1-31-2024

Adila: Fairness-informed Collaborative Team Formation

Hamed Ghasr Loghmani
University of Windsor

Adila: Fairness-informed Collaborative Team Formation

By

Hamed Ghasr Loghmani

A Thesis
Submitted to the Faculty of Graduate Studies
through the School of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science
at the University of Windsor

Windsor, Ontario, Canada

2024

Adila: Fairness-informed Collaborative Team Formation

by

Hamed Ghasr Loghmani

APPROVED BY:

N. Delia
Department of Interdisciplinary and Critical Studies

B. Boufama
School of Computer Science

H. Fani, Advisor
School of Computer Science

January 11, 2024

DECLARATION OF CO-AUTHORSHIP AND PREVIOUS PUBLICATION

I. Co-Authorship

I hereby declare that this thesis incorporates material that is the result of joint research, as follows:

Chapter 2 of the thesis includes the outcome of a publication from my contributions, under the supervision of Dr. Hossein Fani. Chapter 3 incorporates unpublished material co-authored with Reza Barzegar, Gabriel Rueda and Edwin Paul under the supervision of Dr. Hossein Fani. Reza Barzegar contributed through refining the manuscript and team formation literature review. Gabriel Rueda and Edwin Paul, contributed through the implementation of graph generation parts and proof reading the manuscript. Chapter 4 also contains unpublished material co-authored with Mahdis Saeedi, Gabriel Rueda and Edwin Paul under the supervision of Dr. Hossein Fani. Mahdis Saeedi contributed through refining the mathematical formulation, team recommendation literature review and manuscript refining. Gabriel Rueda and Edwin Paul contributed through the implementation of figure generation functions, result averaging and sorting, providing the implementation of gender labeling for our datasets and proof reading the manuscript. Finally, I acknowledge that in all cases the key ideas, primary contributions, experimental designs, data analysis, interpretation, and writing were performed by myself.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my thesis, and have obtained written permission from each of the co-author(s) to include the above material(s) in my thesis.

I certify that, with the above qualification, this thesis, and the research to which it refers, is the product of my own work.

II. Previous Publication

This thesis includes 3 original papers and 2 poster presentations that have been previously published/submitted for publication in peer reviewed journals or conferences, as follows:

Thesis Chapter	Publication title/full citation	Publication Status
Chapter 2	Loghmani, H., Fani, H. (2023). 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. BIAS 2023. Communications in Computer and Information Science, vol 1840. Springer, Cham. https://doi.org/10.1007/978-3-031-37249-0_9	Published
Chapter 3	OpeNTF2: A Framework for Fair Team Formation	Ready for Submission
Chapter 4	A Probabilistic Greedy Attempt to be Fair in Neural Team Recommendation	Ready for Submission

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my thesis. I certify that the above material describes work completed during my registration as a graduate student at the University of Windsor

III. General

I declare that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

ABSTRACT

Team formation aims at forming a collaborative group of experts to accomplish complex tasks, which is a recognized objective in the industry. While state-of-the-art neural team formation models can efficiently analyze massive sets of candidate experts to form effective collaborative teams, they overlook fairness. In this work, we adopt state-of-the-art probabilistic and deterministic greedy reranking algorithms to achieve fairness with respect to (1) popularity or (2) gender in neural models in view of two notions of fairness, demographic parity and equality of opportunity. Specifically, we ensure a minimum representation for experts from the disadvantaged, nonpopular or female, groups by reranking the neural model's ranked list of recommended experts. Our experiments on two large-scale benchmark datasets demonstrate three key findings: (i) neural team formation models heavily suffer from biases toward popular and male experts; (ii) probabilistic greedy reranking algorithms can substantially mitigate such biases while maintaining teams' efficacy; (iii) in the presence of extreme biases, e.g., 95% male vs. 5% female experts in the training datasets, post-processing reranking methods alone fall short, urging further tandem integration of pre-process and in-process debiasing techniques.

ACKNOWLEDGEMENTS

In every journey of life, including this one, each event, whether positive or negative, has led me to this destination. For that, I am grateful for all these experiences collectively. Rather than listing names and turning this into a typical acknowledgment that might go unnoticed, I choose to express my gratitude for the various forms of support I received throughout this journey.

I am deeply grateful to those who illuminated my path when it was dark and patiently held the light as I took my first uncertain steps into this new chapter. I appreciate those who pointed out the obstacles on my journey, preventing my fall, and those who offered a helping hand to lift me back to my feet when I fell nonetheless. I am thankful for the presence of those who gently highlighted my mistakes and kindly forgave them. Additionally, I am grateful for those who showed me their best version, even when I was far from perfection. My heartfelt thanks also go to the strangers who may have sparked an idea in my mind, brightened my day with a warm smile, or offered words of encouragement during a brief conversation. Last but not least, the journey of life would be meaningless without the people we love and care about, from our first steps to our last breath. I want to thank all my beloved people for always being there for me. For turning life into something that makes sense. This is an opportunity to let them know I was resilient on many occasions because of their support. I took that one extra step out of my comfort zone multiple times just to make them proud. And finally, always tried to become a better person because they deserve nothing except perfection.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction to Fairness in Artificial Intelligence

The theory of fairness is a complex and evolving domain in moral and political philosophy, primarily concerned with how resources, opportunities, and rights should be equitably distributed among individuals within a society. It focuses on the intricacies of equality and equity, where equality emphasizes identical treatment and allocation for all, while equity recognizes the diverse needs and circumstances of individuals, advocating for a distribution that addresses these differences to ensure equal opportunities for success.

A significant figure in this theory is philosopher John Rawls, whose *Theory of Justice* introduces the concepts of the *original position* and the *veil of ignorance*. This thought experiment suggests that justice's principles are those that individuals would choose if they were unaware of their position in society, such as their class, race, gender, and personal abilities, ensuring decisions are made impartially. These attributes are often called *sensitive(protected) attributes*. Another perspective within the theory of fairness is utilitarianism, which equates right actions with those that promote overall happiness. This approach, however, sometimes conflicts with individual rights, as it prioritizes the majority's welfare, potentially overlooking what is fair for individuals or minority groups.

In practical applications, fairness theory significantly influences law, education, employment, and healthcare, constantly posing questions about equitable resource

distribution and fair decision-making processes in social, political, and legal institutions. For instance, it informs policies on income distribution, access to education and healthcare, and the design of legal systems to ensure fair procedures. Despite its comprehensive approach, the theory of fairness faces critiques for its potential idealism and the difficulty of applying its principles uniformly across diverse cultural and societal contexts. Critics argue that the theory may oversimplify the complexities inherent in real-world situations and struggle to establish universal fairness principles applicable in all contexts. This ongoing debate and the development of the fairness theory highlight its crucial role in shaping our understanding of justice and equity. As societies evolve and face new ethical challenges, the theory of fairness continues to adapt, offering insights into creating a more equitable and just world, balancing individual needs with collective welfare, and navigating the delicate interplay between universal principles and contextual realities.

1.1 Automated Decision Making

Automated decision-making using artificial intelligence (AI) provides a significant shift in how decisions are made in different domains, finance to healthcare and criminal justice. This technology utilizes the power of AI algorithms to analyze extensive amounts of data, recognize patterns, and make decisions without human intervention. The escalation of employing AI in decision-making processes promises increased efficiency, objectivity, and the ability to process information at a scale unreachable by humans. At the center of automated decision-making is the use of machine learning algorithms. These algorithms are trained on large datasets to recognize patterns and learn from them. After the training phase, they can make predictions or decisions based on new data they witness. For instance, in the education sector, an AI-based algorithm can analyze past data for admission into a university, and make decisions regarding the admission of new students presented to the system. In healthcare, AI systems can process patient data to assist in diagnosing diseases or recommending treatment plans[11].

The main advantage of using AI for decision-making is its ability to examine huge loads of data and also its efficiency. AI systems can process and investigate data more quickly than humans, enabling real-time decision-making that is critical in many areas such as financial trade markets, risk management and prevention systems, and healthcare. Furthermore, AI can provide insights from data that might be overlooked by human experts, leading to more informed and potentially innovative decisions. Moreover, decisions made by AI have the potential to be unbiased, theoretically leading to outcomes that are more objective compared to those influenced by human prejudices. Nonetheless, the use of AI for automated decision-making comes with its own set of challenges. A major issue is the possibility of intrinsic biases within the AI systems, originating from either biased training datasets or biases in the algorithms. Such biases can result in decision outcomes that are unfair or discriminatory, especially in critical fields such as employment selection or legal sentencing. Another challenge is the lack of transparency and explainability in some AI systems, making it difficult to understand or challenge the decisions made by these algorithms[8]. In *Algorithms of Oppression*[7], Safiya Noble explores the connection between the absence of transparency in algorithms and their role in oppressing protected groups. Noble emphasizes that a lack of transparency in the functioning of algorithms and decision-making systems sets the stage for the emergence and perpetuation of biases. She points out that, despite their automated nature, these systems are predominantly overseen by human agents and large corporations. These entities, as critically observed, have a track record of manipulating rules and distorting facts to suit their own interests and profit motives. This critical insight stresses the need for greater ethical scrutiny and transparency in the development and deployment of algorithmic systems[7]. These issues have given rise to an increasing demand for ethical considerations and the establishment of regulatory frameworks to govern the use of AI in decision-making processes. Topics like data privacy, informed consent, and the necessity for explanations play a key role in these debates. There is a growing emphasis among policymakers and industry executives on crafting standards and guidelines aimed at making AI-driven decision-making processes transparent, equitable, and accountable[4]. These concerns

fall into *Responsible AI* research area. As we look to the future, the incorporation of AI in decision-making is anticipated to increase, driven by advancements in AI technology that enable more complex applications. The evolution of AI in decision-making carries the potential for not only enhancing efficiency and effectiveness but also for revolutionizing the methodologies of problem-solving across diverse sectors. Nevertheless, it will be essential to strike a balance between the advantages of AI and its ethical implications and possible risks, to guarantee that this technology contributes positively to the greater good and improves decision-making processes in a way that is responsible and fair.

1.2 Legitimacy of Automated Decision Making

The discussion about legitimacy is not centered on how different groups are treated comparatively, instead it is about whether it is fair to deploy such a system at all in a given scenario. That question, in return, impacts the credibility of the institution implementing it. The majority of organizations require credibility to operate efficiently. Individuals must perceive that an institution generally conforms to societal values. This necessity is quite evident in the context of public entities like governments or educational institutions, which hold direct or indirect accountability to the population. The rationale for private companies needing legitimacy, however, is less obvious. A possible explanation is that the greater the authority a company wields over individuals, the more crucial it is for that exercise of power to be viewed as legitimate. Making decisions that affect people signifies exerting power over them, making it vital to secure legitimacy. In the absence of perceived legitimacy, individuals may resort to different forms of resistance, particularly through legal means. Furthermore, a decline in legitimacy can adversely impact a company's competitive edge in the marketplace. The issue of corporate legitimacy has frequently surfaced in the digital technology sector. Companies that base their business models on personal data, particularly when gathered covertly, have also experienced crises in public perception. Beyond legal repercussions, these companies have faced competition leveraging their weak

privacy policies. For example, Apple's decision to limit Facebook's ability to track iOS users impacted Facebook's earnings[6]. This action was publicly supported, despite strong objections from Facebook, likely because its fundamental business model had suffered a loss of legitimacy. Debates about legitimacy have often been eclipsed by conversations about bias and discrimination within the context of fairness. Advocates frequently opt to concentrate on aspects of distribution as a strategy to challenge legitimacy, as it is typically a more straightforward argument to present. However, this approach can have unintended consequences. Numerous companies have adopted the language of fairness discourse, finding it relatively simple to achieve demographic parity in decisions without truly addressing the underlying issues of legitimacy[10].

1.3 Different Types of Bias in AI-Based Systems

Throughout the stages of model construction, training, and real-world implementation, the biases may manifest for a variety of reasons and can be put in different groups[13]: *i) Pre-existing bias* reflects historical or societal prejudices and stereotypes that are inherently present in the data such as gender disparities in the tech industry. *ii) Technical bias* surface due to limitations, speculations, or built-in qualities of algorithms, methods, or design choices. *iii) Emergent bias* arises during the operation or deployment of an AI-based system which was not initially biased or intended to generate biased outcomes. For instance, an AI chatbot might start generating biased responses after being exposed to biased or discriminative responses from users. Given the increasing prevalence of neural models in automated decision-making, the imperative to address and mitigate biases not only becomes a moral compass for equitable outcomes, but also a catalyst for unlocking the multifaceted benefits of truly fair and representative decision-making systems. So, the key worry is that machine learning and neural models may generate outputs that consistently make already disadvantaged groups and legally protected categories like people with disabilities, racial/ethnic minorities, or underrepresented genders in certain professions even less visible[3]. A variety of methods have been applied in different approaches to mitigate

each type of bias and their categorization will be discussed in the next subsection.

1.3.1 Pre-Existing Bias

Pre-existing bias in ranking refers to the presence of biases or prejudices that exist prior to the ranking process resulting in biases in the ranking outcomes. These biases can be based on factors such as race, gender, ethnicity, religion, socio-economic status, or other personal characteristics that may influence the way an individual or group is evaluated. For example, in a job interview process, if the interviewer has a pre-existing bias towards individuals of a certain race or gender, they may subconsciously rank candidates from that group higher, even if they are less qualified than other candidates. Pre-existing bias can also be influenced by cultural or societal norms that have been internalized by individuals or groups. For instance, if a ranking system gives preference to candidates who have attended prestigious universities, this may reflect a pre-existing bias towards individuals from privileged backgrounds. It is important to identify and address pre-existing bias in ranking to ensure that the process is fair and equitable. This can be done by implementing an objective criteria, increasing transparency in the ranking process, and providing diversity and sensitivity training to those involved in the ranking process[13].

1.3.2 Technical Bias

Technical bias in ranking refers to biases or inaccuracies that arise from flaws in the design or implementation of the ranking system itself. These biases can be unintentional and may result from limitations in the data, algorithms, or methodology used in the ranking process. Statistical bias is one of the most common technical biases. It can arise if the sample size or distribution of data used in the ranking process is not representative of the population being ranked. This can lead to inaccurate and biased results. It is important to identify and address technical bias in ranking to ensure that the process is accurate and fair. This can be done by improving the data quality, addressing algorithmic biases, using more objective criteria, and implementing

statistical checks to identify and correct biases[9]

1.3.3 Emergent Bias

Emergent bias in AI systems refers to biases that arise during the use and interaction with these systems, even if they were not present or intended during the design and initial training phases. This form of bias can emerge due to evolving data inputs, changing societal norms, or the way users interact with the technology. One critical aspect of emergent bias is its dynamic nature. Unlike pre-existing biases that stem from historical data or biased development practices, emergent bias evolves over time. It can be seen in recommendation systems, search engines, and social media algorithms, where user interactions and feedback loops can inadvertently reinforce and amplify certain biases. For instance, a recommendation algorithm might start promoting more polarized content if it leads to increased user engagement, leading to an unintended bias towards extreme viewpoints. Another factor contributing to emergent bias is the shift in societal values and norms. What is considered fair and unbiased at one point in time may change, making previously acceptable algorithms biased by new standards. This temporal dimension of bias necessitates continuous monitoring and updating of AI systems to align with current ethical standards. Addressing emergent bias requires a multi-faceted approach. It involves not only technical solutions, such as algorithmic audits and updates, but also a broader engagement with stakeholders, including users, and policymakers. The goal is to create a feedback mechanism where AI systems can be adjusted in response to identified biases and societal changes. [1]

1.3.4 Bias Mitigation

Debiasing methods can be categorized based on their placement in the pipeline as follows: *(i)Pre-processing* methods modify data or its labels before model training, aiming to enhance data features for better training outcomes. *(ii) In-processing* techniques focus on training models with an emphasis on fairness, without altering the original data; the optimization process balances accuracy with fairness considerations.

While many in-processing strategies resemble pre-processing ones, the distinction lies in the alteration of the training process rather than the input data. On the other hand, *(iii) post-processing* methods seek to improve the fairness of model outputs after training, without adjustments to the data or training procedure. These methods may involve modifying thresholds, scoring rules, or the order of the items to ensure fairness among different groups. Such approaches are particularly beneficial for adding fairness to models that were not initially trained with it in mind. It is notable that these methods can be used in parallel in a pipeline as well. They also ensure a certain level of representation for the disadvantaged group in the ranking and they are typically more straight-forward to implement in a pipeline that was not created based on fairness considerations as initial steps toward making it fairness-aware.

1.4 Notions of Fairness

The *fairness notion* in AI and machine learning refers to a set of principles and methodologies aimed at ensuring that AI systems do not perpetuate or amplify social inequalities. These principles are crucial for designing algorithms that make fair decisions, especially when they impact people's lives. The concept of fairness in AI is complex, as it involves balancing various ethical, legal, and technical considerations. Hence, there are multiple views in categorizing fairness notions and the most common classification is *group/individual* fairness.

1.4.1 Individual Fairness

Individual fairness is a principle which focuses on the ethical necessity to ensure that individuals who are comparably similar are treated equivalently in the decision-making processes. This concept relies fundamentally on the development of a similarity metric, which serves as a benchmark for comparing individuals and subsequently influences the fairness of decisions derived from these comparisons. In other words, *individual fairness* advocates for a consistency in outcomes for individuals who are similar along relevant dimensions. It operates on the principle that an AI system

should make similar predictions or decisions for individuals who are alike in terms of the features considered relevant by the system. The challenge lies in defining an appropriate and just metric of similarity that can be used to assess and ensure such consistency in decision-making. This definition highlights the need for a measurable and objective criterion to compare individuals and make decisions that are just and equitable at the individual level. However, defining what constitutes *similarity* is often subjective and complex, making the implementation of individual fairness in AI a challenging endeavor[2].

1.4.2 Group Fairness

Group fairness represents a crucial aspect of ethical algorithm design, focusing on ensuring equitable treatment and outcomes for different demographic groups. This notion is particularly pertinent in the context of groups defined by sensitive attributes such as race, gender, ethnicity, or socio-economic status. At the heart of group fairness is the pursuit of equality in decision-making processes deployed by AI systems. The goal is to design algorithms that do not inherently favor one group over another, thereby avoiding discrimination. This involves ensuring that the benefits and burdens of AI decisions are distributed evenly across all groups, without bias towards or against any particular demographic. To operationalize group fairness, various approaches have been developed. One key approach is demographic parity, which stipulates that decision rates should be similar across different groups. For instance, in a hiring algorithm, demographic parity would aim for a similar rate of job offers across all racial and gender groups. Another important approach is equality of opportunity, which emphasizes that individuals with similar qualifications should have similar chances of receiving favorable outcomes, independent of group membership. This is especially relevant in contexts like loan approvals or university admissions, where historical data might reflect societal biases[5].

Implementing group fairness is not without challenges. One significant issue is the potential conflict between group fairness and individual fairness, where ensuring equal treatment across groups might lead to unfair outcomes for individuals within those

groups. Additionally, different metrics of group fairness can sometimes be in conflict with each other, making it difficult to achieve all aspects of fairness simultaneously. The relevance and choice of a particular group fairness metric depend heavily on the specific context and the domain of application. In some cases, demographic parity might be the most appropriate goal, while in others, striving for equality of opportunity could be more beneficial.

Understanding and implementing group fairness is essential for mitigating systemic biases in AI systems and promoting social justice. It is particularly vital in sectors like employment, healthcare, and criminal justice, where AI-driven decisions can have profound impacts on people's lives[2].

1.4.2.1 Demographic Parity

Demographic parity is a common conception of non-discrimination. In general, it requires that a decision should not be influenced by any protected attribute[5]. As an instance, approval or denial of a job application by a company shouldn't be influenced by the gender of the applicant. So, in the simplest case where there exist only one protected group and the result of a decision is assumed to be binary, this can be formalized by:

$$P(\text{Approve}|A) = P(\text{Approve}|\bar{A}), \quad P(\text{Deny}|A) = P(\text{Deny}|\bar{A}) \tag{1}$$

where A and $\bar{A}$ refer to protected and non-protected groups respectively.

Demographic parity dictates that the proportion of positive outcomes, such as team selections or job offers, should remain consistent across different demographic groups of experts, with no regard to their popularity, gender, ethnicity, or any other protected characteristic[12]. It implies that if, for instance, 30% of a team or job offer recipients stem from one demographic group, similar percentages should be observable across all other demographic categories to achieve true demographic parity. The underpinning philosophy posits that each individual, irrespective of their demographic attributes, should have an equal shot at positive outcomes generated by automated

decision-making systems.

In a practical context, implementing demographic parity may involve formulating and adhering to policies that meticulously monitor and adjust the outputs of decision-making systems to reflect equal representation of different groups. This may entail a recalibration of algorithms or the incorporation of mechanisms that counterbalance disparities, ensuring that the decision-making process remains uninfluenced by biases related to protected characteristics. Nevertheless, it is crucial to note that while demographic parity promotes a degree of fairness by endorsing equal outcomes across groups, it may not always encapsulate the intricacies of individual merit or qualification. Thus, while it is important in fostering representational equality, supplementary measures may also be needed to ensure that fairness is concurrently upheld on an individual level.

1.4.2.2 Equality of Opportunity

Equality of opportunity is a stronger notion of non-discrimination. In the simplest case where there exists only one protected attribute and the result of a decision is assumed to be binary, a predictor ($\hat{y}$) is said to satisfy equal opportunity with respect to a protected attribute and true outcome y, if the predictor and the protected attribute are independent conditional on the true outcome being Approve (favorable). This can be formalized by:

$$P(\hat{y}|A, y = \text{Approve}) = P(\hat{y}|\bar{A}, y = \text{Approve}) \tag{2}$$

Where A and $\bar{A}$ refer to protected and non-protected groups, $\hat{y}$ is the predicted outcome, and y is the desired outcome[5].

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[2] A. Castelnovo, R. Crupi, G. Greco, et al. "A clarification of the nuances in the fairness metrics landscape". In: *Scientific Reports* 12 (2022), p. 4209. DOI: 10.1038/s41598-022-07939-1.

[3] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through Awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. Cambridge, Massachusetts: Association for Computing Machinery, 2012, pp. 214–226. ISBN: 9781450311151. DOI: 10.1145/2090236.2090255. URL: https://doi.org/10.1145/2090236.2090255.

[4] European Commission's High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines. 2019.

[5] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 3323–3331. ISBN: 9781510838819.

[6] Konger and Chen. "A Change by Apple Is Tormenting Internet Companies, Especially Meta". In: *The New York Times* (2022). URL: https://www.nytimes.com/2022/02/03/technology/apple-privacy-changes-meta.html.

[7] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, 2018, p. 112. ISBN: 978-1-4798-3724-3.

[8] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016.

[9] Gourab K. Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. "Fair Ranking: A Critical Review, Challenges, and Future Directions". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 1929–1942. ISBN: 9781450393522. DOI:

10.1145/3531146.3533238. URL: https://doi.org/10.1145/3531146.3533238.

[10] Julia Powles and Helen Nissenbaum. "The Seductive Diversion of 'Solving' Bias in Artificial Intelligence". In: *Medium* (2018). URL: https://medium.com.

[11] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 2016.

[12] Sahil Verma and Julia Rubin. "Fairness Definitions Explained". In: *Proceedings of the International Workshop on Software Fairness*. FairWare '18. Gothenburg, Sweden: Association for Computing Machinery, 2018, pp. 1–7. ISBN: 9781450357463. DOI: 10.1145/3194770.3194776. URL: https://doi.org/10.1145/3194770.3194776.

[13] Meike Zehlike, Ke Yang, and Julia Stoyanovich. "Fairness in Ranking, Part I: Score-Based Ranking". In: *ACM Comput. Surv.* 55.6 (Dec. 2022). ISSN: 0360-0300. DOI: 10.1145/3533379. URL: https://doi.org/10.1145/3533379.

CHAPTER 2

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

Hamed Loghmani, Hossein Fani

In the 4th International Workshop on Algorithmic Bias in Search and Recommendation, Colocated with the 45th European Conference on Information Retrieval (ECIR 2023), (BIAS'23)

2.1 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[13]. Successful teams have firsthand effects on creating organizational performance in the industry[3, 1, 15], academia [33, 23, 11], law [28, 14], and the healthcare sector [4, 27]. 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 [30].

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)[26], 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 [20, 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 [25, 5, 24]. We can observe the commercial application of machine learning-based algorithmic search for an optimum team in online platforms like LinkedIn[1] 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 [8, 10], disproportionate selection of popular candidates [32, 34, 29], and over/under-representation and racial/gender disparities [19] since they are trained on real-world datasets that already inherit hidden societal biases. On the other hand, social science research

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

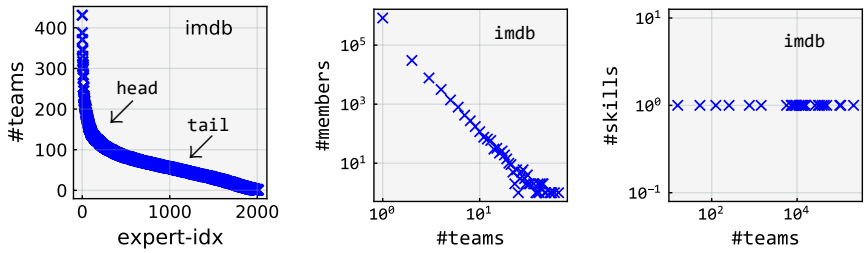


Fig. 2.1.1: Left: Long-tail distribution of casts and crews (experts) in movies (teams). Middle: Long-tail distribution in `log` scale. The figure reads y number of members have x number of teams. Right: uniform distribution of movies over genres (skills)

provides compelling evidence about the synergistic effects of diversity on team performance [31, 21, 12]; 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. [2] 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 [24, 5]. Rad et al. [24] employed uncertainty in learnable parameters by variational Bayesian neural model, and Dashti et al. [5] applied *virtually* negative samples from popular experts during the neural model learning procedure. However, they overlook substantiating the attribution by evidence using fairness metrics.

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 [24, 5]. 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`[2] dataset of movies, given a movie as a team of casts and crews such as actors and directors [18, 16], from Fig. 2.1.1(left), we observe a long tail of many nonpopular experts, while few popular experts in the head that dominate. Fig. 2.1.1(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. [9] for quantifying and mitigating popularity bias in state-of-the-art neural team formation methods [5] in terms of normalized discounted cumulative KL-divergence (`ndkl`) for *re*ranking experts in the recommended teams to achieve fairness based on the *demographic parity* [7] 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 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. [9] 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.

[2] imdb.com/interfaces/.

2.2 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 *demographic parity* [7] 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, demographic parity 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 demographic parity 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.

2.3 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 [5] 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.

2.3.1 Setup

2.3.1.1 Dataset.

Our testbed includes `imdb`[18, 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. 2.1.1, 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 [6, 24]. 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 2.3.1 reports additional point-wise statistics on the dataset before and after preprocessing.

2.3.1.2 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 nonpopular otherwise. As seen in Table 2.3.1, this number is `62.45` and the popularity ratio (popular/nonpopular) is `0.426/0.574`.

2.3.1.3 Baselines.

Our neural team formation baselines include variational Bayesian neural network [24] with unigram negative sampling strategy in minibatches [5] (`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 [22], 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

Table 2.3.1: 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

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) [9].

2.3.1.4 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, $rel(i) = 1$ 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 [9] (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)[9]:

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

2.3.2 Results

In response to **RQ1**, i.e., whether state-of-the-art neural team formation models produce fair teams of experts, from Table 2.3.2, 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 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 *demographic parity*. 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 2.3.2, 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.

Table 2.3.2: 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[5, 24]					
		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[24, 5]					
		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

2.4 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] 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". In: *Mathematics* 8.10 (2020). ISSN: 2227-7390. DOI: 10.3390/math8101804. URL: https://www.mdpi.com/2227-7390/8/10/1804.

[2] Giorgio Barnabò, Adriano Fazzone, Stefano Leonardi, and Chris Schwiegelshohn. "Algorithms for Fair Team Formation in Online Labour Marketplaces". In: *Companion Proceedings of The 2019 World Wide Web Conference*. 2019, pp. 484–490.

[3] K.M. Bursic. "Strategies and benefits of the successful use of teams in manufacturing organizations". In: *IEEE Transactions on Engineering Management* 39.3 (1992), pp. 277–289. DOI: 10.1109/17.156562.

[4] Maxine Craig and Debi McKeown. "How to build effective teams in healthcare". In: *Nursing times* 111.14 (2015), pp. 16–18. ISSN: 0954-7762. URL: http://europepmc.org/abstract/MED/26182585.

[5] Arman Dashti, Saeed Samet, and Hossein Fani. "Effective Neural Team Formation via Negative Samples". In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*. Ed. by Mohammad Al Hasan and Li Xiong. ACM, 2022, pp. 3908–3912. DOI: 10.1145/3511808.3557590. URL: https://doi.org/10.1145/3511808.3557590.

[6] Arman Dashti, Saeed Samet, and Hossein Fani. "Effective Neural Team Formation via Negative Samples". In: *Proceedings of the 31st ACM International Conference on Information amp; Knowledge Management*. CIKM '22. Atlanta, GA, USA: Association for Computing Machinery, 2022, pp. 3908–3912. ISBN: 9781450392365.

[7] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through Awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. Cambridge, Massachusetts: Association for Computing Machinery, 2012, pp. 214–226. ISBN: 9781450311151. DOI: 10.1145/2090236.2090255. URL: https://doi.org/10.1145/2090236.2090255.

[8] 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*. ACM, 2012, pp. 214–226. DOI: 10.1145/2090236.2090255. URL: https://doi.org/10.1145/2090236.2090255.

[9] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. "Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search". In: *KDD*. ACM, 2019, pp. 2221–2231.

[10] 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*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 2125–2126. ISBN: 9781450342322. DOI: 10.1145/2939672.2945386. URL: https://doi.org/10.1145/2939672.2945386.

[11] 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". In: *American Psychologist* 73 (May 2018), pp. 532–548. DOI: 10.1037/amp0000319.

[12] Bas Hofstra, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. "The diversity–innovation paradox in science". In: *Proceedings of the National Academy of Sciences* 117.17 (2020), pp. 9284–9291.

[13] Damon Horowitz and Sepandar D. Kamvar. "The Anatomy of a Large-Scale Social Search Engine". In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. Raleigh, North Carolina, USA: Association for Computing Machinery, 2010, pp. 431–440. ISBN: 9781605587998. DOI: 10.1145/1772690.1772735. URL: https://doi.org/10.1145/1772690.1772735.

[14] Jia Hu and Robert C. Liden. "Making a Difference in the Teamwork: Linking Team Prosocial Motivation to Team Processes and Effectiveness". In: *Academy of Management Journal* 58 (2014), pp. 1102–1127.

[15] Almagul Kairgalievna and Nurul Mohammad Zayed. "THE EFFECT OF TEAMWORK ON EMPLOYEE PRODUCTIVITY". In: 2021.

[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*. 2011, pp. 985–994.

[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*. 2012, pp. 1269–1272. DOI: 10.1109/ICDE.2012.124.

[18] Mehdi Kargar, Lukasz Golab, Divesh Srivastava, Jaroslaw Szlichta, and Morteza Zihayat. "Effective Keyword Search Over Weighted Graphs". In: *IEEE Trans. Knowl. Data Eng.* 34.2 (2022), pp. 601–616. DOI: 10.1109/TKDE.2020.2985376. URL: https://doi.org/10.1109/TKDE.2020.2985376.

[19] 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*. 2015, pp. 3819–3828.

[20] Theodoros Lappas, Kun Liu, and Evimaria Terzi. "Finding a team of experts in social networks". In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*. Ed. by John F. Elder IV, Françoise Fogelman-Soulié, Peter A.

Flach, and Mohammed Javeed Zaki. ACM, 2009, pp. 467–476. DOI: 10.1145/1557019.1557074. URL: https://doi.org/10.1145/1557019.1557074.

[21] Jakob Lauring and Florence Villesèche. "The performance of gender diverse teams: what is the relation between diversity attitudes and degree of diversity?" In: *European Management Review* 16.2 (2019), pp. 243–254.

[22] Quoc Le and Tomas Mikolov. "Distributed Representations of Sentences and Documents". In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML'14. Beijing, China: JMLR.org, 2014, II–1188–II–1196.

[23] Erin Leahey. "From Sole Investigator to Team Scientist: Trends in the Practice and Study of Research Collaboration". In: *Annual Review of Sociology* 42.1 (2016), pp. 81–100. DOI: 10.1146/annurev-soc-081715-074219. eprint: https://doi.org/10.1146/annurev-soc-081715-074219. URL: https://doi.org/10.1146/annurev-soc-081715-074219.

[24] Radin Hamidi Rad, Hossein Fani, Mehdi Kargar, Jaroslaw Szlichta, and Ebrahim Bagheri. "Learning to Form Skill-based Teams of Experts". In: *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. Ed. by Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux. ACM, 2020, pp. 2049–2052. DOI: 10.1145/3340531.3412140. URL: https://doi.org/10.1145/3340531.3412140.

[25] Radin Hamidi Rad, Aabid Mitha, Hossein Fani, Mehdi Kargar, Jaroslaw Szlichta, and Ebrahim Bagheri. "PyTFL: A Python-based Neural Team Formation Toolkit". In: *CIKM*. ACM, 2021, pp. 4716–4720. DOI: 10.1145/3459637.3481992. URL: https://doi.org/10.1145/3459637.3481992.

[26] Fahimeh Rahmanniyay, Andrew Junfang Yu, and Javad Seif. "A multi-objective multi-stage stochastic model for project team formation under uncertainty in time requirements". In: *Comput. Ind. Eng.* 132 (2019), pp. 153–165. DOI: 10.

1016/j.cie.2019.04.015. URL: https://doi.org/10.1016/j.cie.2019.04.015.

[27] 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". In: *American Psychologist* 73.4 (2018). Cited by: 297; All Open Access, Green Open Access, pp. 433–450. DOI: 10.1037/amp0000298.

[28] Peter D. Sherer. "Leveraging Human Assets in Law Firms: Human Capital Structures and Organizational Capabilities". In: *ILR Review* 48.4 (1995), pp. 671–691.

[29] Jianing Sun, Wei Guo, Dengcheng Zhang, Yingxue Zhang, Florence Regol, Yaochen Hu, Huifeng Guo, Ruiming Tang, Han Yuan, Xiuqiang He, et al. "A framework for recommending accurate and diverse items using bayesian graph convolutional neural networks". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 2030–2039.

[30] 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". In: *Psychological Science* 25.8 (2014), pp. 1581–1591.

[31] Cara Tannenbaum, Robert P Ellis, Friederike Eyssel, James Zou, and Londa Schiebinger. "Sex and gender analysis improves science and engineering". In: *Nature* 575.7781 (2019), pp. 137–146.

[32] Emre Yalcin and Alper Bilge. "Investigating and counteracting popularity bias in group recommendations". In: *Information Processing & Management* 58.5 (2021), p. 102608.

[33] Julie Younglove-Webb, Barbara Gray, Charles William Abdalla, and Amy Purvis Thurow. "The Dynamics of Multidisciplinary Research Teams in Academia". In: *The Review of Higher Education* 22.4 (1999), pp. 425–440. ISSN: 1090-7009.

[34] Ziwei Zhu, Yun He, Xing Zhao, and James Caverlee. "Popularity bias in dynamic recommendation". In: *SIGKDD*. 2021, pp. 2439–2449.

CHAPTER 3

OpeNTF2: A Framework for Fair Team Formation

Hamed Loghmani, Reza Barzegar, Gabriel Rueda, Edwin Paul, Hossein Fani

3.1 introduction

Team formation aims to automate the forming of collaborative teams of experts whose combined skills, applied in coordinated ways, successfully solve difficult tasks. It falls under social information retrieval (Social IR) research where the *right* group of experts are searched and hired [12, 11]. Forming teams is challenging for the large pool of candidate experts from diverse cultural backgrounds and personality traits, along with the unknown synergistic balance among them. To address the complex nature of the task, algorithmic approaches have been proposed, among which neural models build up the following in scalability and inference efficacy [18, 3, 19, 17, 21, 20, 16]. Neural team formation learns the relationships between experts and their skills within the context of teams through an iterative and online learning procedure on past instances of teams as training samples. To support neural team formation research with a reproducible and open-sourced platform, Rad et al. released pytfl [20], a python-based library that transforms a dataset of teams into a heterogeneous graph structure followed by the model training and inference steps. Pytfl, however, struggles with large-scale datasets and lacks modularity for ease of customization and extension to new methods and datasets from emerging domains. We previously open-sourced OpeNTF [5], a modularized and scalable benchmark framework, which includes two

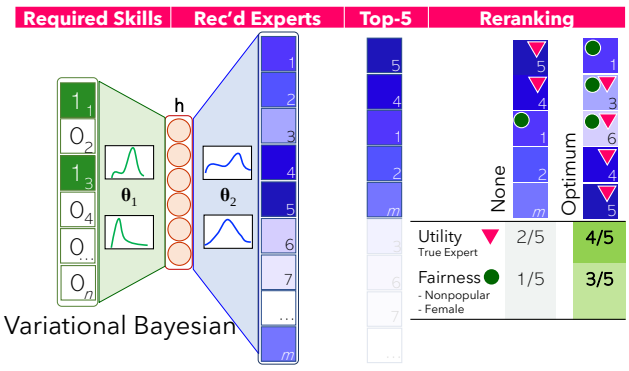


Fig. 3.1.1: `OpeNTF2` fairness-aware reranking flow.

reference neural architectures, including feedforward neural network and the state-of-the-art variational Bayesian neural network, both designed to accept sparse or dense vector representations of skills in the input layer. In OpeNTF, the definition of experts and skills within a team and the team's label of success can be readily modified in different domains through inheritance. In the absence of labels for unsuccessful teams, OpeNTF can follow the closed-world assumption; it presumes existing instances of teams in the training dataset as successful and sample subsets of experts who have not collaborated yet as unsuccessful teams (virtually negative samples) [4, 16]. Further, OpeNTF hosts three large-scale training datasets from varying domains, including `dblp` where publications in computer science are the teams, authors are the experts, and fields of study are the skills, `imdb` where movies are the teams, cast and crew are the experts, and genres are the skills, and `uspt` where patents are the teams, inventors are the experts, and patents' subclasses are the experts. The use of `imdb` in the context of team recommendation research should *not* be mistaken for its applications in movie recommendation systems or movie sentiment analysis; here, the objective is to assemble a team of actors and staff for a movie project rather than a movie suggestion[14, 13]. In sum, OpeNTF can benchmark `16+1` baselines on `3` large-scale datasets.

However, the main focus of existing team formation models and libraries is maximizing the success rate (utility) by tailoring the recommended experts for a team to the required skills only, largely oblivious to the fairness of recommended experts, while it has been well-explored that machine learning models that produce recommendations suffer from unfair biases and result in discrimination and reduced visibil-

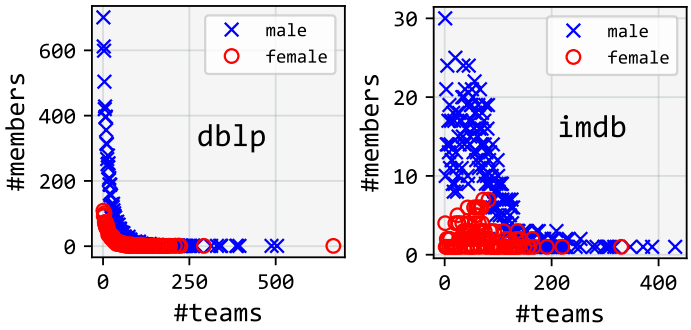


Fig. 3.1.2: Distribution of genders in `dblp` and `imdb`.

ity for an already disadvantaged group [6, 9], disproportionate selection of popular candidates [23, 26, 22], and over/under-representation and racial/gender disparities. Oddly, there is little to no fairness-aware algorithmic method that mitigates societal biases in team formation algorithms except that of the recent work by Barnab'o et al. [2] that proves fair team formation is NP-complete; therefore, computationally prohibitive for practical use.

In this paper, we bring forth `OpeNTF2` that extends the prior version in this major direction. To counter unfairness in the recommended list of experts for teams, in `OpeNTF2`, we adopted three deterministic debiasing reranking algorithms [8] as well as the state-of-the-art probabilistic reranking algorithm [25] based on two alternative notions of fairness, that is, equality of opportunity and demographic parity, which enables further post-processing reranking refinements to the list of recommended experts, as seen in a figure 3.1.1, to reassure the desired fair outcome in terms of popularity and gender. Additionally, `OpeNTF2` features modern transformer-based models and a new large-scale `github` dataset from open-source software repositories where software developers are the experts and programming languages are the skills. Contrary to existing datasets, `github` has a limited variety of skills (programming languages) which are, by and large, employed by many repositories (teams) and follow a more fair distribution, opening a new benchmark challenge for reproducibility and generalizability of team formation models in varying domains with distinct distributions.

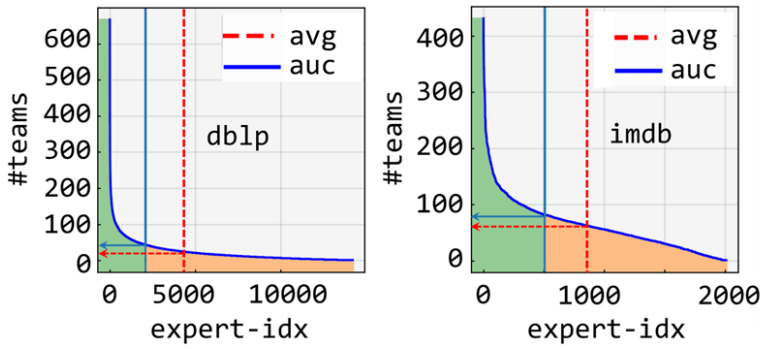


Fig. 3.1.3: Identifying popularity status in `dblp` and `imdb`.

3.2 Debiasing Algorithms

`OpeNTF2` integrated state-of-the-art debiasing reranking algorithms by Geyik et al. [8] including deterministic greedy (`greedy`), greedy conservative (`conservative`), and the relaxed variant of the greedy conservative (`relaxed`). The algorithms aim to meet minimum and maximum representation constraints for each protected attribute (e.g., gender or popularity) at every position in the ranked list. They work by greedy selection of the next best candidate among the attribute values that need more representation to meet the constraints. The algorithms differ in how conservatively they ensure the minimum representation constraints are met - some wait longer before selecting candidates from attribute values close to violating these constraints.

In addition, we adopt a probabilistic greedy reranking algorithm by Zehlike et al. [25], namely `fa*ir`, which uses a multinomial distribution, allowing for the handling of several protected groups simultaneously. They developed *'ranked group fairness'* criterion to ensure each segment of the ranking has a proportion of protected candidates that is statistically close to a set minimum threshold. Their method utilizes a *fairness tree* data structure for an efficient fair ranking verification and construction. Notably, `fa*ir` significantly improves representation for protected groups with minimal utility loss compared to utility-based rankings. This approach also offers flexibility in exploring fairness-utility trade-offs, providing more equitable visibility for multiple disadvantaged groups in top-k rankings.

Finally, we Incorporated fairness evaluation metrics such as ndkl [24] and skew [8]

```
class Reranking:
 def get_stats(teamsvecs, ...):
 def rerank(preds, labels, ratios,
    algorithm='greedy', ...):
 def eval_fairness(preds, labels, metrics, ...):
 def eval_utility(reranked_preds, press,
    metrics, ...):
```

Fig. 3.2.1: `Reranking` driver code.

into our pipeline. Fairness evaluation metrics measure the difference between the distribution of recommended teams over popularity or gender labels and a reference unbiased desired distribution. In tandem with utility metrics, which measure the efficacy of the recommended teams with respect to teams' success rate, such as precision or recall, `OpeNTF2` allows exploring the trade-offs between notions of fairness on the one hand, and success rate on the other hand for a team formation method. `OpeNTF2` supports two alternative notions of fairness, as explained hereafter.

3.2.1 Demographic Parity

This fairness notion requires the top-k recommended subset of experts to reflect the same distribution of popular/nonpopular or female/male as in the entire set of experts, *irrespective* of an expert having the required skills. For instance, the distribution of females and males in the entire dataset of `imdb` is `29.6%` and `70.4%`, respectively. A debiasing algorithm should preserve the same `1:2` ratio in the top-k ranked list of recommended experts between females and males. This notion of fairness forego qualifications and is known to have limitations [7, 10]. Notably, a random baseline that assigns experts to teams from a uniform distribution of experts regardless of popularity or gender labels is an *ideally* fair model yet at the cost of very low success rates for the recommended teams.

3.2.2 Equality of Opportunity

This notion of fairness requires *qualified* experts to have an equal chance of being selected for the team regardless of their popularity or gender status. For example, in `imdb`, given '*sport*' as the required genre (skill), `18.6%` of our qualified experts, i.e., acted in at least *one 'sport'* movie, are female and the rest are male. A debiasing algorithm should preserve the matching ratio of `1:4` between females and males in the top-k recommended list of experts. It is required to satisfy demographic parity among cast and crew who have participated in at least one movie in this genre. Basically, `OpeNTF2`'s Adila dampen a bias by adjusting the distributions of popular/male and nonpopluar/female experts in the top-k recommended experts for a team according to their ratio in the training dataset (demographic parity) or based on their ratio among qualified experts (equality of opportunity) 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.

3.2.3 Protected Attributes

`OpeNTF2` rely on experts' labels of popularity or gender to measure the biases and run the debiasing algorithms. 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 [27]. As shown in Figure 3.1.3, `OpeNTF2` has adopted two alternatives: *i*) `avg` where 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. As seen in Figure 3.1.3 for `imdb`, a random expert has participated in `62.45` movies. So, experts who participated in more movies are considered popular. *ii*) `auc` where 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.

Contrary to popularity, gender is accorded personally (self-identified). While `uspt` dataset includes gender labels, other training datasets lack gender labels in part or

```
# ./src/mdl/nmt.py
from mdl.ntf import Ntf
class Nmt(Ntf):
 def __init__(self):
  super(Ntf, self).__init()
 def learn(self, splits, path):
  cli_cmd = 'onmt_train '
  cli_cmd += f'-config
      {path}/fold{foldidx}/config.yml'
  subprocess.Popen(shlex.split(cli_cmd)).wait()
 def run(self, splits, vecs, cmd, ...):
  .... #loading model configs
  if 'train' in cmd:
   input_data,output_data=self.prepare_data(vecs)
   model_path = self.build_vocab(input_data,
    output_data, ...)
   self.learn(splits, model_path)
```

Fig. 3.2.2: `Ntf` class definition.

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 [1], based on the first name of the experts for `dblp` as well as those that are missing in `imdb`. Their approach is quite simple but effective. They gathered occurrences of names from all around the world on the web, and then for each name they returned a probability in addition to a gander and count value. It means the probability of the given gender is based on the count value which is the number of presence of that name in their database. The gender distributions for the `imdb` and `dblp` datasets are illustrated in Figure 3.3.2. It is evident from the data that both datasets exhibit a significant gender bias. Specifically, in the `imdb` dataset, the male-to-female ratio is `0.868` to `0.132`. Similarly, the `dblp` dataset shows a male-to-female ratio of `0.877` to `0.123`. This disproportionate representation underscores the importance of addressing and rectifying such biases.

Table 3.2.1: Results of debiasing algorithms for popularity on `imdb`.

		%ndkl	%ndkl	%map10	%ncdg10
		before↓	after↓	Δ↑	Δ↑
demographic parity					
bnn	det-cons	67.53	16.59	-0.36	-0.82
	det-greedy		16.60	-0.36	-0.82
	det-relaxed		16.35	-0.36	-0.82
	fa*ir		17.27	0.00	00.00
bnn-emb	det-cons	74.67	15.71	-0.48	-1.03
	det-greedy		15.72	-0.48	-1.03
	det-relaxed		15.43	-0.48	-1.03
	fa*ir		17.53	0.00	0.00
equality of opportunity					
bnn	det-cons	61.74	19.85	-0.35	-0.81
	det-greedy		20.11	-0.35	-0.81
	det-relaxed		19.70	-0.35	-0.81
	fa*ir		16.61	0.00	0.00
bnn-emb	det-cons	68.57	18.94	-0.48	-1.03
	det-greedy		19.21	-0.48	-1.03
	det-relaxed		18.77	-0.48	-1.03
	fa*ir		16.88	0.00	0.00

3.2.4 Benchmark Results

`OpeNTF2` has been benchmarked for debiasing algorithms on `imdb` and `dblp` to mitigate popularity and gender biases based on the two notions of fairness. As seen

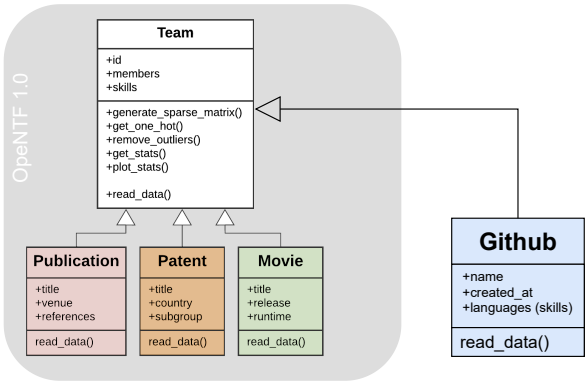


Fig. 3.3.1: OpeNTF2 dataset class inheritance

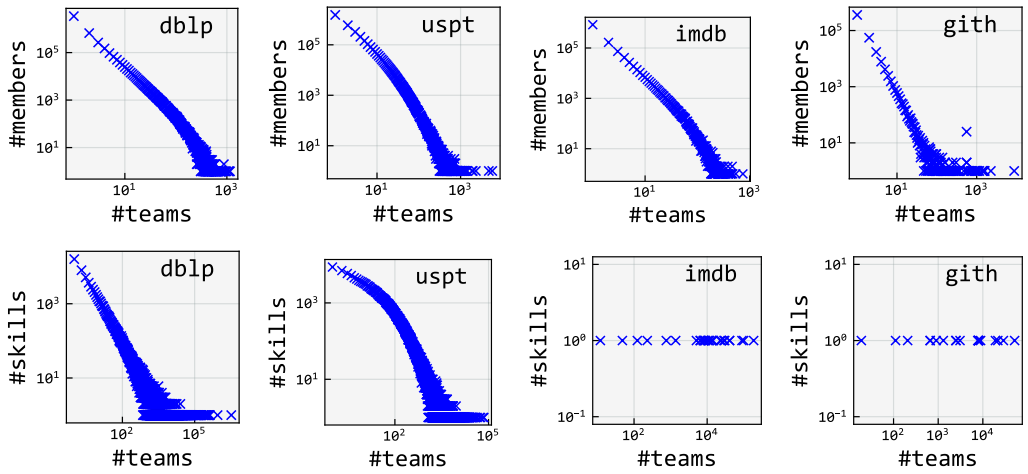
in Table 3.2.1, `fa*ir` were able to mitigate popularity bias significantly in terms of `ndkl` while avoiding utility loss based on the `map@10` and `ndcg@10`, before and after mitigation process. The full results, are accessible at `OpeNTF2`'s codebase.

3.3 Additional Features

We further extend `OpeNTF` with transformer-based neural models and a large-scale dataset from a new domain.

3.3.1 Transformer-based Models

The team formation problem can be viewed as a special case of sequence-to-sequence task from a subset of skills as the source sequence to a subset of experts as the target sequence with a compromise in the order of elements. `OpeNTF2` has integrated opennmt-py [15] through the wrapper class `Nmt`, to utilize modern transformers and encoder-decoder models with multiple rnn cells of type lstm or gru and different attention mechanisms. `Nmt` prepares the required source and target element sets and calls opennmt-py's executables by spawning a new process via python's `subprocess`. Furthermore, since `Nmt` wrapper inherits from `Ntf`, such models can utilize temporal training strategy through `tNtf`.



	dblp	uspt	imdb	github
#teams	4,877,383	7,068,508	507,034	132,851
#experts	5,022,955	3,508,807	876,981	452,606
#skills	89,504	241,961	28	20
%popular experts (avg)	18.00%	20.20%	5.00%	18.80%
%female experts	24.20%	13.50%	29.60%	-

Fig. 3.3.2: Distribution of teams over experts (top) and skills (bottom) along with stats in training datasets.

3.3.2 Dataset

Previously, OpeNTF hosted dblp dataset of computer science research papers with authors as experts and field of study (fos) as skills, uspt dataset of US patents with inventors as experts and patent subclasses as skills, and imdb dataset of movies with cast and crew as experts and genres as skills. While dblp and uspt have similar distributions of experts and skills in teams, imdb has been the only dataset with different distribution of skills in teams due to the limited number of skills (movie genres), as seen in Figure 3.3.2. In OpeNTF2, we also contribute github dataset of open-source software as teams of software developers crawled from github repositories, including contributors as the experts and programming languages as the skills as well as complementary information such as stargazer count, number of forks, and creation date. We used google cloud bigquery to obtain the list of repositories. We examine GitHub MemberEvent data from 2020 (starting on Jan 1) until it 750k unique repository names. MemberEvent are recorded every time an individual becomes a

```
python -u main.py
       -data ../data/raw/dblp/dblp.v12.json \
       -domain dblp \
       -model fnn bnn fnn_emb bnn_emb \
       -fairness greedy conservative relaxed fa-ir \
       -attribute gender popularity
```

Fig. 3.3.3: OpeNTF2's quickstart command.

contributor to a repository. The crawling process took 6 months given the 1,500 api call per hour limit. Github dataset follows similar distributions as in imdb, having two datasets in two alternative categories of varying distributions, i.e., dblp and uspt vs. imdb and github. Figure 3.3.2 demonstrates the distributions of experts and skills in teams as well as some stats for all datasets including the github dataset. Figure 3.3.3 shows OpeNTF2's quickstart command for fair and temporal team formation on github. Also, Figure 3.3.1 shows github's class definition.

3.4 Conclusion and Future Work

This paper presented OpeNTF2 with key extensions to its initial release as the first open-source python-based benchmark library for neural team formation research. OpeNTF2 features *i*) fairness-aware reranking algorithms to mitigate popularity and gender disparities in the training datasets of the team formation problem, *ii*) transformers and encoder-decoder models with rnn cells and attention mechanisms, and *iv*) a new large-scale dataset of open-source software repositories. OpeNTF's future focus will mainly be on fair team formation. We aim to include other notions of fairness such as equalized odds, and the respective debiasing methods and evaluation metrics. Finally, we plan to detect and mitigate potential racial bias in training datasets.

References

[1] https://genderize.io/. [Online; accessed 16-June-2023].

[2] Giorgio Barnabò, Adriano Fazzone, Stefano Leonardi, and Chris Schwiegelshohn. "Algorithms for Fair Team Formation in Online Labour Marketplaces". In: *Companion Proceedings of The 2019 World Wide Web Conference*. 2019, pp. 484–490.

[3] Arman Dashti, Saeed Samet, and Hossein Fani. "Effective Neural Team Formation via Negative Samples". In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*. Ed. by Mohammad Al Hasan and Li Xiong. ACM, 2022, pp. 3908–3912. DOI: 10.1145/3511808.3557590. URL: https://doi.org/10.1145/3511808.3557590.

[4] Arman Dashti, Saeed Samet, and Hossein Fani. "Effective Neural Team Formation via Negative Samples". In: *Proceedings of the 31st ACM International Conference on Information amp; Knowledge Management*. CIKM '22. Atlanta, GA, USA: Association for Computing Machinery, 2022, pp. 3908–3912. ISBN: 9781450392365.

[5] Arman Dashti, Karan Saxena, Dhwani Patel, and Hossein Fani. "OpeNTF: A Benchmark Library for Neural Team Formation". In: *CIKM*. ACM, 2022, pp. 3913–3917. DOI: 10.1145/3511808.3557526. URL: https://doi.org/10.1145/3511808.3557526.

[6] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through awareness". In: *ITCS*. 2012, pp. 214–226.

[7] 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*. ACM, 2012, pp. 214–226. DOI: 10.1145/2090236.2090255. URL: https://doi.org/10.1145/2090236.2090255.

[8] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. "Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search". In: *KDD*. ACM, 2019, pp. 2221–2231.

[9] 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*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 2125–2126. ISBN: 9781450342322. DOI: 10.1145/2939672.2945386. URL: https://doi.org/10.1145/2939672.2945386.

[10] Moritz Hardt, Eric Price, and Nati Srebro. "Equality of Opportunity in Supervised Learning". In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. 2016, pp. 3315–3323.

[11] Damon Horowitz and Sepandar D. Kamvar. "Searching the village: models and methods for social search". In: *Commun. ACM* 55.4 (2012), pp. 111–118. DOI: 10.1145/2133806.2133830. URL: https://doi.org/10.1145/2133806.2133830.

[12] Damon Horowitz and Sepandar D. Kamvar. "The anatomy of a large-scale social search engine". In: *WWW*. ACM, 2010, pp. 431–440. DOI: 10.1145/1772690.1772735. URL: https://doi.org/10.1145/1772690.1772735.

[13] 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*. 2011, pp. 985–994.

[14] Mehdi Kargar, Lukasz Golab, Divesh Srivastava, Jaroslaw Szlichta, and Morteza Zihayat. "Effective Keyword Search Over Weighted Graphs". In: *IEEE Trans. Knowl. Data Eng.* 34.2 (2022), pp. 601–616. DOI: 10.1109/TKDE.2020.2985376. URL: https://doi.org/10.1109/TKDE.2020.2985376.

[15] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. "OpenNMT: Open-Source Toolkit for Neural Machine Translation". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, System Demonstrations*. Association for Computational Linguistics, 2017, pp. 67–72. DOI: 10.18653/v1/P17-4012. URL: https://doi.org/10.18653/v1/P17-4012.

[16] Radin Hamidi Rad, Ebrahim Bagheri, Mehdi Kargar, Divesh Srivastava, and Jaroslaw Szlichta. "Retrieving Skill-Based Teams from Collaboration Networks". In: *SIGIR*. ACM, 2021, pp. 2015–2019. DOI: 10.1145/3404835.3463105. URL: https://doi.org/10.1145/3404835.3463105.

[17] Radin Hamidi Rad, Ebrahim Bagheri, Mehdi Kargar, Divesh Srivastava, and Jaroslaw Szlichta. "Subgraph Representation Learning for Team Mining". In: *WebSci*. ACM, 2022, pp. 148–153. DOI: 10.1145/3501247.3531578. URL: https://doi.org/10.1145/3501247.3531578.

[18] Radin Hamidi Rad, Hossein Fani, Ebrahim Bagheri, Mehdi Kargar, Divesh Srivastava, and Jaroslaw Szlichta. "A Variational Neural Architecture for Skill-Based Team Formation". In: *ACM Trans. Inf. Syst.* (Apr. 2023). Just Accepted. ISSN: 1046-8188. DOI: 10.1145/3589762. URL: https://doi.org/10.1145/3589762.

[19] Radin Hamidi Rad, Hossein Fani, Mehdi Kargar, Jaroslaw Szlichta, and Ebrahim Bagheri. "Learning to Form Skill-based Teams of Experts". In: *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. Ed. by Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux. ACM, 2020, pp. 2049–2052. DOI: 10.1145/3340531.3412140. URL: https://doi.org/10.1145/3340531.3412140.

[20] Radin Hamidi Rad, Aabid Mitha, Hossein Fani, Mehdi Kargar, Jaroslaw Szlichta, and Ebrahim Bagheri. "PyTFL: A Python-based Neural Team Formation Toolkit".

In: *CIKM*. ACM, 2021, pp. 4716–4720. DOI: 10.1145/3459637.3481992. URL: https://doi.org/10.1145/3459637.3481992.

[21] Radin Hamidi Rad, Shirin Seyedsalehi, Mehdi Kargar, Morteza Zihayat, and Ebrahim Bagheri. "A Neural Approach to Forming Coherent Teams in Collaboration Networks". In: *EDBT*. OpenProceedings.org, 2022, 2:440–2:444. DOI: 10.48786/edbt.2022.37. URL: https://doi.org/10.48786/edbt.2022.37.

[22] Jianing Sun, Wei Guo, Dengcheng Zhang, Yingxue Zhang, Florence Regol, Yaochen Hu, Huifeng Guo, Ruiming Tang, Han Yuan, Xiuqiang He, et al. "A framework for recommending accurate and diverse items using bayesian graph convolutional neural networks". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 2030–2039.

[23] Emre Yalcin and Alper Bilge. "Investigating and counteracting popularity bias in group recommendations". In: *Information Processing Management* 58.5 (2021), p. 102608. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2021.102608. URL: https://www.sciencedirect.com/science/article/pii/S0306457321001047.

[24] Ke Yang and Julia Stoyanovich. "Measuring Fairness in Ranked Outputs". In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. SSDBM '17. Chicago, IL, USA: Association for Computing Machinery, 2017. ISBN: 9781450352826.

[25] Meike Zehlike, Tom Sühr, Ricardo Baeza-Yates, Francesco Bonchi, Carlos Castillo, and Sara Hajian. "Fair Top-k Ranking with Multiple Protected Groups". In: *Inf. Process. Manage.* 59.1 (Jan. 2022). ISSN: 0306-4573. DOI: 10.1016/j.ipm.2021.102707. URL: https://doi.org/10.1016/j.ipm.2021.102707.

[26] Ziwei Zhu, Yun He, Xing Zhao, and James Caverlee. "Popularity bias in dynamic recommendation". In: *SIGKDD*. 2021, pp. 2439–2449.

[27] Ezra W. Zuckerman and John T. Jost. “What Makes You Think You’re so Popular? Self-Evaluation Maintenance and the Subjective Side of the ”Friendship Paradox””. In: *Social Psychology Quarterly* 64.3 (2001), pp. 207–223. ISSN: 01902725. (Visited on 06/14/2023).

CHAPTER 4

A Probabilistic Greedy Attempt to be Fair in Neural Team Recommendation

Hamed Loghmani, Mahdis Saeedi, Gabriel Rueda, Edwin Paul, Hossein Fani

4.1 Introduction

As modern tasks have surpassed the capacity of individuals, recommending collaborative teams of experts, whose combined skills applied in coordinated ways toward a common goal could yield success, has been a surge of research interest in many disciplines, including psychology [17, 42, 36, 26], the science of team science (scits) [46, 57], and industrial engineering [13, 12]. Team recommendation can be seen as social information retrieval (Social IR) where the right group of experts, rather than the right documents, are searched and hired to solve the task at hand.

To replace the tedious, error-prone, and suboptimal manual team formation by a human selector, who falls short for an overwhelming number of experts, and fails to consider a multitude of criteria to optimize simultaneously[13], a rich body of various computational methods, from operations research[2, 13, 58, 63, 65, 8, 18, 32, 62], social network analysis[38, 34, 23, 56], and recently, machine learning[50, 54, 14, 49, 47, 16, 51, 48] have been proposed. Specifically, neural models have been proposed to learn the distributions of experts and their skill sets in the context of successful

and unsuccessful teams in training datasets to recommend future successful teams. Such models have brought state-of-the-art efficacy while enhancing efficiency, taking the stage and becoming canonical in team recommendation literature[50, 47, 44, 16, 14].

Most critically, manual team formation also suffers from human selector's hidden personal and societal biases [52], whose mitigation has received little to no attention in the literature. Indeed, the primary focus of existing team recommenders 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*. It has been well-explored that machine learning-based recommenders produce unfair biases in their ranked list of recommendations, leading to discrimination and reduced visibility for an already disadvantaged group [27], disproportionate selection of popular experts [40], and over/under-representation and racial/gender disparities. These biases, far from being random, originate mainly from training datasets. For instance, training sets in team recommendation suffer from popularity bias; that is, the majority of *nonpopular* experts have scarcely participated in the (successful) teams, whereas few popular experts are in many teams [28, 15]. Therefore, popular experts would receive more attention and are more frequently recommended by a machine learning model, leading to systematic discrimination against already disadvantaged nonpopular experts.

In this paper, we propose incorporating the notions of fairness in tandem with expertise to facilitate recommending merit-based teams while equal opportunity and fairness is also maximized. Specifically, we leverage a probabilistic fairness-aware reranking method [68] to adjust the ordering of experts in the final ranked list of recommendations 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 belongs to *post-processing* category of methods, which seek to improve the fairness of model's outputs after training, without adjustments to the data or training procedure. 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 the best of our knowledge, there is no fairness-aware approach in the neural team recommendation method except that of Loghmani et al. [40], who applied deterministic greedy reranking algorithms to mitigate popularity bias in neural team recommendation models. They showed such deterministic methods can mitigate bias but at the cost of substantial drop in team's likelihood for success. Our experiments on two large-scale datasets demonstrate that:
(1) With respect to popularity bias, our approach substantially mitigates bias while maintaining the success rate of the recommended teams.
(2) With respect to gender, however, our approach's impact on fairness has been marginal due to the highly skewed distribution of male vs. female experts in the training dataset.

4.2 Related Work

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

4.2.1 Neural Team Formation

Among the proposed team recommendation methods, we focus on the neural models as the cutting-edge computational methods which offer efficiency and effectiveness due to the inherently iterative and online learning procedure. This line of research starts with Sapienza et al. [54] who employed an autoencoder neural network architecture for the team recommendation problem in online games. Subsequently, other researchers have continued this line, exploring alternative neural-based architectures. Rad et al. [28] used a variational Bayesian neural model that incorporated uncertainty via probabilistic weights to address overfitting in highly skewed training datasets. Rad et al. [47] also employed graph neural network to learn dense vector

representations of skills followed by the variational Bayesian model, obtaining performance improvements. Dashti et al. [15] proposed negative sampling heuristics to improve the efficiency of neural model training even more via *virtually* unsuccessful teams. Successful as they are, neural team recommenders overlook fairness, contrary to rich body of fairness-aware methods in other disciplines such as in healthcare [11, 25, 1, 39], information retrieval [6, 19, 45], image processing and classification [31, 33, 41, 37, 64] and finally, ranking and recommendations [68, 66, 24, 55, 53, 67, 43, 4].

4.2.2 Fairness-aware Recommendation

Amongst fairness-aware recommendation methods in machine learning, post-processing methods offer the benefits of implementing fairness without modification to the model architecture or negative impact on the predictive power of a model. Further, they are model agnostic, enabling their application across various models [22]. They permit adjustments and tuning of fairness criteria based on ongoing evaluations and different fairness definitions without model retraining [10, 20]. The ease of implementation comes to the scene as well, with post-processing methods often presenting a simpler, more straightforward path to implementation than in-processing methods. Such advantages have encouraged post-processing methods for the fair team recommendation problem [69].

Post-processing reranking algorithms ensure a fair-distributed representations of diverse groups across all ranking positions[9]. Within the literature, diverse methods, including integer programming problems [5, 55, 60] and algorithmic approaches [24, 68, 66] have been applied to address this challenge. Integer programming-based methods, however, consist of a large number of variables and constraints, hence, practically prohibitive. Some other approaches proposed algorithms to create a fair rerank of an original ranked list. For instance, Geyik et al. [24] attain a specified distribution of protected attributes like gender and age among the top-ranked items in the final recommended list. Through extensive simulations, these algorithms are evaluated against varying attribute values and desired distributions, exploring the

trade-off between fairness and ranking utility. They utilize various measures such as `skew` and normalized discounted cumulative KL-divergence (`ndkl`) to quantify bias in rankings, also for evaluation. Contrary to the deterministic approaches, Zehlike et al. [68] took a probabilistic approach and presented an algorithm to produce a top-k ranking while maintaining fairness towards *multiple* protected groups.

4.3 Fair Neural Team Recommendation

In this section, we introduce the necessary notations and definitions for neural team formation, on the one hand, and fairness on the other hand. Then, we provide a general problem statement for recommending a fair team.

4.3.1 Neural Team Recommendation

Definition 1 (Team)**.** *Given a set of skills $\mathcal{S}$ and a set of experts $\mathcal{E}$, a team of experts $E \subseteq \mathcal{E}; E \neq \emptyset$ that collectively cover a skill set $S \subseteq \mathcal{S}; S \neq \emptyset$ shown by (S, E) along with its success status y where $y \in \{0, 1\}$. Further, $\mathcal{T} = \{(S, E)_y : y \in \{0, 1\}\}$ indexes all previous teams, successful and unsuccessful.*

Definition 2 (Team Recommendation)**.** *For a given subset of desired skills, the goal of the team recommendation problem is to recommend an optimal subset of experts E^* that their collaboration as a team leads to success, i.e., $(S, E^*)_{y=1}$, and ignore potentially unsuccessful subset of experts E', i.e., $(S, E')_{y=0}$. 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 $f_\theta : \mathcal{P}(\mathcal{S}) \rightarrow \mathcal{P}(\mathcal{E}), f_\theta(S) = E^*$.*

The output of a neural team recommender is a list of *all* experts where each expert $e \in \mathcal{E}$ is assigned a score, which is the probability of its membership in the final recommended team and can be ordered decreasingly to form a ranking. The recommended team is a subset of experts $E \in \mathcal{E}$ with the top-k highest probability scores.

Definition 3 (Neural Team Recommendation)**.** *Given a subset of skills S and all previous teams $\mathcal{T}$ as the training set, neural team recommendation estimates $f_\theta(\boldsymbol{s})$ 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 θ in f_θ over $\mathcal{T}$, that is, $argmax_\theta p(\theta|\mathcal{T})$.*

4.3.2 Notions of Fairness

To eschew varied interpretations and to provide actionable criteria to design and evaluate computational 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 a group level like females vs. males. In this paper, as we aim to provide fairness for groups of expert candidates when recommending a team, we focus on group-based notions of fairness, including demographic parity [7] and equality of opportunity [29].

Given a protected attribute $a_V = \{v_1, \cdots, v_n\}$, e.g., gender= $\{0 : male, 1 : female\}$, we divide experts into groups per attribute value, each referred to as protected group $g_{a:v}$, e.g., females and males, such that $\forall e, e' \in g_{a:v}$, $e_{a:v}$ and $e'_{a:v}$. Notions of group fairness are then defined as follows:

Definition 4 (Demographic Parity †)**.** *Demographic parity requires decisions for members of protected groups to be oblivious to the value of the protected attribute [29]. Formally,*

$$\forall d \in D, i \neq j \in a_V \qquad p(\hat{d}|e_{a:i}) = p(\hat{d}|e'_{a:j}) \tag{1}$$

where D is the set of decisions and $\hat{d}$ is the predicted decision for the correct decision d, a_V is the domain set of a protected attribute a, and e_i is an expert whose value of protected attribute a is i.

Assuming decisions are about the boolean membership status of experts in the recommended subset of experts E, i.e., $\{e \in E, e \neq E\}$ and protected attributes are

either gender= $\{0 : male, 1 : female\}$ or popularity= $\{0 : popular, 1 : nonpopular\}$, equation 1 becomes

$$\forall e_0, e_1' \in \mathcal{E} \quad [p(e_0 \in E) = p(e_1' \in E)] \wedge [p(e_0 \neq E) = p(e_1' \neq E)] \tag{2}$$

where $\mathcal{E}$ is the set of all candidates. 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 [61]. However, demographic parity overlooks experts' qualifications in protected groups; no condition for experts exists in equation 1 and equation 2.

Definition 5 (Equalized Odds)**.** *Equalized odds [29] is a stronger notion of fairness as it applies demographic parity on subsets of protected groups whose members are qualified enough to receive the true decision in a boolean decision set and protected attribute:*

$$\forall d \in D = \{0, 1\} \quad p(\hat{d}|e_0, d) = p(\hat{d}|e_1, d) \tag{3}$$

Assuming D is the set of decisions about team membership, we prioritize fairness for the decision $d = 1$, that is, recommending an expert to be in a team as an advantaged outcome, moving on from the discrimination against $d = 0$ (not recommended). This leads to a relaxed version of equalized odds, referred to as equality of opportunity, as follows.

Definition 6 (Equality of Opportunity ‡)**.** *Equality of opportunity [29] applies equalized odds only for the true positive decision in a boolean decision set and protected attribute:*

$$p(\hat{d}|e_0, d = 1) = p(\hat{d}|e_1, d = 1) \tag{4}$$

Adopting equality of opportunity for fair team recommendation problem, a qualified expert for a team with a required subset of skills S can be defined, oblivious to the expert's value of the protected attribute, as an expert who has experience in at least

one skill $s \in S$ in at least one team in the past. Hence,

$$p(e_0 \in E | e_0, e_0.S \cap S \neq \emptyset) = p(e_1 \in E | e_1, e_1.S \cap S \neq \emptyset) \tag{5}$$

where $e.S$ is the set of skills for an expert e.

4.3.2.1 Targeted Biases.

We target mitigating the following biases *separately*:

- *Popularity Bias:* The choice of popularity bias in this study is motivated due to the fact that team recommendation training sets exhibit the popularity bias; the majority of experts have limited involvement in successful teams, referred to as nonpopular experts, while a small number of experts, the popular ones, are part of many teams. In our work, we assign the protected attribute `popularity` to experts whose value is 0 if an expert's participation in teams exceeds the dataset's average and 1 otherwise.
- *Gender Bias:* Addressing gender bias in team recommendation models is a critical step toward promoting diversity and inclusion in the workplace. The literature has highlighted several reasons why gender bias is important in team recommendation including enhancing team creativity [30], improving team performance [3], and meeting legal and ethical obligations [21].

Definition 7 (Fair Team)**.** *Let E be a subset of experts. E's fairness is determined by the fair identifier function I as follows:*

$$I(E) = \begin{cases} 1 & E \text{ is a fair subset} \\ 0 & E \text{ is not a fair subset} \end{cases} \tag{6}$$

A subset of experts E in a team (S, E) is fair with respect to the notion of demographic parity *iff* equation 2 holds:

$$I(E) = 1^{\dagger} \iff \text{equation 2} \tag{7}$$

Alternatively, a subset of experts E in a team (S, E) is fair with respect to the notion of equality of opportunity *iff* equation 5 holds:

$$I(E) = 1^{\ddagger} \iff \text{equation 5} \tag{8}$$

4.4 Proposed Reranking Method

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

$$g(E) = E^* \text{ such that } I(g(f_\theta(S))) = I(g(E)) = I(E^*) \approx 1 \tag{9}$$

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. Let $E_{r,k}$ is the first k experts of E ranked by the probability values of the team recommender model for f_θ, denoted by ranking r, $|E_{r,k}|_p$ is the number of protected members in the $E_{r,k}$, $F(|E|_p; |E|, p)$ is the cumulative distribution function for a binomial distribution with parameters $|E|$ and p, $F^{-1}(\alpha; k, p)$ is the inverse function of F that computes the minimum number of required protected experts in $E_{\hat{r},k}$, $g_{a:1} = \{e_1^{(1)}, \cdots, e_1^{(l)}\}$ is the set of experts in the *positive* protected group (e.g., females). Then, the function $g(E)$ creates a new ranking r' based on the following:

$$\begin{cases} \hat{g}(e_i^{(k)}) = e_i^{(k)} & F^{-1}(\alpha; 1, p) \leq |E_{r,k}|_p \\ \hat{g}(e_i^{(k)}) = e_1^{(1)}, & F^{-1}(\alpha; 1, p) > |E_{r,k}|_p \\ \hat{g}(e_i^{(k+1)}) = e_1^{(2)}, & \\ .... & \\ \hat{g}(e^{(K+m)}) = e_1^{(m)}, & \end{cases} \tag{10}$$

where $m = F^{-1}(\alpha; 1, p) - |E_{r,k}|_p$

We adopt Zehlik et al. [66]'s $fa*ir$ algorithm as our reranking estimator for g for fair team recommendation, denoted $\hat{g}$, wherein the proportion of protected experts in every prefix of the top-k reranking r' remains statistically above or indistinguishable from a given minimum, i.e., $I(g(f_\theta(S))) \approx 1$. We evaluate this claim empirically on two large-scale datasets using two fairness metrics, namely `ndkl` and `skew`. Meanwhile, we evaluate the efficacy (utility) of teams by information retrieval metrics, including `map@k` and `ndcg@k`.

Our research unfolds in two pipelined steps:

1. initiating the training of a state-of-the-art neural team recommendation model to estimate f_θ, designed to generate expert recommendations for given skill subsets,

2. employing cutting-edge reranking algorithms to reorder the experts with an aim to elevate fairness, while trying to preserve accuracy in the recommendations.

From Definition 4.4 we conclude that the $F^{-1}(\alpha, |E|_p, p)$ is the minimum number of required protected experts in a team with size $|E|$ in order to have at least p percent of protected members in the subset of experts E. We calculate this value for each prefix of the team from $k = 1, \cdots, 100$. An instance of the aforementioned table is illustrated in Table 4.4.5.

4.5 Experiments

This section presents the details of research questions that we are going to answer in this paper.

`RQ1`, delves into the potential biases in the output of state-of-the-art neural team recommendation models. Our goal is to determine if these models, when recommending teams of experts, perpetuate biases, particularly concerning popularity and gender as protected attributes.

Table 4.4.1: Results for imdb with respect to popularity protected attribute.

			demographic parity					
		%ndkl	%ndkl	skew before→ 0		skew after→ 0		%map10	%ncdg10
		before↓	after↓	nonprotected	protected	nonprotected	protected	Δ↑	Δ↑
bnn	det-cons[24]	67.53	16.59	0.7472	-4.0195	-0.5449	0.2668	-0.36	-0.82
	det-greedy[24]		16.60			-0.5449	0.2668	-0.36	-0.82
	det-relaxed[24]		16.35			-0.5330	0.2629	-0.36	-0.82
	fa*ir[68]		17.27			0.2151	-0.1981	0.00	00.00
bnn-emb	det-cons[24]	74.67	15.71	0.7870	-4.3045	-0.4673	0.2374	-0.48	-1.03
	det-greedy[24]		15.72			-0.4673	0.2374	-0.48	-1.03
	det-relaxed[24]		15.43			-0.4560	0.2329	-0.48	-1.03
	fa*ir[68]		17.53			0.2167	-0.1985	0.00	0.00
		equality of opportunity							
bnn	det-cons[24]	61.74	19.85	0.6816	-3.9636	-0.6284	0.3288	-0.35	-0.81
	det-greedy[24]		20.11			-0.6289	0.329	-0.35	-0.81
	det-relaxed[24]		19.70			-0.6249	0.3277	-0.35	-0.81
	fa*ir[68]		16.61			0.198	-0.2052	0.00	0.00
bnn-emb	det-cons[24]	70.61	18.94	0.7283	-4.1799	-0.5553	0.3022	-0.48	-1.03
	det-greedy[24]		19.17			-0.5558	0.3024	-0.48	-1.03
	det-relaxed[24]		18.68			-0.5534	0.3014	-0.48	-1.03
	fa*ir[68]		18.15			0.2016	-0.2056	0.00	0.00

RQ2, examines if state-of-the-art probabilistic reranking algorithms are capable of enhancing the fairness of neural team recommendation models without compromising on their utility.

RQ3 investigates how effectiveness post-processing methods are in mitigating severe pre-existing biases within training datasets, and under what conditions these methods uphold the integrity and utility of the generated models across various application domains.

4.5.1 Datasets

Our experimental testbed consists of two large scale datasets: imdb and dblp. These datasets offer a comprehensive view of the domain and provide a robust foundation for our analyses. Our experimental dataset consists of data from imdb [35], where

Table 4.4.2: Results for `imdb` with respect to gender protected attribute.

		demographic parity							
		%ndkl	%ndkl	skew before→ 0		skew after→ 0		%map10	%ncdg10
		before↓	after↓	nonprotected	protected	nonprotected	protected	Δ↑	Δ↑
bnn	det-cons[24]	4.13	4.44	0.0009	-0.0494	0.0394	-0.3441	-0.36	-0.81
	det-greedy[24]		3.96			0.0419	-0.3693	-0.36	-0.81
	det-relaxed[24]		4.00			0.0404	-0.3559	-0.36	-0.81
	fa*ir[68]		4.09			0.0006	-0.0451	0.00	0.00
bnn-emb	det-cons[24]	4.92	8.34	0.014	-0.1338	0.0625	-0.6183	-1.17	-1.03
	det-greedy[24]		3.99			0.0385	-0.3311	-1.17	-1.03
	det-relaxed[24]		4.21			0.0371	-0.3185	-1.17	-1.03
	fa*ir[68]		4.88			0.0136	-0.1280	0.00	0.00
		equality of opportunity							
bnn	det-cons[24]	3.42	3.97	0.0051	-0.0626	0.0313	-0.274	-0.42	-0.99
	det-greedy[24]		3.67			0.0300	-0.2598	-0.42	-0.99
	det-relaxed[24]		3.71			0.0289	-0.2502	-0.42	-0.99
	fa*ir[68]		3.39			0.0044	-0.0522	0.00	0.00
bnn-emb	det-cons[24]	4.00	4.17	0.0141	-0.1367	0.0288	-0.2453	-0.54	-1.20
	det-greedy[24]		3.89			0.0262	-0.2219	-0.54	-1.20
	det-relaxed[24]		3.93			0.0250	-0.2125	-0.54	-1.20
	fa*ir[68]		3.96			0.0135	-0.1304	0.00	0.00

each movie entry contains information about the cast and crew (such as actors and director) as well as the movie's genres. We view each movie as a team, with the cast and crew being the team members and the movie genres representing the skills needed. The `dblp.v12` [59] dataset is derived from the dblp computer science bibliography, a comprehensive collection of bibliographic references of the computer science domain. Dblp offers open information on principal computer science journals and conference proceedings. Within the `dblp.v12` version, one can typically expect to find attributes such as authors, the paper's title, its publication year, the publishing venue, referenced papers, and often an abstract summarizing the content. In the team recommendation problem, each publication can be viewed as a team, where the authors are the experts and the fields of study represent the set of skills covered by that team.

Table 4.4.3: Results for dblp with respect to popularity protected attribute.

		demographic parity							
		%ndkl	%ndkl	skew before→ 0		skew after→ 0		%map10	%ncdg10
		before↓	after↓	nonprotected	protected	nonprotected	protected	Δ↑	Δ↑
bnn	det-cons[24]	109.56	14.64	1.1343	-19.9704	0.6484	-0.5462	-0.28	-0.58
	det-greedy[24]		14.64			0.6484	-0.5462	-0.28	-0.58
	det-relaxed[24]		18.31			0.6413	-0.5360	-0.28	-0.58
	fa*ir[68]		19.71			0.2639	-0.1524	0.00	00.00
bnn-emb	det-cons[24]	110.31	14.09	1.1415	-20.7584	0.6262	-0.5161	-0.28	-0.58
	det-greedy[24]		14.09			0.6262	-0.5161	-0.28	-0.58
	det-relaxed[24]		17.65			0.6189	-0.5063	-0.28	-0.58
	fa*ir[68]		19.61			0.2686	-0.1531	0.00	0.00
		equality of opportunity							
bnn	det-cons[24]	102.01	13.12	1.0560	-19.9253	0.5773	-0.5113	-0.28	-0.58
	det-greedy[24]		13.16			0.5773	-0.5113	-0.28	-0.58
	det-relaxed[24]		16.15			0.5729	-0.5050	-0.28	-0.58
	fa*ir[68]		18.96			0.2499	-0.1631	0.00	0.00
bnn-emb	det-cons[24]	102.85	12.65	1.0641	-20.6268	0.5555	-0.4813	-0.28	-0.58
	det-greedy[24]		12.67			0.5555	-0.4813	-0.28	-0.58
	det-relaxed[24]		15.63			0.5512	-0.4752	-0.28	-0.58
	fa*ir[68]		18.39			0.2526	-0.1645	0.00	0.00

4.5.1.1 Protected Attribute Distibutions

An expert is labeled as 'popular' if their team participation exceeds the dataset's average. For the imdb dataset, this average stands at 62.45 teams, while for the dblp dataset, it's 23.02 teams. In imdb, the proportion of popular to nonpopular experts is 0.426 to 0.574, while in dblp, it's 0.313 to 0.687. Regarding gender distribution, as illustrated in Fig.4.5.1 dblp has a male-to-female ratio of 0.858 to 0.142, and imdb has a slightly different ratio of 0.877 to 0.123.

4.5.2 Baselines

As our team recommendation baseline, we utilized a variation of OpeNTF[15]a Bayesian neural network (bnn) with variational inference [28], optimizing using the Kullback-Leibler (KL) divergence. We employed two distinct variations of this baseline model.

Table 4.4.4: Results for `dblp` with respect to gender protected attribute.

		demographic parity							
		%ndkl	%ndkl	skew before→ 0		skew after→ 0		%map10	%ncdg10
		before↓	after↓	nonprotected	protected	nonprotected	protected	Δ↑	Δ↑
bnn	det-cons[24]	11.80	4.92	-0.0895	0.4274	-0.0774	0.3945	-0.28	-0.58
	det-greedy[24]		3.72			-0.0774	0.3946	-0.28	-0.58
	det-relaxed[24]		6.52			-0.0078	-0.2066	-0.28	-0.58
	fa*ir[68]		8.39			-0.0014	-0.1162	0.00	0.00
bnn-emb	det-cons[24]	7.29	4.97	-0.0638	0.3084	-0.0734	0.3781	-0.28	-0.58
	det-greedy[24]		3.59			-0.0734	0.3784	-0.28	-0.58
	det-relaxed[24]		4.92			-0.0415	0.0871	-0.28	-0.58
	fa*ir[68]		6.83			-0.0368	0.1384	0.00	0.00
		equality of opportunity							
bnn	det-cons[24]	18.97	9.19	-0.1439	0.882	-0.1357	0.864	-0.28	-0.58
	det-greedy[24]		7.70			-0.1357	0.864	-0.28	-0.58
	det-relaxed[24]		9.53			-0.1343	0.8577	-0.28	-0.58
	fa*ir[68]		18.97			-0.1438	0.8818	0.00	0.00
bnn-emb	det-cons[24]	15.93	9.24	-0.1192	0.7646	-0.1348	0.8609	-0.28	-0.58
	det-greedy[24]		7.61			-0.1348	0.8609	-0.28	-0.58
	det-relaxed[24]		10.16			-0.1325	0.8504	-0.28	-0.58
	fa*ir[68]		15.93			-0.1191	0.7644	0.00	0.00

Table 4.4.5: A sample table of $F^{-1}(\alpha, |T|_p, p)$ values for $F^{-1}(0.1, 10, 0.6)$

position	1	2	3	4	5	6	7	8	9	10
min #of protected experts	0	0	1	1	2	2	3	3	4	4

Firstly, one that leverages pre-trained dense vector representations for the input skill subsets, denoted as (`-emb`), and secondly, a variation without any embedding. Both of these baseline models incorporate the use of negative sampling heuristics. In our efforts to establish fairness-aware reranking baselines, we employed three specific deterministic greedy reranking algorithms: `det-greedy`, `det-cons`, and `det-relaxed`, as detailed in [24]. Moreover, we integrated the adopted variant of `fa*ir` from [68], adapted especially for the purpose of team recommendation , serving as our primary reranking benchmark.

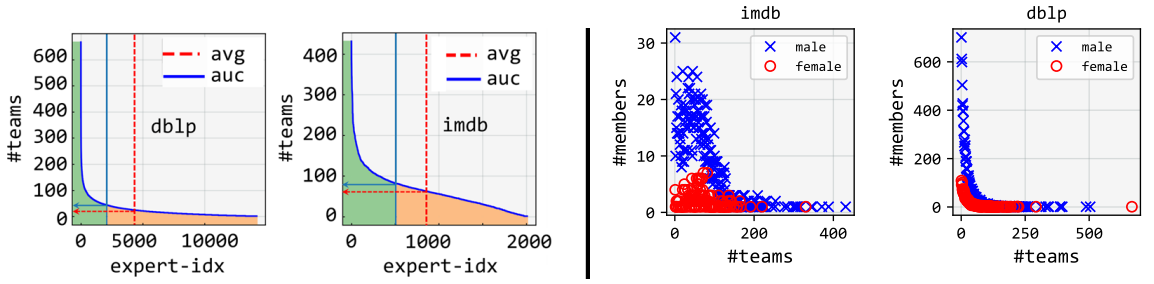


Fig. 4.5.1: Distribution of popular experts (left) and gender in `dblp` and `imdb` (right).

4.5.3 Evaluation Strategy and Metrics

To showcase the effectiveness of our approach, we partitioned our dataset by randomly selecting `15%` of the teams to constitute the test set. We then employed 5-fold cross-validation on the remaining `85%` of the teams, ensuring a rigorous training and validation process. This process yields one specialized model for each fold. For performance assessment, we adopted `map` (Mean Average Precision) and `ndcg` (Normalized Discounted Cumulative Gain) as our evaluation metrics. Furthermore, to ascertain the fairness of our team compositions, we turned to `ndkl`(Normalized Discounted Kullback-Leibler) and `skew` measures. Notably, we undertook a comparative fairness and utility evaluation, assessing the teams both prior to and following the reranking procedure.

4.5.4 The Effect of Skewness in Qualified Group

When the qualified set manifests skewness, our conventional expectations are vulnerable to bias while employing post-processing methods. Post-processing methods may fall short of adequately addressing stark skewness since they predominantly deal with the symptoms(biased output)rather than targeting the root(biased data or learning processes). When the bias escalates in severity, it becomes necessary to contemplate the utilization of pre-processing and in-processing methods in tandem with post-processing to achieve an optimal outcome. This integrative approach necessitates careful examination throughout the entire pipeline, guaranteeing that biases are sufficiently addressed and mitigated, thereby ensuring that outputs uphold the main tenets of fairness and equity.

4.5.5 Results

From Table 4.4.1 and 4.4.3, when using popularity as the protected attribute in `imdb` and `dblp`, we achieved significant improvements in fairness across fairness metrics without compromising utility metrics. On the other hand, from Table 4.4.2 and Table 4.4.4, where gender is the protected attribute and based on the employed fairness notions, the teams were already well-balanced. Consequently, we observed no significant enhancements in both fairness and utility. In response to `RQ1`, upon a comprehensive fairness evaluation conducted prior to reranking, it has been determined that the output from our team recommendation methods can exhibit bias towards certain protected attributes. As a result, specific protected groups may experience disadvantages. It is crucial to address these biases to ensure equitable outcomes for all participants, regardless of their background or attributes. From a comparative perspective, our results show that while fa*ir can maintain utility in terms of `map@10` after reranking, deterministic reranking algorithms can not as previously addressed by Loghmani et al. [40]. The same pattern is witnessed for `ndcg@10` as well in `imdb` and `dblp` with popularity or gender as the protected attribute based on demographic parity and equality of opportunity.

For `RQ2`, when considering popularity as the protected attribute, our findings confirm its influence. By adopting the `fa*ir` algorithm, we observed a marked improvement in fairness without compromising the overall utility. However, in the context of gender, the situation was more challenging. The initial bias in our dataset was so pronounced that, even though the algorithm identified teams as fair based on our chosen fairness criteria, in reality, they were not. This indicates the importance of continuously refining our fairness notions and the need for more comprehensive evaluations in cases of deeply rooted biases.

Finally, regarding `RQ3`, in our analysis, we determined that while post-processing methods can be notably effective in addressing biases, their efficacy diminishes when they are employed single-handedly. Specifically, when confronting *extreme* biases in data, these methods struggle to rectify them without a consequential loss in utility. It

is worth noting that the term extreme is context-sensitive; its interpretation can vary based on the nature of the problem, the expertise and number of involved specialists, as well as the size of the selected team. To achieve a more balanced and optimal outcome, a holistic approach that integrates pre-processing, in-processing, and post-processing methods is strongly recommended. The literature, such as Zehlike et al. [69], articulates that when bias reaches such extreme levels, post-processing methods alone tend to be insufficient in effectively mitigating it. Consequently, no tangible improvements in terms of gender bias were observed in these two datasets.

4.6 Concluding Remarks

In this paper, we explored the possibility of employing the adopted version of the `fa*ir` algorithm to mitigate popularity and gender biases in neural team formation baselines. Our results confirmed that we could notably enhance fairness regarding popularity bias while retaining utility. On the other hand, the severe bias in the datasets and consequently the teams in terms of gender bias, caused our post-processing method to fall short. Hence, for gender bias, no considerable changes were witnessed regarding fairness or utility. We recognize the importance of a comprehensive fairness-aware pipeline, emphasizing the need to address biases right from the initial stages of pre-processing. This approach should be synergistically combined with other essential phases, including in-processing and post-processing. Although post-processing methods offer value, their standalone capability can be constrained, especially in extreme cases where they might inadvertently introduce reverse discrimination. To enhance the robustness and fairness of our team recommendation models, we plan to incorporate in-processing methods within our pipeline. Furthermore, we are committed to mitigating potential biases in the datasets using pre-processing techniques. In our pursuit of building a holistic framework, we are also considering the integration of additional fairness notions, such as equalized odds, to furnish us with a more comprehensive toolkit for evaluating and mitigating biases.

References

[1] Muhammad Aurangzeb Ahmad, Arpit Patel, Carly Eckert, Vikas Kumar, and Ankur Teredesai. "Fairness in Machine Learning for Healthcare". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '20. Virtual Event, CA, USA: Association for Computing Machinery, 2020, pp. 3529–3530. ISBN: 9781450379984. DOI: 10.1145/3394486.3406461. URL: https://doi.org/10.1145/3394486.3406461.

[2] Adil Baykasoglu, Türkay Dereli, and Sena Das. "Project Team Selection Using Fuzzy Optimization Approach". In: *Cybern. Syst.* 38.2 (2007), pp. 155–185. DOI: 10.1080/01969720601139041. URL: https://doi.org/10.1080/01969720601139041.

[3] Suzanne T. Bell, Anton J. Villado, Marc A. Lukasik, Larisa Belau, and Andrea L. Briggs. "Getting Specific about Demographic Diversity Variable and Team Performance Relationships: A Meta-Analysis". In: *Journal of Management* 37.3 (2011), pp. 709–743. DOI: 10.1177/0149206310365001. eprint: https://doi.org/10.1177/0149206310365001. URL: https://doi.org/10.1177/0149206310365001.

[4] Alex Beutel et al. "Fairness in Recommendation Ranking through Pairwise Comparisons". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Anchorage, AK, USA: Association for Computing Machinery, 2019, pp. 2212–2220. ISBN: 9781450362016. DOI: 10.1145/3292500.3330745. URL: https://doi.org/10.1145/3292500.3330745.

[5] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. "Equity of Attention: Amortizing Individual Fairness in Rankings". In: *CoRR* abs/1805.01788 (2018). arXiv: 1805.01788. URL: http://arxiv.org/abs/1805.01788.

[6] Amin Bigdeli, Negar Arabzadeh, Shirin SeyedSalehi, Morteza Zihayat, and Ebrahim Bagheri. "Gender Fairness in Information Retrieval Systems". In: *Pro-*

ceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '22. Madrid, Spain: Association for Computing Machinery, 2022, pp. 3436–3439. ISBN: 9781450387323. DOI: 10.1145/3477495.3532680. URL: https://doi.org/10.1145/3477495.3532680.

[7] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. "Building classifiers with independency constraints". In: *2009 IEEE international conference on data mining workshops*. IEEE. 2009, pp. 13–18.

[8] Manoel B. Campêlo, Tatiane Fernandes Figueiredo, and Ana Silva. "The sociotechnical teams formation problem: a mathematical optimization approach". In: *Ann. Oper. Res.* 286.1 (2020), pp. 201–216. DOI: 10.1007/s10479-018-2759-5. URL: https://doi.org/10.1007/s10479-018-2759-5.

[9] Carlos Castillo. "Fairness and transparency in ranking". In: *ACM SIGIR Forum*. Vol. 52. 2. ACM New York, NY, USA. 2019, pp. 64–71.

[10] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. "Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*. Ed. by danah boyd and Jamie H. Morgenstern. ACM, 2019, pp. 319–328. DOI: 10.1145/3287560.3287586. URL: https://doi.org/10.1145/3287560.3287586.

[11] Richard J Chen, Judy J Wang, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. "Algorithmic fairness in artificial intelligence for medicine and healthcare". In: *Nature biomedical engineering* 7.6 (2023), pp. 719–742.

[12] Shi-Jie (Gary) Chen. "An Integrated Methodological Framework for Project Task Coordination and Team Organization in Concurrent Engineering". In: *Concurr. Eng. Res. Appl.* 13.3 (2005), pp. 185–197.

[13] Shi-Jie (Gary) Chen and Li Lin. "Modeling team member characteristics for the formation of a multifunctional team in concurrent engineering". In: *IEEE Trans. Engineering Management* 51.2 (2004), pp. 111–124. DOI: 10.1109/TEM.2004.826011. URL: https://doi.org/10.1109/TEM.2004.826011.

[14] Arman Dashti, Saeed Samet, and Hossein Fani. "Effective Neural Team Formation via Negative Samples". In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*. Ed. by Mohammad Al Hasan and Li Xiong. ACM, 2022, pp. 3908–3912. DOI: 10.1145/3511808.3557590. URL: https://doi.org/10.1145/3511808.3557590.

[15] Arman Dashti, Saeed Samet, and Hossein Fani. "Effective Neural Team Formation via Negative Samples". In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. CIKM '22. Atlanta, GA, USA: Association for Computing Machinery, 2022, pp. 3908–3912. ISBN: 9781450392365. DOI: 10.1145/3511808.3557590. URL: https://doi.org/10.1145/3511808.3557590.

[16] Arman Dashti, Karan Saxena, Dhwani Patel, and Hossein Fani. "OpeNTF: A Benchmark Library for Neural Team Formation". In: *CIKM*. ACM, 2022, pp. 3913–3917. DOI: 10.1145/3511808.3557526. URL: https://doi.org/10.1145/3511808.3557526.

[17] Leslie A DeChurch and Jessica R Mesmer-Magnus. "The cognitive underpinnings of effective teamwork: a meta-analysis." In: *Journal of applied psychology* 95.1 (2010), pp. 32–53.

[18] Edmund H. Durfee, James C. Boerkoel Jr., and Jason Sleight. "Using hybrid scheduling for the semi-autonomous formation of expert teams". In: *Future Gener. Comput. Syst.* 31 (2014), pp. 200–212. DOI: 10.1016/j.future.2013.04.008. URL: https://doi.org/10.1016/j.future.2013.04.008.

[19] Michael D. Ekstrand, Robin Burke, and Fernando Diaz. "Fairness and Discrimination in Retrieval and Recommendation". In: *Proceedings of the 42nd Inter-*

national ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR'19. Paris, France: Association for Computing Machinery, 2019, pp. 1403–1404. ISBN: 9781450361729. DOI: 10.1145/3331184.3331380. URL: https://doi.org/10.1145/3331184.3331380.

[20] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. "A comparative study of fairness-enhancing interventions in machine learning". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*. Ed. by danah boyd and Jamie H. Morgenstern. ACM, 2019, pp. 329–338. DOI: 10.1145/3287560.3287589. URL: https://doi.org/10.1145/3287560.3287589.

[21] *Fundamental Rights Report 2018*. European Union Agency for Fundamental Rights. Accessed: 2023-10-31. 2018. URL: https://fra.europa.eu/en/publication/2018/fundamental-rights-report-2018.

[22] Pratik Gajane. "On formalizing fairness in prediction with machine learning". In: *CoRR* abs/1710.03184 (2017). arXiv: 1710.03184. URL: http://arxiv.org/abs/1710.03184.

[23] Matthew E. Gaston, John Simmons, and Marie desJardins. "Adapting Network Structure for Efficient Team Formation". In: *AAAI*. Vol. FS-04-02. AAAI Press, 2004, pp. 1–8. URL: https://www.aaai.org/Library/Symposia/Fall/2004/fs04-02-001.php.

[24] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. "Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Anchorage, AK, USA: Association for Computing Machinery, 2019, pp. 2221–2231. ISBN: 9781450362016. DOI: 10.1145/3292500.3330691. URL: https://doi.org/10.1145/3292500.3330691.

[25] Thomas Grote and Geoff Keeling. "Enabling fairness in healthcare through machine learning". In: *Ethics and Information Technology* 24.3 (2022), p. 39.

[26] Rajni Gyanchandani, Bhavika Nathani, and Deepak Jaroliya. "Factors Affecting Team Performance in IT Sector: An Exploratory Analysis". In: *J-GIBS* 11.1 (2019).

[27] 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*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 2125–2126. ISBN: 9781450342322. DOI: 10.1145/2939672.2945386. URL: https://doi.org/10.1145/2939672.2945386.

[28] Radin Hamidi Rad, Hossein Fani, Mehdi Kargar, Jaroslaw Szlichta, and Ebrahim Bagheri. "Learning to Form Skill-Based Teams of Experts". In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. CIKM '20. Virtual Event, Ireland: Association for Computing Machinery, 2020, pp. 2049–2052. ISBN: 9781450368599. DOI: 10.1145/3340531.3412140. URL: https://doi.org/10.1145/3340531.3412140.

[29] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 3323–3331. ISBN: 9781510838819.

[30] Lu Hong and Scott E. Page. "Groups of diverse problem solvers can outperform groups of high-ability problem solvers". In: *Proceedings of the National Academy of Sciences* 101.46 (2004), pp. 16385–16389. DOI: 10.1073/pnas.0403723101. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.0403723101. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0403723101.

[31] Ajil Jalal, Sushrut Karmalkar, Jessica Hoffmann, Alex Dimakis, and Eric Price. "Fairness for image generation with uncertain sensitive attributes". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 4721–4732.

[32] Sunny Joseph Kalayathankal, John T Abraham, and Joseph Varghese Kureethara. "A Fuzzy Approach To Project Team Selection". In: *International Journal of Scientific Technology Research* 8 (2019).

[33] Chen Karako and Putra Manggala. "Using Image Fairness Representations in Diversity-Based Re-Ranking for Recommendations". In: *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. UMAP '18. Singapore, Singapore: Association for Computing Machinery, 2018, pp. 23–28. ISBN: 9781450357845. DOI: 10.1145/3213586.3226206. URL: https://doi.org/10.1145/3213586.3226206.

[34] Mehdi Kargar and Aijun An. "Discovering top-k teams of experts with/without a leader in social networks". In: *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*. Ed. by Craig Macdonald, Iadh Ounis, and Ian Ruthven. ACM, 2011, pp. 985–994. DOI: 10.1145/2063576.2063718. URL: https://doi.org/10.1145/2063576.2063718.

[35] Mehdi Kargar, Lukasz Golab, Divesh Srivastava, Jaroslaw Szlichta, and Morteza Zihayat. "Effective Keyword Search Over Weighted Graphs". In: *IEEE Transactions on Knowledge and Data Engineering* 34.2 (2022), pp. 601–616. DOI: 10.1109/TKDE.2020.2985376.

[36] Jon R. Katzenbach and Douglas K. Smith. "The Wisdom of Teams: Creating the High-Performance Organization". In: 1992.

[37] Kyriakos Kyriakou, Pınar Barlas, Styliani Kleanthous, and Jahna Otterbacher. "Fairness in Proprietary Image Tagging Algorithms: A Cross-Platform Audit on People Images". In: *Proceedings of the International AAAI Conference on Web and Social Media* 13.01 (July 2019), pp. 313–322. DOI: 10.1609/icwsm.v13i01.3232. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/3232.

[38] Theodoros Lappas, Kun Liu, and Evimaria Terzi. "Finding a team of experts in social networks". In: *Proceedings of the 15th ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009. Ed. by John F. Elder IV, Françoise Fogelman-Soulié, Peter A. Flach, and Mohammed Javeed Zaki. ACM, 2009, pp. 467–476. DOI: 10.1145/1557019.1557074. URL: https://doi.org/10.1145/1557019.1557074.

[39] Min Kyung Lee and Katherine Rich. "Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Yokohama, Japan: Association for Computing Machinery, 2021. ISBN: 9781450380966. DOI: 10.1145/3411764.3445570. URL: https://doi.org/10.1145/3411764.3445570.

[40] Hamed Loghmani and Hossein Fani. "Bootless Application of Greedy Re-ranking Algorithms in Fair Neural Team Formation". In: *Advances in Bias and Fairness in Information Retrieval*. Ed. by Ludovico Boratto, Stefano Faralli, Mirko Marras, and Giovanni Stilo. Cham: Springer Nature Switzerland, 2023, pp. 108–118. ISBN: 978-3-031-37249-0.

[41] Baolong Lv, Feng Liu, Yulin Li, Jianhua Nie, Fangfang Gou, and Jia Wu. "Artificial Intelligence-Aided Diagnosis Solution by Enhancing the Edge Features of Medical Images". In: *Diagnostics* 13.6 (2023). ISSN: 2075-4418. DOI: 10.3390/diagnostics13061063. URL: https://www.mdpi.com/2075-4418/13/6/1063.

[42] Jessica R Mesmer-Magnus and Leslie A DeChurch. "Information sharing and team performance: a meta-analysis." In: *Journal of applied psychology* 94.2 (2009), pp. 535–546.

[43] Preetam Nandy, Cyrus DiCiccio, Divya Venugopalan, Heloise Logan, Kinjal Basu, and Noureddine El Karoui. "Achieving Fairness via Post-Processing in Web-Scale Recommender Systems". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 715–725. ISBN:

9781450393522. DOI: 10.1145/3531146.3533136. URL: https://doi.org/10.1145/3531146.3533136.

[44] Hoang Nguyen, Radin Hamidi Rad, Fattane Zarrinkalam, and Ebrahim Bagheri. "DyHNet: Learning dynamic heterogeneous network representations". In: *Inf. Sci.* 646 (2023), p. 119371. DOI: 10.1016/j.ins.2023.119371. URL: https://doi.org/10.1016/j.ins.2023.119371.

[45] Alexandra Olteanu et al. "FACTS-IR: Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval". In: *SIGIR Forum* 53.2 (Mar. 2021), pp. 20–43. ISSN: 0163-5840. DOI: 10.1145/3458553.3458556. URL: https://doi.org/10.1145/3458553.3458556.

[46] Carol R Paris, Eduardo Salas, and Janis A Cannon-Bowers. "Teamwork in multi-person systems: a review and analysis". In: *Ergonomics* 43.8 (2000), pp. 1052–1075.

[47] Radin Hamidi Rad, Ebrahim Bagheri, Mehdi Kargar, Divesh Srivastava, and Jaroslaw Szlichta. "Retrieving Skill-Based Teams from Collaboration Networks". In: *SIGIR*. ACM, 2021, pp. 2015–2019. DOI: 10.1145/3404835.3463105. URL: https://doi.org/10.1145/3404835.3463105.

[48] Radin Hamidi Rad, Ebrahim Bagheri, Mehdi Kargar, Divesh Srivastava, and Jaroslaw Szlichta. "Subgraph Representation Learning for Team Mining". In: *WebSci*. ACM, 2022, pp. 148–153. DOI: 10.1145/3501247.3531578. URL: https://doi.org/10.1145/3501247.3531578.

[49] Radin Hamidi Rad, Hossein Fani, Mehdi Kargar, Jaroslaw Szlichta, and Ebrahim Bagheri. "Learning to Form Skill-based Teams of Experts". In: *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. Ed. by Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux. ACM, 2020, pp. 2049–2052. DOI: 10.1145/3340531.3412140. URL: https://doi.org/10.1145/3340531.3412140.

[50] Radin Hamidi Rad, Aabid Mitha, Hossein Fani, Mehdi Kargar, Jaroslaw Szlichta, and Ebrahim Bagheri. "PyTFL: A Python-based Neural Team Formation Toolkit". In: *CIKM*. ACM, 2021, pp. 4716–4720. DOI: 10.1145/3459637.3481992. URL: https://doi.org/10.1145/3459637.3481992.

[51] Radin Hamidi Rad, Shirin Seyedsalehi, Mehdi Kargar, Morteza Zihayat, and Ebrahim Bagheri. "A Neural Approach to Forming Coherent Teams in Collaboration Networks". In: *EDBT*. OpenProceedings.org, 2022, 2:440–2:444. DOI: 10.48786/edbt.2022.37. URL: https://doi.org/10.48786/edbt.2022.37.

[52] William B. Rouse, Janis A. Cannon-Bowers, and Eduardo Salas. "The role of mental models in team performance in complex systems". In: *IEEE Trans. Syst. Man Cybern.* 22.6 (1992), pp. 1296–1308. DOI: 10.1109/21.199457. URL: https://doi.org/10.1109/21.199457.

[53] Yuta Saito and Thorsten Joachims. "Fair Ranking as Fair Division: Impact-Based Individual Fairness in Ranking". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '22. Washington DC, USA: Association for Computing Machinery, 2022, pp. 1514–1524. ISBN: 9781450393850. DOI: 10.1145/3534678.3539353. URL: https://doi.org/10.1145/3534678.3539353.

[54] Anna Sapienza, Palash Goyal, and Emilio Ferrara. "Deep Neural Networks for Optimal Team Composition". In: *Frontiers Big Data* 2 (2019), p. 14. DOI: 10.3389/fdata.2019.00014. URL: https://doi.org/10.3389/fdata.2019.00014.

[55] Ashudeep Singh and Thorsten Joachims. "Fairness of Exposure in Rankings". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '18. London, United Kingdom: Association for Computing Machinery, 2018, pp. 2219–2228. ISBN: 9781450355520. DOI: 10.1145/3219819.3220088. URL: https://doi.org/10.1145/3219819.3220088.

[56] Mauro Sozio and Aristides Gionis. "The community-search problem and how to plan a successful cocktail party". In: *SIGKDD*. ACM, 2010, pp. 939–948. DOI: 10.1145/1835804.1835923. URL: https://doi.org/10.1145/1835804.1835923.

[57] Daniel Stokols, Kara L Hall, Brandie K Taylor, and Richard P Moser. "The science of team science: overview of the field and introduction to the supplement". In: *American journal of preventive medicine* 35.2 (2008), S77–S89.

[58] Damjan Strnad and Nikola Guid. "A fuzzy-genetic decision support system for project team formation". In: *Appl. Soft Comput.* 10.4 (2010), pp. 1178–1187. DOI: 10.1016/j.asoc.2009.08.032. URL: https://doi.org/10.1016/j.asoc.2009.08.032.

[59] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. "ArnetMiner: extraction and mining of academic social networks". In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*. Ed. by Ying Li, Bing Liu, and Sunita Sarawagi. ACM, 2008, pp. 990–998. DOI: 10.1145/1401890.1402008. URL: https://doi.org/10.1145/1401890.1402008.

[60] Ali Vardasbi, Fatemeh Sarvi, and Maarten de Rijke. "Probabilistic Permutation Graph Search: Black-Box Optimization for Fairness in Ranking". In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '22. Madrid, Spain: Association for Computing Machinery, 2022, pp. 715–725. ISBN: 9781450387323. DOI: 10.1145/3477495.3532045. URL: https://doi.org/10.1145/3477495.3532045.

[61] Sahil Verma and Julia Rubin. "Fairness Definitions Explained". In: *Proceedings of the International Workshop on Software Fairness*. FairWare '18. Gothenburg, Sweden: Association for Computing Machinery, 2018, pp. 1–7. ISBN: 9781450357463. DOI: 10.1145/3194770.3194776. URL: https://doi.org/10.1145/3194770.3194776.

[62] Lin Wang, Yifeng Zeng, Bilian Chen, Yinghui Pan, and Langcai Cao. "Team Recommendation Using Order-Based Fuzzy Integral and NSGA-II in StarCraft". In: *IEEE Access* 8 (2020), pp. 59559–59570. DOI: 10.1109/ACCESS.2020.2982647. URL: https://doi.org/10.1109/ACCESS.2020.2982647.

[63] Hyeongon Wi, Seungjin Oh, Jungtae Mun, and Mooyoung Jung. "A team formation model based on knowledge and collaboration". In: *Expert Syst. Appl.* 36.5 (2009), pp. 9121–9134. DOI: 10.1016/j.eswa.2008.12.031. URL: https://doi.org/10.1016/j.eswa.2008.12.031.

[64] Kyra Yee, Uthaipon Tantipongpipat, and Shubhanshu Mishra. "Image Cropping on Twitter: Fairness Metrics, Their Limitations, and the Importance of Representation, Design, and Agency". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW2 (Oct. 2021). DOI: 10.1145/3479594. URL: https://doi.org/10.1145/3479594.

[65] Armen Zakarian and Andrew Kusiak. "Forming teams: An analytical approach". In: *IIE Transactions* 31 (Jan. 1999), pp. 85–97. DOI: 10.1023/A:1007580823003.

[66] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. "FA*IR: A Fair Top-k Ranking Algorithm". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. CIKM '17. Singapore, Singapore: Association for Computing Machinery, 2017, pp. 1569–1578. ISBN: 9781450349185. DOI: 10.1145/3132847.3132938. URL: https://doi.org/10.1145/3132847.3132938.

[67] Meike Zehlike and Carlos Castillo. "Reducing Disparate Exposure in Ranking: A Learning To Rank Approach". In: *Proceedings of The Web Conference 2020*. WWW '20. Taipei, Taiwan: Association for Computing Machinery, 2020, pp. 2849–2855. ISBN: 9781450370233. DOI: 10.1145/3366424.3380048. URL: https://doi.org/10.1145/3366424.3380048.

[68] Meike Zehlike, Tom Sühr, Ricardo Baeza-Yates, Francesco Bonchi, Carlos Castillo, and Sara Hajian. "Fair Top-k Ranking with Multiple Protected Groups". In:

Inf. Process. Manage. 59.1 (Jan. 2022). ISSN: 0306-4573. DOI: 10.1016/j.ipm.2021.102707. URL: https://doi.org/10.1016/j.ipm.2021.102707.

[69] Meike Zehlike, Ke Yang, and Julia Stoyanovich. "Fairness in Ranking, Part I: Score-Based Ranking". In: *ACM Comput. Surv.* 55.6 (Dec. 2022). ISSN: 0360-0300. DOI: 10.1145/3533379. URL: https://doi.org/10.1145/3533379.

CHAPTER 5

Poster Presentations

To showcase the practical implications of our study, we engaged in the ***University of Windsor's 8th and 9th Annual Computer Science Demo Day***. During this event, we had the opportunity to discuss and exhibit the practical applications of our research project to professionals from various sectors of the technology industry. We opted for a poster presentation format, recognizing its strengths in visual communication and its effectiveness in capturing the interest of attendees. The following sections of this chapter will feature the posters that were displayed.

5.1 University of Windsor's 8th Demo Day

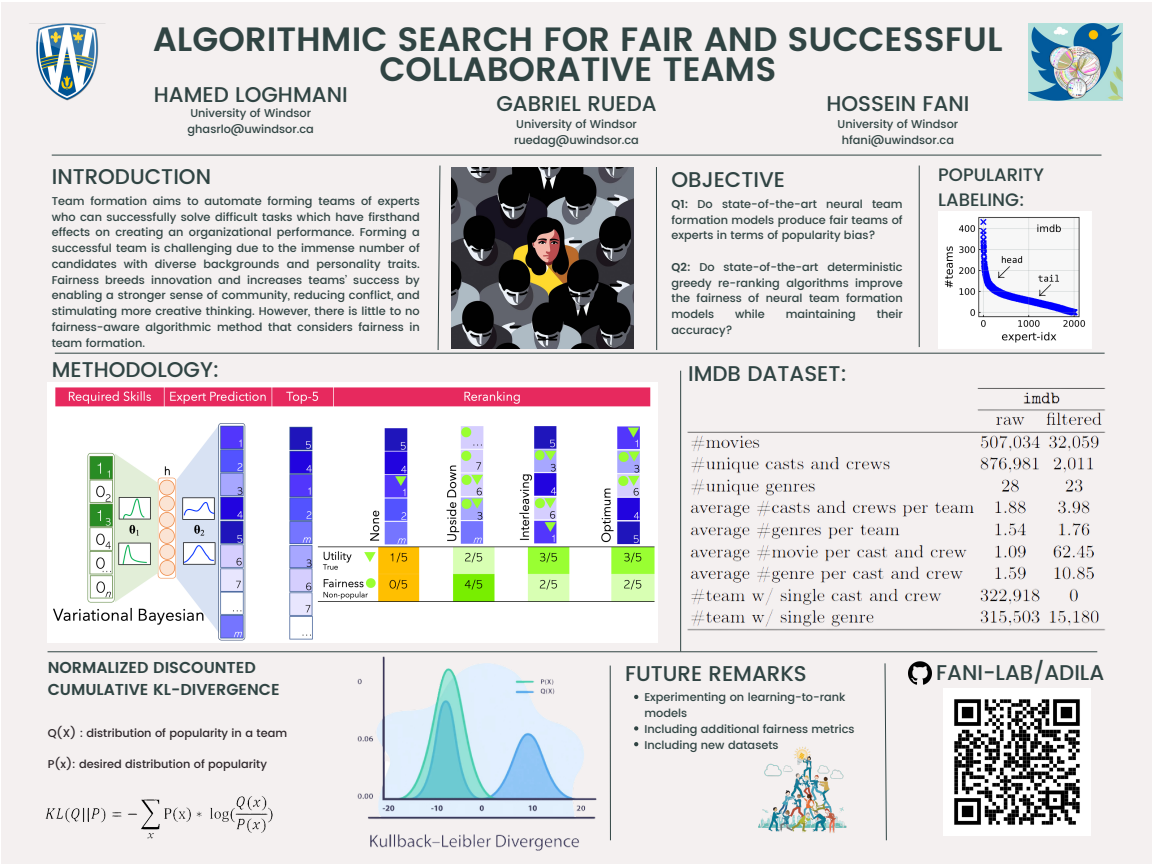


Fig. 5.1.1: The poster we presented at University of Windsor's 8th Annual Computer Science Demo Day

5.2 University of Windsor's 9th Demo Day

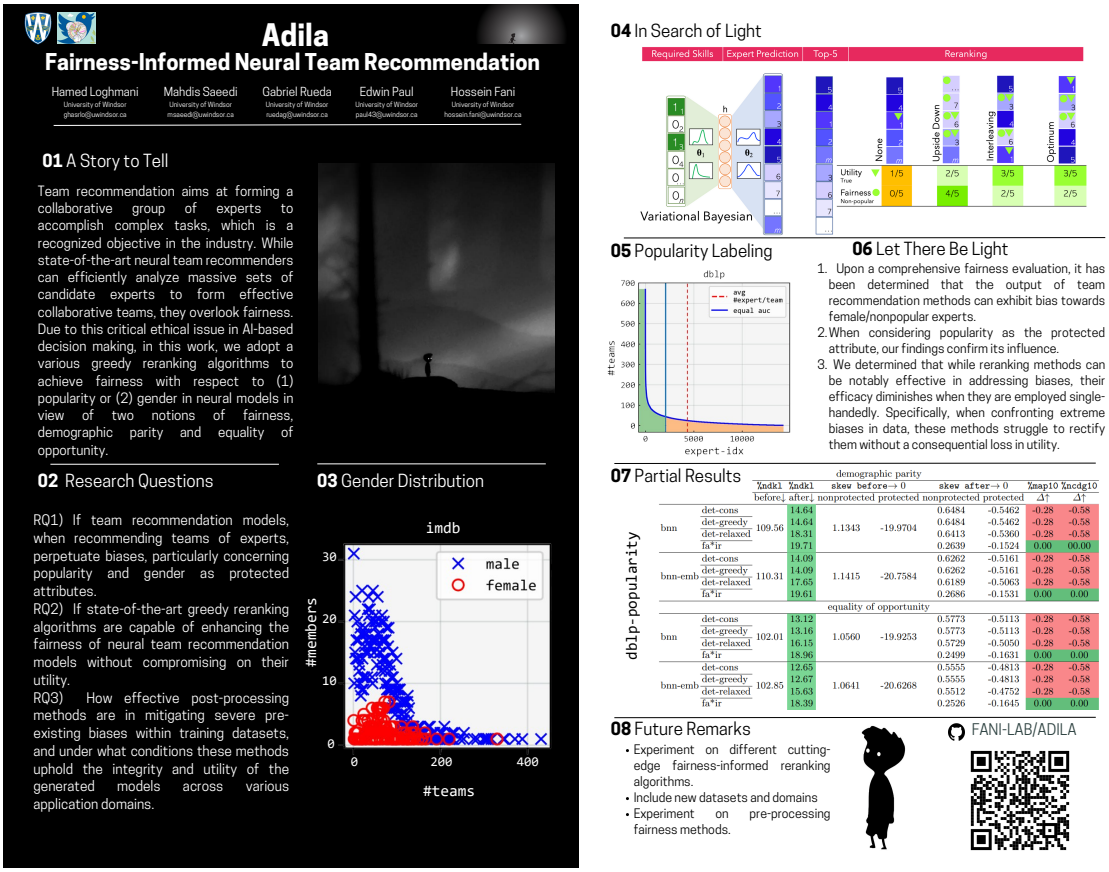


Fig. 5.2.1: The poster we presented at University of Windsor's 9th Annual Computer Science Demo Day

CHAPTER 6

Conclusion

6.1 Research Questions

This section presents the details of research questions that we answered through this thesis.

`RQ1`, delves into the potential biases in the output of state-of-the-art neural team recommendation models. Our goal is to determine if these models, when recommending teams of experts, perpetuate biases, particularly concerning popularity and gender as protected attributes. To conduct a meticulous evaluation, we have employed a Bayesian neural model, which is augmented with negative sampling heuristics[1]. Our investigation adopts two well-known fairness notions(demographic parity and equality of opportunity). We utilize skew and ndkl metrics to measure the fairness scores of the teams predicted by the model. These metrics are instrumental in determining the extent of fairness by examining how individuals with different demographic attributes are treated and provided opportunities within the recommended teams. Demographic parity evaluates whether individuals are selected for teams irrespective of their protected attributes, while equality of opportunity assesses the true positive rate of selection across different demographic groups. By harnessing these notions, we seek to unveil any potential biases and understand the efficacy of the Bayesian neural model in employing a fair team recommendation process.

`RQ2`, examines if state-of-the-art probabilistic and deterministic reranking algorithms are capable of enhancing the fairness of neural team recommendation models without

compromising on their utility. To address this question, we adopt and incorporate fa*ir [4], a distinguished probabilistic reranking algorithm, in addition to three deterministic reranking algorithms det_cons[2], det_greedy[2] and det_relaxed[2] into the predictive framework of our neural model, subsequently measuring the fairness and utility scores of the teams assembled. Our analysis conceptualizes team recommendation as a two-sided marketplace comprising two integral parties: (i) experts with specialized skills, from scientists to artists, and (ii) organizations, such as research laboratories or entities within the entertainment sector, seeking to recruit these experts for their teams. Goal to our exploration is the delicate balance between the success rate(also denoted as utility) and fairness in the teams proposed by neural team recommendation methodologies. We are particularly attentive to biases related to popularity and gender, while ensuring the skills are considered.

`RQ3` investigates how effective post-processing methods are in mitigating severe pre-existing biases within training datasets, and under what conditions these methods uphold the integrity and utility of the generated models across various application domains. Post-processing methods, applied after the training phase, are anticipated to correct biases, aiming to foster equitable and applicable models in real-world scenarios. However, the effectiveness of these methods may vary based on the degree of initial data bias, the domain of application, and the underlying model architecture. This inquiry seeks to identify the favorable conditions for post-processing interventions and to understand how these conditions may vary across different application domains. Key part to this examination is the exploration of potential trade-offs between bias mitigation and the preservation of model performance and utility. Through this exploration, RQ3 aims to contribute towards the broader narrative on ethical AI, shedding light on how post-processing methods can be harnessed to mitigate bias, thus promoting the responsible development and deployment of AI systems across a wide range of domains.

6.2 Results and Limitations

As indicated by the data in Tables 4.4.1 and 4.4.3, utilizing popularity as a protected attribute in the `imdb` and `dblp` datasets led to notable enhancements in various fairness metrics, while simultaneously maintaining the standards of utility metrics. Conversely, Tables 4.4.2 and 4.4.4 reveal that when gender is considered as the protected attribute, the fairness notions applied demonstrated that the teams were already balanced in terms of gender distribution. This resulted in no observable improvements in either fairness or utility metrics.

Addressing **`RQ1`**, a thorough evaluation of fairness prior to the reranking process revealed that biases towards certain protected attributes could emerge from our team recommendation methods. This bias poses a risk of disadvantaging specific protected groups, highlighting the importance of mitigating these biases to guarantee fair outcomes for all individuals, irrespective of their attributes or backgrounds. Comparatively, our findings indicate that while the fa*ir method is capable of preserving utility as measured by `map@10` post-reranking, deterministic reranking algorithms struggle to do so, as previously discussed by Loghmani et al. [3] in the early steps of our research. This trend holds true for `ndcg@10` as well, in both `imdb` and `dblp` datasets, when examining popularity or gender as the protected attribute in the context of demographic parity and equality of opportunity.

In response to **`RQ2`**, our analysis demonstrates a significant impact when popularity is considered as the protected attribute. Utilization of the `fa*ir` algorithm resulted in a notable advancement in fairness metrics, while maintaining overall utility metrics. In contrast, addressing gender as the protected attribute presented a more complex challenge. The initial bias present in our dataset was so substantial that, despite the algorithm's designation of teams as fair under our selected fairness criteria, the actual outcomes did not reflect this fairness. This discrepancy underscores the critical need for ongoing refinement and enhancement of our fairness criteria, particularly in situations where biases are deeply entrenched. It also highlights the necessity for more thorough and multifaceted evaluations to effectively address and mitigate such

ingrained biases in datasets. This approach is essential for ensuring more accurate and truly fair outcomes, especially in contexts where initial biases are significantly present.

In addressing `RQ3`, our investigation has led us to understand that while post-processing interventions can be effective in reducing biases, their impact is somewhat limited when they are the sole strategy employed. This limitation becomes particularly evident in scenarios involving `extreme` biases in datasets. In such cases, these methods often fail to fully correct the biases without a significant trade-off in utility metrics. It is important to recognize that `extreme` is a relative term and its implications can vary depending on several factors such as the specific problem at hand, the level of expertise of those involved, and the size of the dataset or team being considered.

To ensure more equitable and optimal outcomes, it is advisable to adopt a comprehensive approach that combines pre-processing, in-processing, and post-processing techniques. This holistic strategy is crucial, especially in cases where biases are deeply ingrained. The existing literature, including the work of Zehlike et al. [5], supports this view, suggesting that in instances of extreme bias, reliance solely on post-processing is generally inadequate for effective mitigation. As a result, in our analysis of these two datasets, we observed that using only post-processing methods did not lead to significant improvements in addressing gender bias. This finding further reinforces the necessity for a multi-faceted approach to bias mitigation in data-driven decision-making processes.

6.3 Runtime Analysis

In the following section, we perform a runtime analysis on our experiments. Figure 6.3.1 illustrates the runtime performance of various debiasing methods,*det_greedy*, *det_cons*, *det_relaxed*, and *fa*ir*, applied to the *bnn* and *bnn-emb* baselines results from *imdb* dataset. This analysis is conducted against the backdrop of two fairness notions: *equality of opportunity* and *demographic parity*.

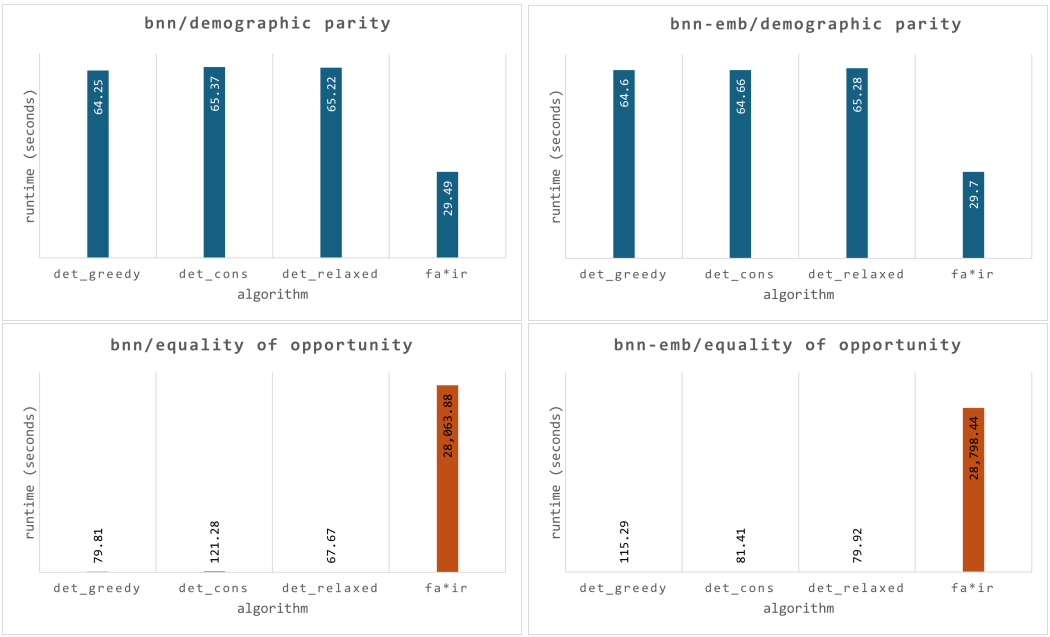


Fig. 6.3.1: Runtime of debiasing algorithms for bnn and bnn-emb baselines on imdb dataset

When assessing these methods under the *demographic parity* notion, a notable observation is the dominant speed of the *fa*ir* method. This method outperforms its counterparts in terms of runtime efficiency across both baselines, highlighting its practical advantage when demographic parity is the chosen fairness notion.

Conversely, when the *equality of opportunity* criterion is employed, the scenario shifts dramatically. Here, the runtime for the *fa*ir* method escalates significantly, surpassing that of the other debiasing techniques. This marked increase in runtime can be attributed to the additional computational steps required when applying *equality of opportunity* in tandem with the *fa*ir* method. Specifically, this involves recurrent recalculations of Table 4.4.5 for each team. This is mainly because when using *fa*ir* with *equality of opportunity*, the minimum proportion of the protected group is determined per the qualified set for each team, while in *demographic parity* it is calculated per dataset and Table 4.4.5 will be only calculated once.

The runtime for our deterministic debiasing method, *det_greedy*, *det_cons*, and *det_relaxed*, was close to each other and neither of them was dominantly faster compared to the others. It is notable that the same pattern of runtimes was witnessed on the *dblp* dataset as well.

6.4 Concluding Remarks and Future Work

In this thesis, we investigated the effectiveness of `fa*ir`, `det_cons`, `det_greedy` and `det_relaxed` algorithms in addressing biases related to popularity and gender within the framework of neural team recommendation baselines. Our findings demonstrated significant improvements in mitigating popularity bias, maintaining utility levels. However, we encountered challenges in rectifying gender bias due to its deep-seated nature in our datasets, resulting in limited success with our post-processing approach. Consequently, no substantial alterations were observed in terms of fairness or utility for gender bias.

This research highlights the criticality of adopting a comprehensive strategy that tackles biases from the earliest stages, starting with pre-processing. Integrating this approach with in-processing and post-processing phases is essential for a more effective bias mitigation process. Our experience suggests that while post-processing techniques are valuable, their effectiveness is limited when used in isolation, particularly in cases of severe biases. These techniques might also risk introducing reverse discrimination in certain scenarios.

To reinforce the fairness and effectiveness of our team recommendation models, we intend to integrate in-processing methods into our existing framework. Additionally, we aim to employ pre-processing techniques to further reduce dataset biases. Our future efforts will also involve exploring and incorporating additional fairness notions, such as equalized odds, thereby enhancing our ability to evaluate and mitigate biases more comprehensively. This holistic approach aims to establish a robust framework for fairness in AI-driven team recommendation, ensuring more equitable outcomes across different demographic groups.

References

[1] Arman Dashti, Saeed Samet, and Hossein Fani. “Effective Neural Team Formation via Negative Samples”. In: *Proceedings of the 31st ACM International*

Conference on Information & Knowledge Management. CIKM '22. Atlanta, GA, USA: Association for Computing Machinery, 2022, pp. 3908–3912. ISBN: 9781450392365. DOI: 10.1145/3511808.3557590. URL: https://doi.org/10.1145/3511808.3557590.

[2] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. "Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Anchorage, AK, USA: Association for Computing Machinery, 2019, pp. 2221–2231. ISBN: 9781450362016. DOI: 10.1145/3292500.3330691. URL: https://doi.org/10.1145/3292500.3330691.

[3] Hamed Loghmani and Hossein Fani. "Bootless Application of Greedy Re-ranking Algorithms in Fair Neural Team Formation". In: *Advances in Bias and Fairness in Information Retrieval*. Ed. by Ludovico Boratto, Stefano Faralli, Mirko Marras, and Giovanni Stilo. Cham: Springer Nature Switzerland, 2023, pp. 108–118. ISBN: 978-3-031-37249-0.

[4] Meike Zehlike, Tom Sühr, Ricardo Baeza-Yates, Francesco Bonchi, Carlos Castillo, and Sara Hajian. "Fair Top-k Ranking with Multiple Protected Groups". In: *Inf. Process. Manage.* 59.1 (Jan. 2022). ISSN: 0306-4573. DOI: 10.1016/j.ipm.2021.102707. URL: https://doi.org/10.1016/j.ipm.2021.102707.

[5] Meike Zehlike, Ke Yang, and Julia Stoyanovich. "Fairness in Ranking, Part I: Score-Based Ranking". In: *ACM Comput. Surv.* 55.6 (Dec. 2022). ISSN: 0360-0300. DOI: 10.1145/3533379. URL: https://doi.org/10.1145/3533379.

VITA AUCTORIS

NAME:	Hamed Ghasr Loghmani
PLACE OF BIRTH:	Ahvaz, Iran
YEAR OF BIRTH:	1997
EDUCATION:	
	NODET (National Organization for Development of Exceptional Talents), Highschool Diploma, Ahvaz, Iran, 2016
	Shahid Chamran University, B.Sc in Computer Engineering, Ahvaz, Iran, 2021
	University of Windsor, M.Sc in Computer Science, Windsor, Ontario, 2024