Bootless Application of Greedy Re-ranking Algorithms in Fair Neural Team Formation

Hamed Loghmani and Hossein Fani
[0000-0002-3857-4507], [0000-0002-6033-6564]

University of Windsor, Canada
{ghasrlo, hfani}@uwindsor.ca

Abstract. Team formation aims to automate forming teams of experts who can successfully solve difficult tasks, which have firsthand effects on creating organizational performance. While state-of-the-art neural team formation methods are able to efficiently analyze massive collections of experts to form effective collaborative teams, they largely ignore the fairness in recommended teams of experts. Fairness breeds innovation and increases teams' success by enabling a stronger sense of community, reducing conflict, and stimulating more creative thinking. In this paper, we study the application of state-of-the-art deterministic greedy re-ranking algorithms to mitigate the potential popularity bias in the neural team formation models based on *equality of opportunity*. Our experiments show that, first, neural team formation models are biased toward popular experts. Second, although deterministic re-ranking algorithms mitigate popularity bias substantially, they severely hurt the efficacy of teams. The code to reproduce the experiments reported in this paper is available at `https://github.com/fani-lab/Adila/tree/bias23`[1].

Introduction

Algorithmic search for collaborative teams, also known as team formation, aims to automate forming teams of experts whose combined skills, applied in coordinated ways, can successfully solve complex tasks such as producing the next blockbuster *'thriller'* with a touch of *'sci-fi'* in the movie industry. Team formation can be seen as social information retrieval (Social IR) where the right group of talented people are searched and hired to solve the task at hand [1,2]. Successful teams have firsthand effects on creating organizational performance in the industry [3,4,5], academia [6,7,8], law [9,10], and the healthcare sector [11,12]. Forming a successful team whose members can effectively collaborate and deliver the outcomes within the constraints such as planned budget and timeline is challenging due to the immense number of candidates with various backgrounds, personality traits, and skills, as well as unknown synergistic balance among them; not *all* teams with the best experts are necessarily successful [13].

[1] A feminine Arabic given name, عادلة, meaning just and fair.

Historically, teams have been formed by relying on human experience and instinct, resulting in suboptimal team composition due to (1) an overwhelming number of candidates, and (2) hidden societal biases, among other reasons. To address the former, the earliest algorithmic methods of team formation were conceived in the *i*) Operations Research (OR) [14], where multiple objective functions must be optimized in a large search space of *all* possible combinations of skillful experts, given constraints for human and non-human factors as well as scheduling preferences. Such work, however, was premised on the mutually independent selection of experts and overlooked the organizational and collaborative ties among experts. Next, *ii*) social network analysis has been employed to fill the gap by the network representation of the experts with links that shows collaborations in the past [15,16,17]. They search for the optimum teams over *all* possible subnetworks, which is daunting. Recently, *iii*) a paradigm shift to machine learning has been observed, opening doors to the analysis of massive collections of experts coming from different fields. Machine learning approaches efficiently learn relationships between experts and their skills in the context of successful (positive samples) and unsuccessful teams (negative samples) from all past instances to excel at recommending teams of experts [18,19,20]. We can observe the commercial application of machine learning-based algorithmic search for an optimum team in online platforms like LinkedIn[2] to help the industry browse the enormous space of experts and form *almost surely* successful teams.

However, the primary focus of existing machine learning-based methods in team formation is the maximization of the success rate (utility) by tailoring the recommended experts for a team to the required skills only, largely ignoring the *fairness* in recommended experts. Indeed, it has been well-explored that machine learning methods that produce recommendations suffer from unfair biases. They result in discrimination and reduced visibility for an already disadvantaged group [21,22], disproportionate selection of popular candidates [23,24,25], and over/under-representation and racial/gender disparities [26] since they are trained on real-world datasets that already inherit hidden societal biases. On the other hand, social science research provides compelling evidence about the synergistic effects of diversity on team performance [27,28,29]; diversity breeds innovation and increases teams' success by enabling a stronger sense of community and support, reducing conflict, and stimulating more creative thinking.

Surprisingly, there is little to no fairness-aware algorithmic method that mitigates societal biases in team formation algorithms except that of the recent work by Barnabò et al. [30] that proves fair team formation is NP-complete; therefore, computationally prohibitive for practical use. Recent state-of-the-art neural team formation models have *weakly* attributed their performance gain to mitigating popularity bias inherent in the underlying real-world training data [20,19]. Rad et al. [20] employed uncertainty in learnable parameters by variational Bayesian neural model, and Dashti et al. [19] applied *virtually* negative samples from popular experts during the neural model learning procedure. However, they overlook substantiating the attribution by evidence using fairness metrics.

[2] business.linkedin.com/talent-solutions

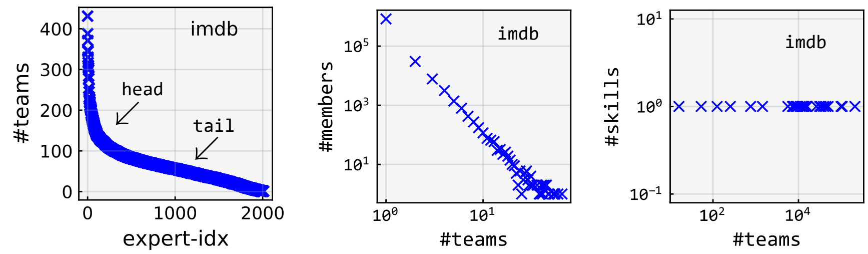

Fig. 1. Left: Long-tail distribution of casts and crews (experts) in movies (teams). Middle: Long-tail distribution in `log` scale. The figure reads y number of members have x number of teams. Right: uniform distribution of movies over genres (skills)

A purely diversity-centric design for team formation algorithms that solely overfit to satisfy diversity, neglecting the team's success, is also unfair to the organizations, e.g., a team of *non*popular individuals who cannot accomplish the tasks. In this paper, we propose to model team formation as a two-sided marketplace between two stakeholders: i) *experts* who hold skills, e.g., artists, and ii) *organizations* who recruit experts for their teams, e.g., entertainment industries. We investigate the trade-off between success rate (utility) and fairness in the recommended teams by neural team formation methods in terms of popularity bias, given the required skills. The choice of popularity bias in this study is motivated due to: (1) training sets in team formation suffer from popularity bias; that is, the majority of experts have scarcely participated in the (successful) teams (nonpopular experts), whereas few experts (popular ones) are in many teams [20,19]. Therefore, popular experts receive higher scores and are more frequently recommended by the machine learning model, leading to systematic discrimination against already disadvantaged nonpopular experts. Statistically, popularity bias can be observed as long tail distribution (power law). For instance, in `imdb`[3] dataset of movies, given a movie as a team of casts and crews such as actors and directors [31,16], from Fig. 1(left), we observe a long tail of many nonpopular experts, while few popular experts in the head that dominate. Fig. 1(middle) shows the same observation in `log` scale based on y number of experts participating in x number of teams. (2) Moreover, experts' labels of being popular or otherwise can be calculated from datasets based on their position in the statistical distribution; that is, those in the *'tail'* are assumed to be nonpopular experts, while those in the *'head'* are the popular ones.

In this paper, we employ the framework by Geyik et al. [32] for quantifying and mitigating popularity bias in state-of-the-art neural team formation methods [19] in terms of normalized discounted cumulative KL-divergence (`ndkl`) for *re*ranking experts in the recommended teams to achieve fairness based on the *equality of opportunity* (equalized odds) [33] depending on the distribution of teams over popular and nonpopular experts in the training datasets. Meanwhile, we measure the impact of the popularity bias mitigation on the success

[3] imdb.com/interfaces/

rate (utility) of the recommended teams using information retrieval metrics, namely mean average precision (`map`) and normalized discounted cumulative gain (`ndcg`). Our early results on `imdb` using three re-ranking algorithms by Geyik et al. [32] demonstrate that (1) state-of-the-art Bayesian neural models fall short in producing fair teams of experts in terms of popularity, and (2) state-of-the-art deterministic re-ranking algorithms improve the fairness of neural team formation models but at the cost of a substantial decrease in accuracy of predicted teams in terms of success rate. Our findings encourage further development of fairness-aware re-ranking methods for the task of team formation.

Research Methodology

Ranking is the primary output interface of the neural team formation model for producing expert recommendations where all available experts are recommended for a given required subset of skills but with different scores, usually a probability value after a `softmax` layer, and the final recommended experts are selected among the top-k highest scores. This enables further post-processing refinements like re-ranking the list of recommended items to improve fairness in the recommended list. Therefore, our research includes two pipelined steps: *i*) training state-of-the-art neural team formation model to produce experts recommendations for given subsets of skills while measuring the accuracy and diversity of top-k experts as the optimum team, and *ii*) applying state-of-the-art re-ranking algorithms to reorder the top-k experts and to improve fairness while maintaining accuracy. For example, when two or more experts have been assigned the same probability score in the final ranked list by a model, a re-ranking algorithm can prioritize nonpopular experts over popular ones and reassign new higher scores.

We follow the *equality of opportunity* [33] notion of fairness; that is, for being a member of a team (a preferred label that benefits an expert), a neural team formation model should predict an expert's membership with equal odds based on the underlying training dataset for all popular and nonpopular experts. In other words, equality of opportunity measures whether the experts who should qualify for a team are equally *likely* regardless of their popularity status. For instance, given the percentage of popular experts to nonpopular ones is 10% to 90%, the neural model satisfies equality of opportunity for forming a team of k experts should the team include $k \times 10\%$ popular and $k \times 90\%$ nonpopular experts. It is noteworthy that a `random` baseline that assigns experts to teams from a uniform distribution of experts regardless of popularity labels is an *ideally* fair model yet at the cost of very low success rates for the predicted teams.

Intuitively, a few popular experts who participated in many training instances of teams reinforce a neural model to forget about the majority nonpopular experts for their scarce number of teams, leading to popularity bias. As a result, a new predicted team would only include experts from the minority popular experts ($k \times 100\%$), which is disproportionate compared to their population size (10%). In this paper, we aim to dampen the popularity bias by adjusting the distributions of popular and nonpopluar experts in the top-k recommended experts for a team according to their ratio in the training dataset via deterministic algorithms and study the impacts on the team's quality in terms of success rate; that is measuring the accuracy of top-k experts for teams whose all $k \times 100\%$ members are popular experts compared to teams with $k \times 10\%$ popular and $k \times 90\%$ nonpopular experts.

Experiments

In this section, we lay out the details of our experiments and findings toward answering the following research questions:
RQ1: Do state-of-the-art neural team formation models produce fair teams of experts in terms of popularity bias? To this end, we benchmark state-of-the-art Bayesian neural model with negative sampling heuristics [19] and measure the fairness scores of predicted teams.
RQ2: Do state-of-the-art deterministic greedy re-ranking algorithms improve the fairness of neural team formation models while maintaining their accuracy? To this end, we apply three deterministic greedy re-ranking algorithms on the neural model predictions and measure the diversity and utility scores afterwards.

Setup

Dataset. Our testbed includes `imdb`[31,16] dataset where each instance is a movie consisting of its cast and crew such as actors and director, as well as the movie's genres. We consider each movie as a team whose members are the cast and crew, and the movie's genres are the skills. The choice of `imdb` in team formation literature is not to be confused with its use cases in recommender systems or 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. As shown in Fig. 1, we can observe a long tail in the distributions of teams over experts; many casts and crews have participated in very few movies. However, the distribution with respect to the set of skills follows a more fair distribution. Specifically, `imdb` has a limited variety of skills (genres) which are, by and large, employed by many movies. We filter out singleton and sparse movies with less than `3` members as well as casts and crews who relatively participated in very few movies, as suggested by [34,20]. The latter also reduced the computational complexity of the neural models in their last layer where the size equals the number of experts. We ensured that the preprocessing step made no major change to the statistical distributions of the dataset. Table 1 reports additional point-wise statistics on the dataset before and after preprocessing.

Popularity Labels. We label an expert as popular if she participated in more than the average number of teams per expert over the whole dataset, and non-popular otherwise. As seen in Table 1, this number is `62.45` and the popularity ratio (popular/nonpopular) is `0.426/0.574`.

Table 1. Statistics of the raw and preprocessed `imdb` dataset.

	imdb	
	raw	filtered
#movies	507,034	32,059
#unique casts and crews	876,981	2,011
#unique genres	28	23
average #casts and crews per team	1.88	3.98
average #genres per team	1.54	1.76
average #movie per cast and crew	1.09	62.45
average #genre per cast and crew	1.59	10.85
#team w/ single cast and crew	322,918	0
#team w/ single genre	315,503	15,180

Baselines. Our neural team formation baselines include variational Bayesian neural network [20] with unigram negative sampling strategy in minibatches [19] (`bnn`) and Kullback-Leibler optimization. The model includes a single hidden layer of size `d=100`, `leaky relu` and `sigmoid` are the activation functions for the hidden and the output layers, respectively, and `Adam` is the optimizer. The input and output layers are sparse occurrence vector representations (one-hot encoded) of skills and experts of size $|\mathcal{S}|$ and $|\mathcal{E}|$, respectively. Moreover, we also used pre-trained dense vector representations for the input skill subsets (`-emb`). Adapted from paragraph vectors of Le and Mikolov [35], we consider each team as a document and the skills as the document's words. We used the distributed memory model to generate the real-valued embeddings of the subset of skills with a dimension of `d=100`. We evaluate baselines with and without the application of re-ranking methods (`before`, `after`). To have a minimum level of comparison, we also add a model that randomly assigns experts to a team (`random`). The re-ranking methods include the *i*) score maximizing greedy mitigation algorithm (`greedy`), *ii*) greedy conservative mitigation algorithm (`conservative`), and *iii*) the relaxed variant of greedy conservative algorithm (`relaxed`) [32].

Evaluation Strategy and Metrics. To demonstrate prediction effectiveness, we randomly select `15%` of teams for the test set and perform `5`-fold cross-validation on the remaining teams for model training and validation that results in one trained model per each fold. Let (s, e) a team of experts e for the required skills s from the test set, we compare the top-k ranked list of experts e', predicted by the model of each fold for the input skills 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) including mean average precision (`map`) and normalized discounted cumulative gain (`ndcg`) at top-`{2,5,10}`. Formally,

$$\texttt{ap}(k) : \frac{\sum_{i=1}^{k} \texttt{p}(i) \times \delta_e(i)}{|e \cap e'|} \tag{1}$$

where $\texttt{p}(k) = \frac{|e \cap e'|}{k}$ is the precision, i.e., how many of the k predicted experts e' 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. For normalized discounted cumulative gain (`ndcg`),

$$\texttt{dcg}(k) = \sum_{i=1}^{k} \frac{\texttt{rel}(i)}{\texttt{log}(i+1)} \tag{2}$$

where $\texttt{rel}(i)$ 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, $rel(i) = \texttt{1}$ if $i \in e$ and `0` otherwise, and e.q.(2) becomes:

$$\texttt{dcg}(k) = \sum_{i=1}^{k} \frac{\delta_e(i)}{\texttt{log}(i+1)} \tag{3}$$

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.,

$$\texttt{ndcg}(k) = \frac{\sum_{i=1}^{k} \frac{\delta_e(i)}{\texttt{log}(i+1)}}{\sum_{i=1}^{|e|} \frac{1}{\texttt{log}(i+1)}} \tag{4}$$

To evaluate fairness, we used `ndkl` with no cutoff [32] (the lower, the better) with being 0 in the ideal fair cases. Formally, let $d_{e'}$ the distribution of popular and nonpopular experts in the predicted top-k experts e' (the proportions of popular and nonpopular experts) and d_e the ideal fair distribution for a test instance of a team (s, e), the Kullback–Leibler (`kl`) divergence of $d_{e'}$ from d_e is:

$$\texttt{kl}(d_{e'}(k)||d_e(k)) = \sum_{i=1}^{k} d_{e'}(i) \, \texttt{log} \frac{d_{e'}(i)}{d_e(i)} \tag{5}$$

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 (`ndkl`)[32]:

$$\texttt{ndkl}(d_{e'}) = \frac{\sum_{k=1}^{|e|} \frac{1}{\texttt{log}(k+1)} \, \texttt{kl}(d_{e'}(k)||d_e(k))}{\sum_{i=1}^{|e|} \frac{1}{\texttt{log}(i+1)}} \tag{6}$$

Results

In response to **RQ1**, i.e., whether state-of-the-art neural team formation models produce fair teams of experts, from Table 2, we observe that state-of-the-art Bayesian neural models with negative sampling (`bnn` and `bnn_emb`) suffer from popularity bias having regard to their high `ndkl` compared to `random` baseline

Table 2. Average performance of 5-fold on test set in terms of fairness (ndkl; the lower, the better) and utility metrics (map and ndcg, the higher, the better)

bnn[19,20]		greedy		conservative		relaxed	
	before	after	Δ	after	Δ	after	Δ
ndcg2↑	0.695%	0.126%	-0.569%	0.091%	-0.604%	0.146%	-0.550%
ndcg5↑	0.767%	0.141%	-0.626%	0.130%	-0.637%	0.130%	-0.637%
ndcg10↑	1.058%	0.247%	-0.811%	0.232%	-0.826%	0.246%	-0.812%
map2↑	0.248%	0.060%	-0.188%	0.041%	-0.207%	0.063%	-0.185%
map5↑	0.381%	0.083%	-0.298%	0.068%	-0.313%	0.079%	-0.302%
map10↑	0.467%	0.115%	-0.352%	0.101%	-0.366%	0.115%	-0.352%
ndlkl↓	0.2317	0.0276	-0.2041	0.0276	-0.2041	0.0273	-0.2043

bnn_emb[20,19]		greedy		conservative		relaxed	
	before	after	Δ	after	Δ	after	Δ
ndcg2↑	0.921%	0.087%	-0.834%	0.121%	-0.799%	0.087%	-0.834%
ndcg5↑	0.927%	0.117%	-0.810%	0.150%	-0.777%	0.117%	-0.810%
ndcg10↑	1.266%	0.223%	-1.043%	0.241%	-1.025%	0.223%	-1.043%
map2↑	0.327%	0.034%	-0.293%	0.057%	-0.270%	0.034%	-0.293%
map5↑	0.469%	0.059%	-0.410%	0.084%	-0.386%	0.059%	-0.410%
map10↑	0.573%	0.093%	-0.480%	0.111%	-0.461%	0.093%	-0.480%
ndkl↓	0.2779	0.0244	-0.2535	0.0244	-0.2535	0.0241	-0.2539

random		greedy		conservative		relaxed	
	before	after	Δ	after	Δ	after	Δ
ndcg2↑	0.1711%	0.136%	-0.035%	0.205%	0.034%	0.205%	0.034%
ndcg5↑	0.1809%	0.170%	-0.011%	0.190%	0.009%	0.190%	0.009%
ndcg10↑	0.3086%	0.258%	-0.051%	0.283%	-0.026%	0.283%	-0.026%
map2↑	0.0617%	0.059%	-0.003%	0.089%	0.028%	0.089%	0.028%
map5↑	0.0889%	0.095%	0.006%	0.110%	0.021%	0.110%	0.021%
map10↑	0.1244%	0.121%	-0.003%	0.140%	0.016%	0.140%	0.016%
ndkl↓	0.0072	0.0369	0.0296	0.0366	0.0293	0.0366	0.0294

before applying deterministic re-ranking algorithms, thus answering **RQ2** negatively. Indeed, the random baseline which blindly assigns experts to teams is following the experts' popularity label distribution in the training dataset, and hence, yields the best fair model based on *equality of opportunity* (equalized odds). However, random baseline has the lowest utility metric values while bnn and bnn_emb achieve the highest.

In response to **RQ2**, i.e., whether state-of-the-art deterministic re-ranking algorithms improve the fairness of neural team formation models while maintaining their accuracy, from Table 2, although applying all re-ranking algorithms resulted in lower ndkl values by increasing the diversity of experts in the recommended teams, they substantially reduced the teams' accuracy at the same time for all neural models in terms of all utility metrics, proving the ineffectiveness of deterministic greedy re-ranking algorithms for the task of team formation. Among the re-ranking algorithms, `relaxed` is the best since it decreases the `ndkl` of neural models the most while the drop in the utility metrics is the lowest compared to the other two algorithms.

Concluding Remarks

We focused on the problem of fair team formation. We showed that state-of-the-art neural models, which can efficiently learn relationships between experts and their skills in the context of successful and unsuccessful teams from all past instances, suffer from popularity bias. To mitigate the popularity bias while maintaining the success rates of recommended teams, we applied three state-of-the-art deterministic re-ranking algorithms to reorder the final ranked list of experts against the popular experts in favour of nonpopular ones. We found that while deterministic re-ranking algorithms improve the fairness of neural team formation models, they fall short of maintaining accuracy. Our future research directions include *i*) investigating other fairness factors like demographic attributes, including age, race, and gender; and *ii*) developing machine learning-based models using Learning-to-Rank (L2R) techniques to mitigate popularity bias as opposed to deterministic greedy algorithms.

References

1. Damon Horowitz and Sepandar D. Kamvar. The anatomy of a large-scale social search engine. In *WWW*, page 431–440, New York, NY, USA, 2010. Association for Computing Machinery.
2. Damon Horowitz and Sepandar D. Kamvar. Searching the village: Models and methods for social search. *Commun. ACM*, 55(4):111–118, apr 2012.
3. K.M. Bursic. Strategies and benefits of the successful use of teams in manufacturing organizations. *IEEE Transactions on Engineering Management*, 39(3):277–289, 1992.
4. Gholamreza Askari, Nader Asghri, Madjid Eshaghi Gordji, Heshmatolah Asgari, José António Filipe, and Adel Azar. The impact of teamwork on an organization's performance: A cooperative game's approach. *Mathematics*, 8(10), 2020.
5. Almagul Kairgalievna and Nurul Mohammad Zayed. The effect of teamwork on employee productivity. 2021.
6. Julie Younglove-Webb, Barbara Gray, Charles William Abdalla, and Amy Purvis Thurow. The dynamics of multidisciplinary research teams in academia. *The Review of Higher Education*, 22(4):425–440, 1999.
7. Erin Leahey. From sole investigator to team scientist: Trends in the practice and study of research collaboration. *Annual Review of Sociology*, 42(1):81–100, 2016.
8. Kara Hall, Amanda Vogel, Grace Huang, Katrina Serrano, Elise Rice, Sophia Tsakraklides, and Stephen Fiore. The science of team science: A review of the empirical evidence and research gaps on collaboration in science. *American Psychologist*, 73:532–548, 05 2018.
9. Peter D. Sherer. Leveraging human assets in law firms: Human capital structures and organizational capabilities. *ILR Review*, 48(4):671–691, 1995.

10. Jia Hu and Robert C. Liden. Making a difference in the teamwork: Linking team prosocial motivation to team processes and effectiveness. *Academy of Management Journal*, 58:1102–1127, 2014.
11. Maxine Craig and Debi McKeown. How to build effective teams in healthcare. *Nursing times*, 111(14):16—18, 2015.
12. Michael A. Rosen, Deborah DiazGranados, Aaron S. Dietz, Lauren E. Benishek, David Thompson, Peter J. Pronovost, and Sallie J. Weaver. Teamwork in healthcare: Key discoveries enabling safer, high-quality care. *American Psychologist*, 73(4):433 – 450, 2018. Cited by: 297; All Open Access, Green Open Access.
13. Roderick I. Swaab, Michael Schaerer, Eric M. Anicich, Richard Ronay, and Adam D. Galinsky. The too-much-talent effect: Team interdependence determines when more talent is too much or not enough. *Psychological Science*, 25(8):1581–1591, 2014.
14. Fahimeh Rahmanniyay, Andrew Junfang Yu, and Javad Seif. A multi-objective multi-stage stochastic model for project team formation under uncertainty in time requirements. *Computers & Industrial Engineering*, 132:153–165, 2019.
15. Theodoros Lappas, Kun Liu, and Evimaria Terzi. Finding a team of experts in social networks. In *SIGKDD 2009*, pages 467–476. ACM, 2009.
16. Mehdi Kargar and Aijun An. Discovering top-k teams of experts with/without a leader in social networks. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 985–994, 2011.
17. Mehdi Kargar and Aijun An. Efficient top-k keyword search in graphs with polynomial delay. In *2012 IEEE 28th International Conference on Data Engineering*, pages 1269–1272, 2012.
18. Radin Hamidi Rad, Aabid Mitha, Hossein Fani, Mehdi Kargar, Jaroslaw Szlichta, and Ebrahim Bagheri. Pytfl: A python-based neural team formation toolkit. In *CIKM*, pages 4716–4720. ACM, 2021.
19. Arman Dashti, Saeed Samet, and Hossein Fani. Effective neural team formation via negative samples. In *CIKM*, pages 3908–3912. ACM, 2022.
20. Radin Hamidi Rad, Hossein Fani, Mehdi Kargar, Jaroslaw Szlichta, and Ebrahim Bagheri. Learning to form skill-based teams of experts. In *CIKM '20*, pages 2049–2052. ACM, 2020.
21. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226. ACM, 2012.
22. Sara Hajian, Francesco Bonchi, and Carlos Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 2125–2126. ACM, 2016.
23. Emre Yalcin and Alper Bilge. Investigating and counteracting popularity bias in group recommendations. *Inf. Process. Manag.*, 58(5):102608, 2021.
24. Ziwei Zhu, Yun He, Xing Zhao, and James Caverlee. Popularity bias in dynamic recommendation. In *KDD*, pages 2439–2449. ACM, 2021.
25. Jianing Sun, Wei Guo, Dengcheng Zhang, Yingxue Zhang, Florence Regol, Yaochen Hu, Huifeng Guo, Ruiming Tang, Han Yuan, Xiuqiang He, and Mark Coates. A framework for recommending accurate and diverse items using bayesian graph convolutional neural networks. In *KDD*, pages 2030–2039. ACM, 2020.
26. Matthew Kay, Cynthia Matuszek, and Sean A. Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, pages 3819–3828. ACM, 2015.

27. Cara Tannenbaum, Robert P Ellis, Friederike Eyssel, James Zou, and Londa Schiebinger. Sex and gender analysis improves science and engineering. *Nature*, 575(7781):137–146, 2019.
28. Jakob Lauring and Florence Villesèche. The performance of gender diverse teams: what is the relation between diversity attitudes and degree of diversity? *European Management Review*, 16(2):243–254, 2019.
29. Bas Hofstra, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, 117(17):9284–9291, 2020.
30. Giorgio Barnabò, Adriano Fazzone, Stefano Leonardi, and Chris Schwiegelshohn. Algorithms for fair team formation in online labour marketplaces. In *WWW*, pages 484–490, 2019.
31. Mehdi Kargar, Lukasz Golab, Divesh Srivastava, Jaroslaw Szlichta, and Morteza Zihayat. Effective keyword search over weighted graphs. *IEEE Trans. Knowl. Data Eng.*, 34(2):601–616, 2022.
32. Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *KDD*, pages 2221–2231. ACM, 2019.
33. Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.
34. Arman Dashti, Saeed Samet, and Hossein Fani. Effective neural team formation via negative samples. In *CIKM*, page 3908–3912, New York, NY, USA, 2022. Association for Computing Machinery.
35. Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, ICML'14, page II–1188–II–1196. JMLR.org, 2014.