contrary to popularity, gender is self-identified. While `uspt` dataset includes gender labels, other training datasets lack gender labels in part or whole. In `imdb`, although we inferred the gender of some cast and crew by their role identified as actor or actress, gender labels for other experts were missing. In `dblp`, no gender label for the experts has been provided. Therefore, we utilized genderize[81], based on the first name of the experts for `dblp` as well as those that are missing in `imdb`. As seen in Figure 1, datasets are heavily biased toward male experts. Specifically, `dblp` has a male-to-female ratio of `0.858` to `0.142`, `imdb` has a slightly different ratio of `0.877` to `0.123` and `uspt` has a ratio of `0.862` to `0.138`.

4.2 | Baselines

4.2.1 | Neural Team Recommendation

We compare the impact of our proposed probabilistic reranking method on mitigating neural models' biases using the state-of-the-art variational Bayesian neural network (`bnn`)[20,21,27] with a single hidden layer of size `d=128`, `leaky relu` as the activation function for the hidden layer, and Kullback-Leibler (KL) divergence as the optimizer. For the input layer, we used sparse occurrence vector representations (one-hot encoded) of skills of size $|\mathcal{S}|$ as well as pretrained dense vector representations

(-emb)[26]. The output layer is the sparse occurrence vector representations (one-hot encoded) of experts of size $|\mathcal{S}|$ and $|\mathcal{E}|$, respectively. We randomly select 15% of teams for the test set and perform 5-fold cross-validation on the remaining teams for model training over 20 epochs that results in one trained model per each fold. Given a team (S, E) from the test set, we select the top-k=100 experts with the highest probabilities as the recommended team $\mathrm{E} = f_{\boldsymbol{\theta}}(\mathrm{S})$ by the model of each fold.

4.2.2 | Fairness-aware Reranking

Our fairness-aware reranking baselines include three deterministic greedy reranking algorithms det-greedy, det-cons, and det-relaxed by Geyik et al.[57] as well as our proposed probabilistic reranking method with the significance level α=0.1.

4.3 | Evaluation Strategy

To measure the effectiveness of our proposed method, it is essential to evaluate fairness in the teams to ensure equitable treatment of all individuals and prevent systematic discrimination against protected groups. Measuring fairness metrics before and after team recommendation helps identify potential biases in the recommendation algorithm, quantifies the effectiveness of fairness interventions, and provides transparency about how well the system promotes fairness. These evaluations also help balance different notions of fairness and demonstrate compliance with ethical guidelines or organizational diversity goals while maintaining accountability in the team recommendation process. Team utility must be considered alongside fairness because teams ultimately need to accomplish tasks effectively. Measuring utility before and after fairness interventions is mandatory to understand potential performance trade-offs, and identify possible performance improvements through diversity, and demonstrate to stakeholders that fairness can be achieved while maintaining team effectiveness. Hence, we measure fairness and utility both before and after applying our methodologies to our baselines.

4.3.1 | Before Mitigating Bias

To answer our research questions, we evaluate the efficacy of the model in recommending *correct* experts for each team of our test set by comparing the ranked list of experts, predicted by the model of each fold, with the observed subset of experts E and report the average performance of models on all folds in terms of information retrieval metrics including mean average precision (map) and normalized discounted cumulative gain (ndcg) at top-10, as explained in Section 4.3.3. Furthermore, to assess the fairness of the predicted teams, we report the average fairness metrics, including normalized discounted Kullback-Leibler (ndkl)[65], skew[57], and expo[59] for popularity and gender attributes with respect to demographic parity and equality of opportunity, as formalized in Section 4.3.3. While skew is a symmetric metric to measure the symmetrical distribution of data among protected attributes, ndkl is an *a*symmetric measure of differences between the actual and desired distributions of protected attributes. For both of these metrics, the closer the value to 0, the more unbiased the distribution. In contrast to ndkl and skew that are not utility-aware, or in other words, they will not consider utility in their evaluations, expo calculates the exposure of different protected groups with respect to their utility.

4.3.2 | After Mitigating Bias

Given the predicted ranked list of experts $\hat{\mathrm{E}}$ by the model of each fold for the observed subset of experts E for a team of the test set, we apply fairness-aware reranking methods on $\hat{\mathrm{E}}$ to produce an unbiased ranked list of experts $\hat{g}(\hat{\mathrm{E}})$. We then evaluate the new reranked list in terms of information retrieval metrics and fairness metrics. We presume that fairness-aware reranking baselines improve the fairness metrics without discounting the informational retrieval metrics.

In total, we compare {bnn, bnn-emb} baselines for {popularity, gender} protected attributes with respect to {demographic parity, equality of opportunity} notions of fairness *before* and *after* applying fairness-aware reranking methods {det-greedy, det-cons, det-relaxed, **our method**} in terms of {map, ndcg} information retrieval metrics and {skew, ndkl, expo} fairness metrics on {dblp, imdb, uspt} datasets.

4.3.3 | Fairness and Utility Metrics

Let (S, E) a team of experts E for the required set of skills S from the test set, we compare the top-k ranked list of experts, predicted by the model of each fold for the input skills S, i.e., $f_\theta(\mathrm{S})$, with the observed subset of experts E and report the average performance of models on all folds in terms of utility metrics (the higher, the better) as formalized below.

1) **Mean Average Precision (`map`)**

$$\mathtt{ap}(k) : \frac{\sum_{i=1}^{k} \mathtt{pr}(i) \times \delta_{\mathrm{E}}(i)}{|\mathrm{E} \cap f_\theta(\mathrm{S})|} \tag{12}$$

where $\mathtt{pr}(k) = \frac{|\mathrm{E} \cap f_\theta(\mathrm{S})|}{k}$ is the precision, i.e., how many of the k predicted experts are correctly identified from the test instance of the team E and $\delta_e(i)$ returns `1` if the i-th predicted expert is in E. Finally, we report the mean of average precisions (`map`) on all test instances of teams.

2) **Normalized Discounted Cumulative Gain (`ndcg`)**

$$\mathtt{dcg}(k) = \sum_{i=1}^{k} \frac{\mathtt{rel}(i)}{\mathtt{log}(i+1)} \tag{13}$$

where $\mathtt{rel}(i)$ captures the degree of relevance for the predicted expert at position i. In our problem setting, however, all members of a test team are considered of the same importance. Therefore, $rel(i) = \mathtt{1}$ if $i \in \mathrm{E}$ and `0` otherwise, and Equation (13) becomes:

$$\mathtt{dcg}(k) = \sum_{i=1}^{k} \frac{\delta_{\mathrm{E}}(i)}{\mathtt{log}(i+1)} \tag{14}$$

This metric can be *normalized* relative to the ideal case when the top-k predicted experts include members of the test team E at the lowest possible ranks, i.e.,

$$\mathtt{ndcg}(k) = \frac{\sum_{i=1}^{k} \frac{\delta_{\mathrm{E}}(i)}{\mathtt{log}(i+1)}}{\sum_{i=1}^{|\mathrm{E}|} \frac{1}{\mathtt{log}(i+1)}} \tag{15}$$

Utility metrics are, however, oblivious to the protected attributes of experts, and hence, overlook whether the set of top-k predicted experts is a fair team (Section 3.2.4). To evaluate fairness, we use well-known fairness metrics as follows:

1) **Normalized Discounted KL Divergence (`ndkl`)**[65], which builds upon the foundation of Kullback-Leibler (KL) divergence to measure the expectation of the logarithmic difference between two discrete probability distributions, (the lower, the better) with being 0 in the ideal equal distributions. However, it advances a step further by incorporating a discounting factor, which allows it to assign varying levels of importance to different elements within the distributions being compared. This discounting is particularly valuable in scenarios where the order or priority of elements matters, as a result of which `ndkl` has been extensively employed in recommendation systems and information retrieval[82,83,57,84,85,32,86]. Additionally, `ndkl` includes a normalization component, which scales the results to a more interpretable range and facilitates comparisons across different baselines. Formally, let $p = \frac{|\mathrm{E}_{r,i}|_\eta}{i}$ be the distribution of a protected group in the top-i predicted experts by a ranker r, e.g., the proportions of nonpopular or female experts, and q the ideal fair distribution for a test instance of a team (S, E), the KL divergence of q from p is:

$$\mathtt{kl}(p||q) = \sum_{i=1}^{k} p(i)\ \mathtt{log}\frac{p(i)}{q(i)} \tag{16}$$

where q can be manually set, like 50%, or calculated based on the overall distribution (proportion) of the protected group in the entire dataset, i.e., $q = \frac{|\mathcal{E}|_\eta}{|\mathcal{E}|}$. This metric has a minimum value of `0` when both distributions are identical up to position i. A higher value indicates a greater divergence between the two distributions, and the metric is always non-negative. We report the *normalized discounted cumulative* KL-divergence $(\mathtt{ndkl})$[87]:

$$\texttt{ndkl}(p) = \frac{\sum_{i=1}^{|\mathrm{E}|} \frac{1}{\texttt{log}(i+1)} \texttt{kl}(p||q)}{\sum_{i=1}^{|\mathrm{E}|} \frac{1}{\texttt{log}(i+1)}} \tag{17}$$

2) **`Skew@k`**[57] is the logarithmic ratio of (1) the proportion of items, herein the experts, from a protected group among the top-*k* predicted experts to (2) the ideal fair proportion for that group. Similar to `ndkl`, given $p = \frac{|\mathrm{E}_{r,i}|_\eta}{i}$ as the distribution of a protected group in the top-*i* predicted experts by a ranker *r*, and $q = \frac{|\mathcal{E}|_\eta}{|\mathcal{E}|}$ the ideal fair distribution, without loss of generality to any desired distribution for a test instance of a team (S, E):

$$\texttt{skew@}i(r) = \log_e(\frac{p}{q}) \tag{18}$$

A negative `skew@`i corresponds to a lesser than desired representation of experts with the protected group in the top-*k* results, while a positive `skew@i` corresponds to favoring such experts. The `log` makes this metric symmetric around zero with respect to ratios for and against a specific protected group and is particularly useful for assessing whether a ranker tends to favor certain protected groups disproportionately[88].

3) **Utility-aware Exposure (`expo`)**[59] measures the ratio of exposure to success across different protected groups in the top-*k* predicted experts[56,89,90,30,91,92,93,94,95,96]. In recommender systems, exposure for each protected group is defined as the expected probability that an expert of a protected group will be presented at top-*k* position. This metric quantifies the likelihood of visibility for each member of protected group and provides a measure of how equitably exposure is distributed across different protected groups. Given the top-*k* ranked list of predicted experts for a team E, the exposure for an expert is calculated as:

$$\texttt{expo}(e) = \frac{1}{\texttt{log}(i+1)}; e \in \mathrm{E} \tag{19}$$

The simplest fair exposure among groups is that the average exposures of the experts in protected groups are equal. Given a protected attribute $a = \{a_1, ..., a_n\}$, the average exposure for the experts of a protected group $\mathcal{G}_{a_i}$ based on a protected attribute value a_i, e.g., female experts $\mathcal{G}_1$ vs. male experts $\mathcal{G}_0$, at the top-*k*:

$$\mu_{\texttt{expo}}(a_i) = \frac{1}{|\mathrm{E} \cap \mathcal{G}_{a_i}|} \sum_{e \in \mathrm{E} \cap \mathcal{G}_{a_i}} \texttt{expo}(e) \tag{20}$$

The expected utility of an expert in position-based ranking models can be determined by the probability of the expert appearing in a given position. To integrate utility in the exposure metric, an average probability score over experts of a protected group is also calculated as:

$$\mu_{utility}(a_i) = \frac{1}{|\mathrm{E} \cap \mathcal{G}_{a_i}|} \sum_{e \in \mathrm{E} \cap \mathcal{G}_{a_i}} v_{\mathrm{E}}(e) \tag{21}$$

where v_{E} is a vector of probabilities in Equation 3 based on which a ranker select the top-*k* predicted experts as the recommended team. This is a form of group fairness that considers the utility associated with experts. The exposure value for a protected group is the ratio of average exposure values and average probability scores for a protected group, as follows:

$$\texttt{expo}(a_i) = \frac{\mu_{\texttt{expo}}(a_i)}{\mu_{utility}(a_i)} \tag{22}$$

Finally, the overall `expo` value is calculated for a protected attribute based on the ratio of exposure value for a pair of protected groups as:

$$\texttt{expo}(a) = \frac{\texttt{expo}(a_1)}{\texttt{expo}(a_0)} \tag{23}$$

4.4 | Results

We report the comparative experimental results on three benchmark datasets `dblp`, `imdb`, and `uspt` in Tables 3, 4 and 5, respectively. Our analysis reveals several key findings. Foremost, we observe that neural team recommendation baselines

TABLE 3 Average performance of 5-fold neural models on test set on dblp dataset. For the metrics, ndkl, the lower the better (↓), skew, the closer to 0 the better (→0), and map and ndcg, the higher the better (↓).

		%ndkl	%ndkl	skew before→0		skew after→0		%map10	%ndcg10
		before↓	after↓	nonprotected	protected	nonprotected	protected	Δ ↑	Δ ↑
popularity, demographic parity									
bnn	det-cons	109.56	14.64	1.13	-19.97	0.64	-0.54	-0.28	-0.58
	det-greedy		14.64			0.64	-0.54	-0.28	-0.58
	det-relaxed		18.31			0.64	-0.53	-0.28	-0.58
	our method		19.71			0.26	-0.15	0.00	00.00
bnn-emb	det-cons	110.31	14.09	1.14	-20.75	0.62	-0.51	-0.28	-0.58
	det-greedy		14.09			0.62	-0.51	-0.28	-0.58
	det-relaxed		17.65			0.61	-0.50	-0.28	-0.58
	our method		19.61			0.26	-0.15	0.00	0.00
popularity, equality of opportunity									
bnn	det-cons	102.01	13.12	1.05	-19.92	0.57	-0.51	-0.28	-0.58
	det-greedy		13.16			0.57	-0.51	-0.28	-0.58
	det-relaxed		16.15			0.57	-0.50	-0.28	-0.58
	our method		18.96			0.24	-0.16	0.00	0.00
bnn-emb	det-cons	102.85	12.65	1.06	-20.62	0.55	-0.48	-0.28	-0.58
	det-greedy		12.67			0.55	-0.48	-0.28	-0.58
	det-relaxed		15.63			0.55	-0.47	-0.28	-0.58
	our method		18.39			0.25	-0.16	0.00	0.00
gender, demographic parity									
bnn	det-cons	11.80	4.92	-0.08	0.42	-0.07	0.39	-0.28	-0.58
	det-greedy		3.72			-0.07	0.39	-0.28	-0.58
	det-relaxed		6.52			0.00	-0.20	-0.28	-0.58
	our method		8.39			0.00	-0.11	0.00	0.00
bnn-emb	det-cons	7.29	4.97	-0.06	0.30	-0.07	0.37	-0.28	-0.58
	det-greedy		3.59			-0.07	0.37	-0.28	-0.58
	det-relaxed		4.92			-0.04	0.08	-0.28	-0.58
	our method		6.83			-0.03	0.13	0.00	0.00
gender, equality of opportunity									
bnn	det-cons	18.97	9.19	-0.14	0.88	-0.13	0.86	-0.28	-0.58
	det-greedy		7.70			-0.13	0.86	-0.28	-0.58
	det-relaxed		9.53			-0.13	0.85	-0.28	-0.58
	our method		18.97			-0.14	0.88	0.00	0.00
bnn-emb	det-cons	15.93	9.24	-0.11	0.76	-0.13	0.86	-0.28	-0.58
	det-greedy		7.61			-0.13	0.86	-0.28	-0.58
	det-relaxed		10.16			-0.13	0.85	-0.28	-0.58
	our method		15.93			-0.11	0.76	0.00	0.00

exhibit significant biases with respect to two protected attributes: popularity (favoring already well-known experts) and gender (showing systematic underrepresentation of female experts). Before applying our debiasing reranking method, these baselines demonstrated a clear tendency to amplify the pre-existing biases in the data. Specifically, experts with a high volume of papers in dblp, many movies in imdb, and patents in uspt would be disproportionately recommended, while qualified experts with lower visibility were systematically overlooked. Similarly, the gender distribution in the recommended teams was significantly skewed, indicating algorithmic bias rather than a mere reflection of domain demographics. In terms of ndkl and skew metrics, Tables 3, 4 and5 reveal popularity bias across different domains and baselines. The ndkl metric, where 0 represents the ideal value indicating perfect alignment with the desired distribution, consistently showed substantial positive values across *all* baselines. Specifically, we observed ndkl values ranging from 102.01 to 110.31 for dblp, 61.74 to 74.67 for imdb, and 42.11 to 110.13 for uspt datasets, indicating divergence from the desired distribution obtained by demographic parity and equality of opportunity fairness notions. This deviation demonstrates that the baselines disproportionately favored popular experts, effectively marginalizing less frequently occurring experts in the recommended teams. The consistency of the observed pattern indicates that popularity bias is not confined to any single domain but rather represents a systematic issue in neural team recommendation baselines before our fairness interventions.

The skew metric, which is symmetric around 0, representing the desired representation, provides a more detailed view of representation disparities. Positive values indicate over-representation and negative values signal under-representation of protected groups. This metric also shows consistent patterns of popularity bias across our experimental settings. Tables 3, 4 and5 demonstrate that across all domains and baselines before our fairness considerations, there is a systematic over-representation of

TABLE 4 Average performance of `5`-fold neural models on test set on `imdb` dataset. For the metrics, `ndkl`, the lower the better (↓), `skew`, the closer to `0` the better (→`0`), and `map` and `ndcg`, the higher the better (↓).

		`%ndkl`	`%ndkl`	`skew` before→`0`		`skew` after→`0`		`%map10`	`%ndcg10`
		before↓	after↓	nonprotected	protected	nonprotected	protected	Δ ↑	Δ ↑
popularity, demographic parity									
bnn	det-cons	67.53	16.59	0.74	-4.01	-0.54	0.26	-0.36	-0.82
	det-greedy		16.60			-0.54	0.26	-0.36	-0.82
	det-relaxed		16.35			-0.53	0.26	-0.36	-0.82
	our method		17.27			0.21	-0.19	0.00	00.00
bnn-emb	det-cons	74.67	15.71	0.7870	-4.30	-0.46	0.23	-0.48	-1.03
	det-greedy		15.72			-0.46	0.23	-0.48	-1.03
	det-relaxed		15.43			-0.45	0.23	-0.48	-1.03
	our method		17.53			0.21	-0.19	0.00	0.00
popularity, equality of opportunity									
bnn	det-cons	61.74	19.85	0.68	-3.96	-0.62	0.32	-0.35	-0.81
	det-greedy		20.11			-0.62	0.32	-0.35	-0.81
	det-relaxed		19.70			-0.62	0.32	-0.35	-0.81
	our method		16.61			0.19	-0.20	0.00	0.00
bnn-emb	det-cons	70.61	18.94	0.72	-4.17	-0.55	0.30	-0.48	-1.03
	det-greedy		19.17			-0.55	0.30	-0.48	-1.03
	det-relaxed		18.68			-0.55	0.30	-0.48	-1.03
	our method		18.15			0.20	-0.20	0.00	0.00
gender, demographic parity									
bnn	det-cons	4.13	4.44	0.00	-0.04	0.03	-0.34	-0.36	-0.81
	det-greedy		3.96			0.04	-0.36	-0.36	-0.81
	det-relaxed		4.00			0.04	-0.35	-0.36	-0.81
	our method		4.09			0.00	-0.04	0.00	0.00
bnn-emb	det-cons	4.92	8.34	0.01	-0.13	0.06	-0.61	-1.17	-1.03
	det-greedy		3.99			0.03	-0.33	-1.17	-1.03
	det-relaxed		4.21			0.03	-0.31	-1.17	-1.03
	our method		4.88			0.01	-0.12	0.00	0.00
gender, equality of opportunity									
bnn	det-cons	3.42	3.97	0.00	-0.06	0.03	-0.27	-0.42	-0.99
	det-greedy		3.67			0.03	-0.25	-0.42	-0.99
	det-relaxed		3.71			0.02	-0.25	-0.42	-0.99
	our method		3.39			0.00	-0.05	0.00	0.00
bnn-emb	det-cons	4.00	4.17	0.01	-0.13	0.02	-0.24	-0.54	-1.20
	det-greedy		3.89			0.02	-0.22	-0.54	-1.20
	det-relaxed		3.93			0.02	-0.21	-0.54	-1.20
	our method		3.96			0.01	-0.13	0.00	0.00

popular experts (indicated by positive `skew` values) joint with consistent under-representation of nonpopular experts (negative `skew`). This bidirectional deviation from the ideal case is particularly informative as it quantifies not just the presence of popularity bias, but its direction and magnitude. For instance, in the `dblp` dataset, with demographic parity notion of fairness and `bnn` team recommendation baseline, popular experts showed a `skew` of `1.13`, indicating they received more recommendations than would be expected based on their proportion in the entire dataset. Conversely, nonpopular experts exhibited a `skew` of `-19.97`, suggesting they were recommended less frequently than desired. Similar patterns emerged in both the `imdb` and `uspt` datasets, with popular experts consistently showing positive values and nonpopular experts showing negative values.

> **Finding 1.** Neural team recommendation baselines, regardless of the underlying architecture, withhold substantial biases in both popularity and gender representation before our debiasing intervention.

In response to **RQ1**, whether our proposed probabilistic reranking method can mitigate popularity bias in the recommended team of experts based on demographic parity and equality of opportunity while maintaining the team's likelihood of success, from Tables 3, 4 and 5, we can observe that our method could substantially reduce the bias of the neural team recommendation baselines in terms of `ndkl` and `skew` (closer to `0`) with no change to the information retrieval metrics on all datasets. With respect to gender bias, however, our reranking method falls short; `ndkl` and `skew` values after applying our reranking methods generally remain the same, or become worse (larger value) in several settings like for `bnn` baseline on `uspt` dataset where `ndkl`

TABLE 5 Average performance of 5-fold neural models on test set on `uspt` dataset. For the metrics, `ndkl`, the lower the better (↓), `skew`, the closer to `0` the better (→`0`), and `map` and `ndcg`, the higher the better (↓).

		`%ndkl`	`%ndkl`	`skew` before →`0`		`skew` after →`0`		`%map10`	`%ndcg10`
		before↓	after↓	protected	nonprotected	protected	nonprotected	Δ ↑	Δ ↑
popularity, demographic parity									
bnn	det-cons	90.93	24.36	-10.58	1.05	0.81	0.78	-0.24	-0.53
	det-greedy		24.35			0.81	0.78	-0.24	-0.53
	det-relaxed		28.94			-0.80	0.78	-0.24	-0.53
	our method		18.52			-0.13	-0.59	0.00	0.00
bnn-emb	det-cons	110.13	18.88	-22.79	1.13	-0.61	0.68	-0.67	-1.35
	det-greedy		18.88			-0.61	0.68	-0.67	-1.35
	det-relaxed		24.13			-0.60	0.67	-0.67	-1.35
	our method		19.68			-0.15	0.22	0.00	0.00
popularity, equality of opportunity									
bnn	det-cons	42.11	16.12	-9.95	0.44	-0.32	0.24	-0.24	-0.53
	det-greedy		20.39			-0.32	0.24	-0.24	-0.53
	det-relaxed		19.06			-0.32	0.24	-0.24	-0.53
	our method		17.20			-0.14	-0.62	0.00	0.00
bnn-emb	det-cons	52.44	14.50	-22.16	0.53	-0.18	0.18	-0.67	-1.35
	det-greedy		19.48			-0.18	0.18	-0.67	-1.35
	det-relaxed		18.04			-0.18	0.18	-0.67	-1.35
	our method		14.65			-0.21	0.13	0.00	0.00
gender, demographic parity									
bnn	det-cons	6.50	3.73	0.30	-0.08	0.13	-0.02	-0.24	-0.53
	det-greedy		2.67			0.13	-0.02	-0.24	-0.53
	det-relaxed		3.51			0.13	-0.25	-0.24	-0.53
	our method		11.46			0.36	-0.37	0.00	0.00
bnn-emb	det-cons	8.07	4.52	0.54	-0.13	0.13	-0.02	-0.67	-1.35
	det-greedy		2.68			0.13	-0.02	-0.67	-1.35
	det-relaxed		4.37			0.12	-0.02	-0.67	-1.35
	our method		8.32			0.55	-0.141	0.00	0.00
gender, equality of opportunity									
bnn	det-cons	12.08	10.15	0.20	0.00	0.00	0.07	-0.24	-0.53
	det-greedy		9.03			0.00	0.07	-0.24	-0.53
	det-relaxed		9.71			0.00	0.07	-0.24	-0.53
	our method		16.40			0.33	-0.33	0.00	0.00
bnn-emb	det-cons	12.80	10.81	0.44	-0.04	0.01	0.06	-0.67	-1.35
	det-greedy		9.15			0.01	0.06	-0.67	-1.35
	det-relaxed		10.49			0.01	0.06	-0.67	-1.35
	our method		12.8			0.47	-0.05	0.00	0.00

has increased from `6.50` to `11.46` after applying our method. We attribute this to the extreme gender bias in datasets (e.g., `0.862` to `0.138` male vs. female ratio in `uspt`) such that the original top-*k* ranked list lacks enough (qualified) female experts.

> **Finding 2.** Our proposed probabilistic reranking can mitigate **popularity** bias while maintaining utility across *all* domains.

However, our method exhibits clear limitations when addressing gender bias. The fairness metrics, `ndkl` and `skew` values post-reranking generally remain the same or become worse. For instance, applying our method to the `bnn` baseline on the `uspt` dataset resulted in a significant degradation, with `ndkl` increasing from `6.50` to `11.46`. This limitation stems from severe pre-existing biases in the underlying data distributions, exemplified by the severe gender imbalance in datasets such as `uspt` with its `0.862` to `0.138` male-to-female ratio. Such extreme gender bias creates a fundamental limitation where the top-*k* ranked lists of experts simply lack a sufficient pool of qualified female experts to draw from. Additionally, our results point to the broader challenge of achieving fairness in domains with extreme underrepresentation, where the space of possible fair solutions is constrained by the available expert pool.

Finding 3. Different types of bias may require distinct mitigation strategies due to the *level* of bias across different protected attributes.

To answer **RQ2**, regarding whether our proposed probabilistic reranking method outperforms deterministic reranking methods, from Tables 3, 4 and5 we observe that our probabilistic method demonstrates superior performance over deterministic rerankers specifically in mitigating popularity bias across various fairness notions on all datasets, as evidenced by the combination of fairness and information retrieval metrics. This superiority is reflected in the consistency of bias reduction and maintaining information retrieval metrics. While deterministic rerankers show some capability in mitigating popularity bias, this comes at the cost of a drastic drop in information retrieval metrics. Specifically, we observe consistent negative values in Δmap@10 and Δndcg@10, indicating drops in recommendation quality. This trade-off between fairness and accuracy highlights an important advantage of our probabilistic approach, which maintains recommendation quality while improving fairness metrics.

The effectiveness of gender bias mitigation strategies demonstrates fewer clear-cut results. Deterministic rerankers achieve marginally better fairness metrics, showing improvements in ndkl and skew values, compared to our probabilistic method. However, this modest improvement in fairness comes at a considerable cost to recommendation accuracy, with deterministic approaches showing a significant loss in information retrieval metrics. This suggests that while deterministic approaches might achieve slightly better results in terms of mitigating gender bias, they sacrifice substantial recommendation quality, questioning their practical applicability.

It is particularly noteworthy that all deterministic rerankers exhibit similar performance patterns across datasets, with no single approach demonstrating clear dominance over the others. This finding aligns with Loghmani et al.'s[32] observations and suggests a fundamental limitation in deterministic approaches. While deterministic approaches might seem appealing due to their simplicity and interpretability, our findings suggest that probabilistic methods offer a better approach to addressing bias, considering both fairness and accuracy.

Finding 4. In the context of neural team recommendation, our proposed probabilistic reranking method consistently outperforms deterministic reranking methods on all datasets and baselines.

For **RQ3**, regarding the analysis of expo metric where the ideal value is 1, indicating that the protected groups' exposure is perfectly proportional to their utility scores, and values less (greater) than 1 suggest bias against (in favor of) the disadvantaged group, from Table 6, we observe that our proposed probabilistic reranking method shows improvements for protected groups across datasets, particularly for popularity as the protected attribute. For nonpopular items, we observe consistent and substantial improvements across all datasets. Before reranking, the baseline bnn exhibited significant bias against nonpopular groups, with expo values substantially less than 1 across all datasets. After applying our probabilistic reranking method, these values adjusted closer to the ideal. This shift indicates that our method effectively balanced the exposure of nonpopular groups relative to their success, mitigating the initial bias. In contrast, deterministic reranking algorithms often overcompensated, resulting in expo values significantly greater than 1. For instance, det-cons produced values of 2.4358 for dblp on bnn baseline with demographic parity. These elevated values suggest reverse discrimination, favouring the disadvantaged group disproportionately.

With respect to gender, the results are, however, relatively poor, similar to **RQ2** and 3, 4 and 5. All post-processing reranking methods fall short of mitigating the gender bias, which can be attributed to the presence of extreme gender bias in the dataset. For instance, with male-to-female ratios as skewed as 0.862 to 0.138 in uspt, the top-*k* ranked list of experts lacks sufficient representation of female experts, making it impossible for post-processing methods alone to achieve fairness without compromising quality. Nonetheless, although our proposed probabilistic method negligibly changes the values of expo for better or worse across different settings, its performance is better than deterministic reranking methods. This is evidenced by how deterministic methods tend to push expo values well above 1, indicating an artificial inflation of female representation that compromises the quality of recommendations.

In summary, our probabilistic reranking method consistently mitigated popularity bias across all datasets and fairness definitions while maintaining utility in terms of expo. Unlike deterministic methods that may overcompensate, and drop utility drastically, our approach presents a balance by proportionally adjusting exposure based on information retrieval metrics. Also, our method successfully improves the ratio of exposure to utility for disadvantaged groups, moving expo values towards the ideal 1 across multiple datasets and fairness definitions.

TABLE 6 Average performance of 5-fold neural models in terms of expo on test set on imdb, dblp, uspt datasets. For this metric expo, the closer to 1 the better (→1)

		imdb		dblp		uspt	
		expo→1		expo→1		expo→1	
		before	after	before	after	before	after
popularity, demographic parity							
bnn	det-cons	0.7737	1.7014	0.2612	2.4358	0.6774	1.2303
	det-greedy		1.7014		2.7576		1.2314
	det-relaxed		1.6491		2.2494		1.1268
	our method		**1.0053**		**1.0122**		**0.9858**
bnn-emb	det-cons	0.7705	1.6838	0.2382	2.5549	0.1564	1.1328
	det-greedy		1.6838		2.5552		1.1327
	det-relaxed		1.6320		2.3588		**1.0296**
	our method		**0.9887**		**1.0316**		1.0544
popularity, equality of opportunity							
bnn	det-cons	0.7737	1.6836	0.2612	2.4456	0.6774	1.3286
	det-greedy		1.6951		2.4488		1.4296
	det-relaxed		1.6434		2.2681		1.3889
	our method		**0.8406**		**1.0139**		**0.9909**
bnn-emb	det-cons	0.7705	1.6677	0.2382	2.5813	0.1564	1.2239
	det-greedy		1.6795		2.5839		1.3298
	det-relaxed		1.6261		2.3951		1.2892
	our method		**0.8377**		**1.0338**		**1.0415**
gender, demographic parity							
bnn	det-cons	**0.9195**	1.1357	0.9741	1.1106	0.9461	0.9925
	det-greedy		1.1640		1.1024		1.0002
	det-relaxed		1.1610		1.1433		0.9481
	our method		0.9066		**0.9750**		**0.9814**
bnn-emb	det-cons	**0.9302**	1.3136	**1.0084**	1.1528	0.9761	1.0175
	det-greedy		1.3003		1.1540		1.0011
	det-relaxed		1.3082		1.1693		1.0128
	our method		0.9233		1.0080		**0.9779**
gender, equality of opportunity							
bnn	det-cons	**0.9195**	1.1314	0.9741	1.1324	0.9461	1.0065
	det-greedy		1.1563		1.1242		1.0144
	det-relaxed		1.1561		1.1597		1.0189
	our method		0.9066		**0.9743**		**0.9784**
bnn-emb	det-cons	0.9302	1.2875	1.0084	1.1728	0.9761	1.0228
	det-greedy		1.2909		1.1742		**1.0047**
	det-relaxed		1.2971		1.2047		1.0191
	our method		**0.9233**		**1.0014**		0.9752

> **Finding 5.** Our probabilistic reranking method's performance in terms of expo is consistent with its fairness metrics namely ndkl and skew across all settings and domains.

Regarding **RQ4**, Tables 3, 4 and 5 demonstrate that each fairness-aware reranker, whether deterministic or probabilistic, follows a similar trend across the dblp, imdb, and uspt datasets, despite these datasets originating from different domains. Specifically, the performance metrics, including fairness measures and utility metrics, remain consistent in terms of their trends when applying the same reranking algorithms to different datasets. This consistency suggests that the inherent patterns of bias and the distribution of protected attributes are similar across these datasets, as illustrated in Figure 2. Moreover, the similarity in trends indicates that the fairness-aware reranking algorithms are robust to domain variations and can generalize well across different domains. This robustness also implies that the underlying biases in rankings are not unique to a particular domain but are pervasive across various domains. Consequently, fairness interventions that are effective in one domain are likely to be effective in others.

> **Finding 6.** Our proposed probabilistic reranking method shows consistent effective performance in terms of both fairness and information retrieval metrics across datasets from different domains.

Lastly, our experiments show that while post-processing reranking methods can effectively address biases, their efficacy may become limited when employed single-handedly when confronting *extreme* biases in a dataset; such methods struggle to rectify biases without a consequential loss in accuracy. A holistic approach that integrates pre-processing, in-processing, and post-processing methods is required to achieve a more balanced and optimal outcome.

5 | CONCLUDING REMARKS

In this paper, we formalized the fair team recommendation problem, where we aim to form an unbiased collaborative group of diverse experts to accomplish complex tasks. While state-of-the-art neural team recommenders can efficiently recommend sets of candidate experts to form effective collaborative teams, they are largely biased toward *male* and *popular* experts, potentially overlooking valuable contributors from underrepresented groups. We proposed a model-agnostic post-processing probabilistic reranking method to mitigate unfair biases in the recommended teams of experts by neural team recommendation models, focusing on maintaining team effectiveness while promoting fairness with respect to demographic parity and equality of opportunity notions of fairness. Our experiments on three large-scale benchmark datasets from different domains, including imdb, dblp, and uspt, showed that: (1) neural team recommenders heavily suffer from biases toward popular and male experts, with popular experts; (2) probabilistic greedy reranking algorithms can substantially mitigate popularity biases while maintaining models' efficacy; (3) Biases appeared across all neural team recommendation architectures, indicating it is a fundamental challenge of these systems rather than a flaw in specific model designs; (4) In the presence of extreme biases, where initial recommendations show more than 80% skew toward certain groups, post-processing reranking methods fall short of achieving fair representation. (5) our probabilistic method dominantly outperforms deterministic baselines and is robust towards domain changes. Our future research direction includes mitigating multiple biases jointly, i.e., gender bias together with popularity bias, and incorporating in-processing methods to address these challenges at the model training stage rather than solely through post-processing adjustments.

6 | ETHICAL CONSIDERATIONS

We utilized genderize[81] to obtain missing genders for individuals in the imdb and all of the experts in the dblp dataset. Genderize is a crowd-based system that predicts a gender for a given name with a probability based on the number of appearances of a name in their database, name patterns and cultural associations, which may be reasonably accurate, it inherently incorporates societal biases and assumptions that can be problematic.

REFERENCES

1. Gómez-Zará D, Das A, Pawlow B, Contractor N. In search of diverse and connected teams: A computational approach to assemble diverse teams based on members' social networks. *PLOS ONE.* 2022;17(11):1-29. doi: 10.1371/journal.pone.0276061
2. Burke CS, Georganta E, Marlow S. A bottom up perspective to understanding the dynamics of team roles in mission critical teams. *Frontiers in Psychology.* 2019;10:441832.
3. Stokols D, Hall KL, Taylor BK, Moser RP. The science of team science: overview of the field and introduction to the supplement. *American journal of preventive medicine.* 2008;35(2):S77–S89.
4. Chen SG. An Integrated Methodological Framework for Project Task Coordination and Team Organization in Concurrent Engineering. *Concurr. Eng. Res. Appl..* 2005;13(3):185–197. doi: 10.1177/1063293X05056462
5. Wang L, Zeng Y, Chen B, Pan Y, Cao L. Team Recommendation Using Order-Based Fuzzy Integral and NSGA-II in StarCraft. *IEEE Access.* 2020;8:59559–59570. doi: 10.1109/ACCESS.2020.2982647
6. Campêlo MB, Figueiredo TF, Silva A. The sociotechnical teams formation problem: a mathematical optimization approach. *Ann. Oper. Res..* 2020;286(1):201–216. doi: 10.1007/s10479-018-2759-5
7. Esgario JGM, Silva dIE, Krohling RA. Application of Genetic Algorithms to the Multiple Team Formation Problem. *CoRR.* 2019;abs/1903.03523.
8. Kalayathankal SJ, Abraham JT, Kureethara JV. A Fuzzy Approach To Project Team Selection. *International Journal of Scientific Technology Research.* 2019;8.
9. Rahman H, Roy SB, Thirumuruganathan S, Amer-Yahia S, Das G. Optimized group formation for solving collaborative tasks. *VLDB J..* 2019;28(1):1–23. doi: 10.1007/s00778-018-0516-7
10. Durfee EH, Jr. JCB, Sleight J. Using hybrid scheduling for the semi-autonomous formation of expert teams. *Future Gener. Comput. Syst..* 2014;31:200–212. doi: 10.1016/j.future.2013.04.008
11. Strnad D, Guid N. A fuzzy-genetic decision support system for project team formation. *Appl. Soft Comput..* 2010;10(4):1178–1187. doi: 10.1016/j.asoc.2009.08.032
12. Wi H, Oh S, Mun J, Jung M. A team formation model based on knowledge and collaboration. *Expert Syst. Appl..* 2009;36(5):9121–9134. doi: 10.1016/j.eswa.2008.12.031

13. Baykasoglu A, Dereli T, Das S. Project Team Selection Using Fuzzy Optimization Approach. *Cybern. Syst.*. 2007;38(2):155–185. doi: 10.1080/01969720601139041
14. Chen SG, Lin L. Modeling team member characteristics for the formation of a multifunctional team in concurrent engineering. *IEEE Trans. Engineering Management.* 2004;51(2):111–124. doi: 10.1109/TEM.2004.826011
15. Zakarian A, Kusiak A. Forming teams: An analytical approach. *IIE Transactions.* 1999;31:85-97. doi: 10.1023/A:1007580823003
16. Kargar M, An A. Discovering top-k teams of experts with/without a leader in social networks. In: Macdonald C, Ounis I, Ruthven I., eds. *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*ACM 2011:985–994.
17. Sozio M, Gionis A. The community-search problem and how to plan a successful cocktail party. In: ACM 2010:939–948
18. Lappas T, Liu K, Terzi E. Finding a team of experts in social networks. In: IV JFE, Fogelman-Soulié F, Flach PA, Zaki MJ., eds. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*ACM 2009:467–476
19. Gaston ME, Simmons J, desJardins M. Adapting Network Structure for Efficient Team Formation. In: . FS-04-02. AAAI Press 2004:1–8.
20. Hamidi Rad R, Fani H, Bagheri E, Kargar M, Srivastava D, Szlichta J. A Variational Neural Architecture for Skill-Based Team Formation. *ACM Trans. Inf. Syst.*. 2023;42(1). doi: 10.1145/3589762
21. Dashti A, Samet S, Fani H. Effective Neural Team Formation via Negative Samples. In: Hasan MA, Xiong L., eds. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*ACM 2022:3908–3912
22. Dashti A, Saxena K, Patel D, Fani H. OpeNTF: A Benchmark Library for Neural Team Formation. In: ACM 2022:3913–3917
23. Rad RH, Seyedsalehi S, Kargar M, Zihayat M, Bagheri E. A Neural Approach to Forming Coherent Teams in Collaboration Networks. In: OpenProceedings.org 2022:2:440–2:444
24. Rad RH, Bagheri E, Kargar M, Srivastava D, Szlichta J. Subgraph Representation Learning for Team Mining. In: ACM 2022:148–153
25. Rad RH, Mitha A, Fani H, Kargar M, Szlichta J, Bagheri E. PyTFL: A Python-based Neural Team Formation Toolkit. In: ACM 2021:4716–4720
26. Rad RH, Bagheri E, Kargar M, Srivastava D, Szlichta J. Retrieving Skill-Based Teams from Collaboration Networks. In: ACM 2021:2015–2019
27. Rad RH, Fani H, Kargar M, Szlichta J, Bagheri E. Learning to Form Skill-based Teams of Experts. In: d'Aquin M, Dietze S, Hauff C, Curry E, Cudré-Mauroux P., eds. *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*ACM 2020:2049–2052
28. Sapienza A, Goyal P, Ferrara E. Deep Neural Networks for Optimal Team Composition. *Frontiers Big Data.* 2019;2:14. doi: 10.3389/fdata.2019.00014
29. Stoyanovich J, Zehlike M, Yang K. Fairness in Ranking: From Values to Technical Choices and Back. In: SIGMOD '23. Association for Computing Machinery 2023; New York, NY, USA:7–12
30. Zehlike M, Sühr T, Baeza-Yates R, Bonchi F, Castillo C, Hajian S. Fair Top-k Ranking with Multiple Protected Groups. *Inf. Process. Manage.*. 2022;59(1). doi: 10.1016/j.ipm.2021.102707
31. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*. 2021;54(6). doi: 10.1145/3457607
32. Loghmani H, Fani H. Bootless Application of Greedy Re-ranking Algorithms in Fair Neural Team Formation. In: Boratto L, Faralli S, Marras M, Stilo G., eds. *Advances in Bias and Fairness in Information Retrieval*Springer Nature Switzerland 2023; Cham:108–118.
33. Calders T, Kamiran F, Pechenizkiy M. Building classifiers with independency constraints. In: IEEE. 2009:13–18.
34. Hardt M, Price E, Srebro N. Equality of Opportunity in Supervised Learning. In: NIPS'16. Curran Associates Inc. 2016; Red Hook, NY, USA:3323–3331.
35. Kou Y, Shen D, Snell Q, et al. Efficient Team Formation in Social Networks based on Constrained Pattern Graph. In: IEEE 2020:889–900
36. Kargar M, Golab L, Srivastava D, Szlichta J, Zihayat M. Effective Keyword Search Over Weighted Graphs. *IEEE Trans. Knowl. Data Eng.*. 2022;34(2):601–616. doi: 10.1109/TKDE.2020.2985376
37. Keane P, Ghaffar F, Malone D. Using machine learning to predict links and improve Steiner tree solutions to team formation problems - a cross company study. *Appl. Netw. Sci.*. 2020;5(1):57. doi: 10.1007/s41109-020-00306-x
38. Zuckerman EW, Jost JT. What Makes You Think You're so Popular? Self-Evaluation Maintenance and the Subjective Side of the "Friendship Paradox". *Social Psychology Quarterly.* 2001;64(3):207–223.
39. Fani H, Barzegar R, Dashti A, Saeedi M. A Streaming Approach to Neural Team Formation Training. In: ECIR'24. 2024.
40. Kaw S, Kobti Z, Selvarajah K. Transfer Learning with Graph Attention Networks for Team Recommendation. In: IEEE 2023:1–8
41. Dong Y, Chawla NV, Swami A. metapath2vec: Scalable Representation Learning for Heterogeneous Networks. In: ACM 2017:135–144
42. Velickovic P, Fedus W, Hamilton WL, Liò P, Bengio Y, Hjelm RD. Deep Graph Infomax. In: OpenReview.net 2019.
43. Chen RJ, Wang JJ, Williamson DF, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering.* 2023;7(6):719–742.
44. Grote T, Keeling G. Enabling fairness in healthcare through machine learning. *Ethics and Information Technology.* 2022;24(3):39.
45. Lee MK, Rich K. Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. In: CHI '21. Association for Computing Machinery 2021; New York, NY, USA
46. Ahmad MA, Patel A, Eckert C, Kumar V, Teredesai A. Fairness in Machine Learning for Healthcare. In: KDD '20. Association for Computing Machinery 2020; New York, NY, USA:3529–3530
47. Bigdeli A, Arabzadeh N, SeyedSalehi S, Zihayat M, Bagheri E. Gender Fairness in Information Retrieval Systems. In: SIGIR '22. Association for Computing Machinery 2022; New York, NY, USA:3436–3439
48. Ekstrand MD, Burke R, Diaz F. Fairness and Discrimination in Retrieval and Recommendation. In: SIGIR'19. Association for Computing Machinery 2019; New York, NY, USA:1403–1404
49. Lv B, Liu F, Li Y, Nie J, Gou F, Wu J. Artificial Intelligence-Aided Diagnosis Solution by Enhancing the Edge Features of Medical Images. *Diagnostics.* 2023;13(6). doi: 10.3390/diagnostics13061063
50. Jalal A, Karmalkar S, Hoffmann J, Dimakis A, Price E. Fairness for image generation with uncertain sensitive attributes. In: PMLR. 2021:4721–4732.
51. Yee K, Tantipongpipat U, Mishra S. Image Cropping on Twitter: Fairness Metrics, Their Limitations, and the Importance of Representation, Design, and Agency. *Proc. ACM Hum.-Comput. Interact.*. 2021;5(CSCW2). doi: 10.1145/3479594

52. Kyriakou K, Barlas P, Kleanthous S, Otterbacher J. Fairness in Proprietary Image Tagging Algorithms: A Cross-Platform Audit on People Images. *Proceedings of the International AAAI Conference on Web and Social Media.* 2019;13(01):313-322. doi: 10.1609/icwsm.v13i01.3232
53. Karako C, Manggala P. Using Image Fairness Representations in Diversity-Based Re-Ranking for Recommendations. In: UMAP '18. Association for Computing Machinery 2018; New York, NY, USA:23–28
54. Nandy P, DiCiccio C, Venugopalan D, Logan H, Basu K, El Karoui N. Achieving Fairness via Post-Processing in Web-Scale Recommender Systems. In: FAccT '22. Association for Computing Machinery 2022; New York, NY, USA:715–725
55. Saito Y, Joachims T. Fair Ranking as Fair Division: Impact-Based Individual Fairness in Ranking. In: KDD '22. Association for Computing Machinery 2022; New York, NY, USA:1514–1524
56. Zehlike M, Castillo C. Reducing Disparate Exposure in Ranking: A Learning To Rank Approach. In: WWW '20. Association for Computing Machinery 2020; New York, NY, USA:2849–2855
57. Geyik SC, Ambler S, Kenthapadi K. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. In: KDD '19. Association for Computing Machinery 2019; New York, NY, USA:2221–2231.
58. Beutel A, Chen J, Doshi T, et al. Fairness in Recommendation Ranking through Pairwise Comparisons. In: KDD '19. Association for Computing Machinery 2019; New York, NY, USA:2212–2220
59. Singh A, Joachims T. Fairness of Exposure in Rankings. In: KDD '18. Association for Computing Machinery 2018; New York, NY, USA:2219–2228
60. Zehlike M, Bonchi F, Castillo C, Hajian S, Megahed M, Baeza-Yates R. FA*IR: A Fair Top-k Ranking Algorithm. In: CIKM '17. Association for Computing Machinery 2017; New York, NY, USA:1569–1578
61. John PG, Vijaykeerthy D, Saha D. Verifying individual fairness in machine learning models. In: PMLR. 2020:749–758.
62. Zehlike M, Yang K, Stoyanovich J. Fairness in Ranking, Part I: Score-Based Ranking. *ACM Comput. Surv.*. 2022;55(6). doi: 10.1145/3533379
63. Altenburger K, De R, Frazier K, Avteniev N, Hamilton J. Are There Gender Differences in Professional Self-Promotion? An Empirical Case Study of LinkedIn Profiles Among Recent MBA Graduates. *Proceedings of the International AAAI Conference on Web and Social Media.* 2017;11(1):460-463. doi: 10.1609/icwsm.v11i1.14929
64. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through Awareness. In: ITCS '12. Association for Computing Machinery 2012; New York, NY, USA:214–226
65. Yang K, Stoyanovich J. Measuring Fairness in Ranked Outputs. In: SSDBM '17. Association for Computing Machinery 2017; New York, NY, USA
66. Mehrotra A, Vishnoi N. Fair ranking with noisy protected attributes. *Advances in Neural Information Processing Systems.* 2022;35:31711–31725.
67. Biega AJ, Gummadi KP, Weikum G. Equity of Attention: Amortizing Individual Fairness in Rankings. In: SIGIR '18. Association for Computing Machinery 2018; New York, NY, USA:405–414
68. Lahoti P, Gummadi KP, Weikum G. iFair: Learning Individually Fair Data Representations for Algorithmic Decision Making. In: 2019:1334-1345
69. Singh A, Joachims T. *Policy Learning for Fairness in Ranking*; Red Hook, NY, USA: Curran Associates Inc. . 2019.
70. Feng Y, Shah C. Has ceo gender bias really been fixed? adversarial attacking and improving gender fairness in image search. In: . 36. 2022:11882–11890.
71. Biega AJ, Gummadi KP, Weikum G. Equity of Attention: Amortizing Individual Fairness in Rankings. *CoRR.* 2018;abs/1805.01788.
72. Barocas S, Hardt M, Narayanan A. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.
73. Juárez J, Brizuela CA. A multi-objective formulation of the team formation problem in social networks: preliminary results. In: Aguirre HE, Takadama K., eds. *GECCO*ACM 2018:261–268
74. Hayano M, Hamada D, Sugawara T. Role and member selection in team formation using resource estimation for large-scale multi-agent systems. *Neurocomputing.* 2014;146:164–172. doi: 10.1016/j.neucom.2014.04.059
75. Majumder A, Datta S, Naidu KVM. Capacitated team formation problem on social networks. In: Yang Q, Agarwal D, Pei J., eds. *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*ACM 2012:1005–1013
76. Li C, Shan M. Team Formation for Generalized Tasks in Expertise Social Networks. In: Elmagarmid AK, Agrawal D., eds. *Proceedings of the 2010 IEEE Second International Conference on Social Computing, SocialCom / IEEE International Conference on Privacy, Security, Risk and Trust, PASSAT 2010, Minneapolis, Minnesota, USA, August 20-22, 2010*IEEE Computer Society 2010:9–16
77. Ani ZC, Yasin A, Husin MZ, Hamid ZA. A method for group formation using genetic algorithm. *International Journal on Computer Science and Engineering.* 2010;2(9):3060–3064.
78. Fitzpatrick E, Askin RG. Forming effective worker teams with multi-functional skill requirements. *Comput. Ind. Eng.*. 2005;48(3):593–608. doi: 10.1016/j.cie.2004.12.014
79. Abdollahpouri H, Mansoury M, Burke R, Mobasher B, Malthouse EC. User-centered Evaluation of Popularity Bias in Recommender Systems. In: Masthoff J, Herder E, Tintarev N, Tkalcic M., eds. *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, Utrecht, The Netherlands, June, 21-25, 2021*ACM 2021:119–129
80. Elahi M, Kholgh DK, Kiarostami MS, Saghari S, Rad SP, Tkalcic M. Investigating the impact of recommender systems on user-based and item-based popularity bias. *Inf. Process. Manag.*. 2021;58(5):102655. doi: 10.1016/J.IPM.2021.102655
81. Genderize . Genderize API. https://genderize.io/; 2023. Accessed: 16-June-2023.
82. Alkhathlan M, Cachel K, Shrestha H, Harrison L, Rundensteiner EA. Balancing Act: Evaluating People's Perceptions of Fair Ranking Metrics. In: ACM 2024:1940–1970
83. Zhu Z, Wang J, Caverlee J. Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems. In: Huang JX, Chang Y, Cheng X, et al., eds. *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*ACM 2020:449–458
84. Moasses R, Rajaei D, Loghmani H, Saeedi M, Fani H. [inline-graphic not available: see fulltext] : Mitigating Gender Bias in Neural Team Recommendation via Female-Advocate Loss Regularization. In: Bellogín A, Boratto L, Kleanthous S, Lex E, Malloci FM, Marras M., eds. *Advances in Bias and Fairness in Information Retrieval - 5th International Workshop, BIAS 2024, Washington, DC, USA, July 18, 2024, Revised Selected Papers*. 2227 of *Communications in Computer and Information Science*. Springer 2024:78–90
85. Cherumanal SP, Spina D, Scholer F, Croft WB. Evaluating Fairness in Argument Retrieval. In: Demartini G, Zuccon G, Culpepper JS, Huang Z, Tong H., eds. *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*ACM 2021:3363–3367
86. Ghosh A, Dutt R, Wilson C. When Fair Ranking Meets Uncertain Inference. In: Diaz F, Shah C, Suel T, Castells P, Jones R, Sakai T., eds. *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15,*

*2021*ACM 2021:1033–1043

87. Geyik SC, Ambler S, Kenthapadi K. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. In: ACM 2019:2221–2231.
88. Seth A, Hemani M, Agarwal C. Dear: Debiasing vision-language models with additive residuals. In: 2023:6820–6829.
89. Morik M, Singh A, Hong J, Joachims T. Controlling Fairness and Bias in Dynamic Learning-to-Rank. In: Huang JX, Chang Y, Cheng X, et al., eds. *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*ACM 2020:429–438
90. Diaz F, Mitra B, Ekstrand MD, Biega AJ, Carterette B. Evaluating Stochastic Rankings with Expected Exposure. In: d'Aquin M, Dietze S, Hauff C, Curry E, Cudré-Mauroux P., eds. *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*ACM 2020:275–284
91. Patro GK, Biswas A, Ganguly N, Gummadi KP, Chakraborty A. FairRec: Two-Sided Fairness for Personalized Recommendations in Two-Sided Platforms. In: Huang Y, King I, Liu T, Steen vM., eds. *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*ACM / IW3C2 2020:1194–1204
92. Qi T, Wu F, Wu C, et al. ProFairRec: Provider Fairness-aware News Recommendation. In: Amigó E, Castells P, Gonzalo J, Carterette B, Culpepper JS, Kazai G., eds. *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*ACM 2022:1164–1173
93. Mansoury M, Abdollahpouri H, Pechenizkiy M, Mobasher B, Burke R. A Graph-Based Approach for Mitigating Multi-Sided Exposure Bias in Recommender Systems. *ACM Trans. Inf. Syst.*. 2022;40(2):32:1–32:31. doi: 10.1145/3470948
94. Balagopalan A, Jacobs AZ, Biega AJ. The Role of Relevance in Fair Ranking. In: Chen H, Duh WE, Huang H, Kato MP, Mothe J, Poblete B., eds. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*ACM 2023:2650–2660
95. Heuss M, Sarvi F, Rijke dM. Fairness of Exposure in Light of Incomplete Exposure Estimation. In: Amigó E, Castells P, Gonzalo J, Carterette B, Culpepper JS, Kazai G., eds. *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*ACM 2022:759–769
96. Wu H, Mitra B, Ma C, Diaz F, Liu X. Joint Multisided Exposure Fairness for Recommendation. In: Amigó E, Castells P, Gonzalo J, Carterette B, Culpepper JS, Kazai G., eds. *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*ACM 2022:703–714