12-2022

Multivariate Fairness for Paper Selection

Reem Alsaffar
University of Arkansas, Fayetteville

Multivariate Fairness for Paper Selection

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Computer Science

by

Reem Alsaffar
Al-Nahrain University
Bachelor of Science in Computer Science, 2003
Al-Nahrain University
Master of Science in Computer Science, 2007

December 2022
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

Susan Gauch, Ph.D.
Dissertation Director

Brajendra Panda, Ph.D.
Committee Member

Laura Kent, Ph.D.
Committee Member

David Andrew, Ph.D.
Committee Member

Abstract

Peer review is the process by which publishers select the best publications for inclusion in a journal or a conference. Bias in the peer review process can impact which papers are selected for inclusion in conferences and journals. Although often implicit, race, gender and other demographics can prevent members of underrepresented groups from presenting at major conferences. To try to avoid bias, many conferences use a double-blind review process to increase fairness during reviewing. However, recent studies argue that the bias has not been removed completely. Our research focuses on developing fair algorithms that correct for these biases and select papers from a more demographically diverse group of authors. To address this, we present fair algorithms that explicitly incorporate author diversity in paper recommendation using multidimensional author profiles that include five demographic features, i.e., gender, ethnicity, career stage, university rank, and geolocation. The Overall Diversity method ranks papers based on an overall diversity score whereas the Multifaceted Diversity method selects papers that fill the highest-priority demographic feature first. We evaluate these algorithms with Boolean and continuous-valued features by recommending papers for SIGCHI 2017 from a pool of SIGCHI 2017, DIS 2017 and IUI 2017 papers and compare the resulting set of papers with the papers accepted by the conference. Both methods increase diversity with small decreases in utility using profiles with either Boolean or continuous feature values. Our best method, Multifaceted Diversity, recommends a set of papers that match demographic parity, selecting authors who are 42.50% more diverse with a 2.45% gain in utility. This approach could be applied when selectin conference papers, journal papers, grant proposals, or other tasks within academia.

Acknowledgments

I want to thank many who helped me along this journey and without whom I would not complete this research.

To my advisor, Dr. Susan Gauch: without your guidance, valuable advice, support and invaluable comments, I would not make it. You have always provided insightful feedback, which pushed me to sharpen my thinking and bring my work to a higher level. Thank you for all your kindness and excellent collaboration.

To my committee members, Dr. Brajendra Panda, Dr. David Andrews, and Dr. Laura Kent: Thank you for your important and useful comments. It has been an honor to present my research and discuss it with you. Likewise, thanks to all staff and faculty at the University of Arkansas for their providing their best work to the university and students.

To my husband, Haydar, and my daughters, Mayar and Haya: your love and support helped me in dark times. You believe in me and always encourage me to bring my best. My wonderful family, you have been amazing, and I can not thank you enough for your support and patience with my craziness.

To my Mother (Ahlam) and my siblings (Mohammed, Hamza, Russul): thank you for supporting me and being there whenever I need you. I am grateful for your understanding and patience because I have been away for several years and missed many events.

Dedications

To the soul of my father, Baqer
باقـــــر

Table of Content

1 Introduction.....9

1.1 Motivation 9
1.2 Goals.....12
1.3 Approaches.....13

2 Related Work 15

2.1 User Profiles 15
2.1.1 Expert Profiles 17
2.1.2 Demographic Profiles 18

2.2 Bias in Academis..... 20

2.3 Bias in Peer Review 21
2.3.1 Paper Assignment Fairness 22

2.4 Fairness..... 23
2.4.1 Demographic Parity.....23
2.4.2 Fairness in Machine Learning 24
2.4.2 Fairness in Ranked Outputs…..... 26

3 Research Plan 29

3.1 System Overview 29

3.2 Demographic Profile Construction based on Protected and Nonprotected groups 30
3.2.1 Data Extraction 30

3.3 Paper Profile Formation..... 35

3.4 Paper Quality Profile..... 36

3.5 Pool Distribution (Demographic Parity..... 37

3.6 Recommending the Papers..... 37
3.6.1 Overall Diversity Method 38
3.6.2 Multi-Faceted Diversity Method 39

4 Experimental and Evaluation..... 41

4.1 Dataset 41

4.2 Baseline and Metrics 44

4.3 Results 45

4.4 Comparison with the Balseline..... 46

4.5 Discussion……..... 51
4.5.1 Diversity Gain Comparison 52
4.5.2 Demographic Parity Comparison 53

5 Conclusion 56

5.1 Summary 56

5.2 Future Work 57

List of Tables

Table 3.1 Demographic features categories ….. ……..……..31

Table 3.2 Career stage weight allocation…….……...…….……...….33

Table 3.3 Countries HDI sample…….……..….…….....34

Table 3.4 Sample of the authors' raw data …….…….……..….35

Table 4.1 Composition of our dataset.…….…….……...…….....…..42

Table 4.2 Demographic participation from protected groups in three current conferences ……..43

Table 4.3 Protected group participation for the recommender algorithms using Boolean and Continuous profiles.…..……..…..…48

Table 4.4 Proportion of recommended papers from each conference.……..…….….49

Table 4.5 Diversity gain and utility savings for the overall diversity and multi-faceted diversity algorithms versus the Baseline for Boolean profiles…….…….….49

Table 4.6 Diversity gain and utility savings for the overall diversity and multi-faceted diversity algorithms versus the Baseline using Continuous values.…….…….….50

Table 4.7 Demographic parity similarity and utility savings for the overall diversity and multi-faceted diversity algorithms versus the baseline (Boolean).……...….51

Table 4.8 Demographic parity similarity and utility savings for the overall diversity and multi-faceted diversity algorithms versus the baseline (Continuous).……...……..….51

Table 4.9 Diversity Gain with Boolean and Continuous weights profiles.……..…..….52

Table 4.10 Demographic Similarity with Boolean and Continuous weights profiles ……..54

List of Figures

Fig 3.1 System framework……... …….……..…..….29

Fig 4.1 Protected Group Membership of Authors for Three Current Conferences…….…..43

Fig 4.2 Improvement in Protected Group Participation between the SIGCHI2017 and our Paper Recommendation Algorithms when using Boolean Profiles……..……..….46

Fig 4.3 Improvement in Protected Group Participation between the SIGCHI2017 and our Paper Recommendation Algorithms when using Continuous Profiles.……..…….....47

Introduction

1.1 Motivation

We are living in the 21st century and the modern world is a very diverse world that asks us to strive to break down barriers to inclusion. Transportation and telecommunication technologies diminish distances everyday so we can live in a global village. It is now common for Americans go to work driving German cars with components made from Korean steel and Malaysian rubber and containing parts made in the United States and Japan. Diversity leads to globalization and vice versa. Organizations need to diversify their workplace and employ people from different ages, genders, nationalities, religions, languages, abilities, and regions so that can survive in this world and have a competitive advantage (Saxena 2014). The New York Times Company is one example of a company that embraces this philosophy and its policies have resulted in a more diverse work environment. In order to reflect the society that they report on, they are improving their recruiting to build a diverse workplace. Last year, people of color made up 48% of their new hires. Women now comprise 52% of their staff, a large increase compared to just to 38% in 2015. At the same time, members of underrepresented minorities now make up 34% of their employees compared to only 27% a few years ago (The New York Times 2021).

However, there is still discrimination against people because of their race, color, gender, religion, national origin, disability, and age (Sugarman, et al., 2018). For instance, in 2014 an article in the New York Times shows that many big companies are still a man's world. At Google, for example, males occupy the highest proportion by far, accounting for 70% of the employees. Moreover, men make up 83% of the engineering employees and 79% of the managers (Manjoo 2014). The company's demographics have not changed much since 2014; in 2021, Google's annual report showed that the company workforce is still white and male environment. 67.8% of Google staff are men and 51.7% are white (Google 2021).

Federal law in the US prohibits discrimination against people because of their race, color, gender, religion, national origin, disability or agentic information and age under the US Equal Employment Opportunity Commission (EEOC) law. These groups have been called *protected* groups (eeoc.gov). However, these protected groups still face the problem of discrimination throughout American society and academia is no exception. Although the organizations and universities are working to achieve diverse students, staff, and faculty, different kinds of biases are still evident, e.g., gender, race, nationality, language. Bias in academia affects admissions at the undergraduate and graduate levels, hiring and promotion of professors, and the selection of university leaders. Some studies show that white men have more job opportunities in universities than women or people of color. For example, a study shows that only 38% of tenure-track positions were awarded to women (Flaherty, 2016).

The situation in Computer Science is very similar and we are a long way from achieving diversity. An article by the CEO and founder of Piazza shows that, in the computer science field, the majority of the students are men and that affects the communication among students such that online tools promote collaboration among males only. This reality, and perception, may discourage women from applying to male or mostly male environments such as Computer Science. ("Women in Computer Science, 2021) (Code.org, 2020) and (Sankar, 2015) document the fact that, of the graduates from Computer Science, only 18% are women and also only 18% are minorities. Women are also underrepresented in Computer Science professional positions. In the US, despite the proportion of the women employees being 45% across all industries and only 30% of the computer scientists in industry, women make up only 21.9% of computer science faculties (Zippia 2019).

These statistics are reflected in the lack of diverse speakers as Computer Science conferences since there are few paper authors that are female and/or members of minority groups. These demographic imbalances are also evident in conference attendees where minorities are underrepresented (Jones et al. 2014). Racial, gender and other types of discrimination during the reviewing stage can also prevent members of minorities presenting at major conferences. Many studies indicate that bias among reviewers and editors might lead to bias in choosing papers for publishing. Fewer women in the committee may affect the decisions when choosing the papers. As an example, in 2016, the proportion of women in mathematical journals and AI and robotics frontiers was very low compared to men (Helmer et al., 2017). There is also evidence that bias can occur in committee decisions. The submissions to biosciences journal eLife between 2012 and 2017 indicate that the reviewers tend to accept papers whose authors have the same gender and from the same geographic region as themselves (Murray et al., 2019). Addressing this, SIGCHI, one of the highest impact ACM conferences, announced that its goal for 2020 is increasing the diversity of its Program Committee (SIGCHI 2019).

To try to avoid bias, many conferences use a double-blind review process, hoping to solve the problem of discrimination and increase fairness during reviewing. However, merely using a double-blind review process fails to solve the problem of discrimination (Cox and Montgomerie, 2019) (Lemire, 2020). Recent studies argue that bias has not been removed completely even when using double-blind reviewing and not all fields have adopted this type of review. In Computer Science for example, when a paper was submitted in a conference, it might be already published on e-print or electronic journals, i.e. arxiv and eccc, or the reviewer might have already seen a talk about the project. Reviewers can frequently guess who the authors are (Barak 2018), so the review process is not actually double-blind. Computer Science and Physics

are two fields that promote sharing and openness among researchers because they are young fields, so it is very easy to infer the authors in these fields even when using double-blind review (Palus 2015). Another experiment by (Cox and Montgomorie, 2019) showed that double-blind review did not produce major differences in the results compared to single-blind review when the authors worked on publications from 2010 to 2019 submitted in ecology journal. Some major conferences that had been using double-blind review have gone back to use single blind since they realized that many systems show hidden fields in the documents that reveal the authors' identities.

Several studies have identified specific demographic features that can be a source of bias and we use these features to model the authors in our data set. The features most frequently identified are gender (Lerback and Hanson, 2017) (Cannon, et al., 2018), Ethnicity (Cannon, et al., 2018), Career Stage (Lerback and Hanson, 2017), University Rank (Flaherty, 2018) (Bowman and Bastedo, 2011) and geolocation (Wu, 2009) (Jacob and Lefgren, 2011).

1.2 Goals

As previously discussed, there is a need for greater diversity in Computer Science discussed above yet studies show that paper submissions to conferences from authors from majority groups are more likely to be accepted than the ones that come from minorities (Jaschik 2016). Thus, our goal is modifying the selection process for papers that have been submitted to conferences in order to improve the diversity of the authors whose work is presented while also minimizing any decrease in the quality of papers presented at the conference. We will focus on publications at major computer science conferences.

In general, papers are rated by a review process on a scale similar to *accept, weak accept, weak reject, reject.* The highest quality papers should obviously be selected, and the lowest

quality rejected, but there is generally a large pool of borderline papers that could be accepted. By considering author diversity as part of the selection criteria, it is our belief that acceptable papers can be chosen with little negative impact on conference quality but a possible large impact on conference diversity. Our approach is based on building a profile for each paper that reflects the paper's overall quality and also models the diversity of the paper authors. The quality profile would be based on the reviewers' ratings whereas the demographic profile would be based on the features discussed earlier. This multi-faceted profile is then used by a recommender module to select papers for inclusion in the conference. Our fair recommender system then recommends papers for inclusion in the conference balancing the goals of increasing the diversity of the authors whose work is selected for presentation while minimizing any decrease in the quality of papers presented. This system can be applied during conferences reviews, journal paper reviews, while awarding grant proposals, and other tasks within academia. To solve this challenge, we have put related goals as described in the following section:

Goal1: Profiling papers based on Boolean and Continuous author demographics.

Goal2: Recommending papers using fair approaches that balance quality and diversity.

Goal3: Achieving Demographic Parity between accepted authors and all authors of submitted papers.

1.3 Approaches

In this dissertation, we present multiple fair recommendation algorithms that balance two aspects of a paper, its quality and the authors' demographic features, when recommending papers to be selected by the conference. Because information about the review process is generally confidential, we simulate the results of the review process by creating pools of papers from related conferences within a specific field that have different impact factors. The highest impact

factor conference papers will play the role of the papers that are rated most highly by the reviewers, the middle impact factor conference papers those with the second best reviews, and papers published at the conference with the lowest of the three impact factors will be treated as papers with lower reviews.

First, we develop demographic profiles to model the authors of each paper in our pool of papers. Our demographic profiles will be comprised of five demographic attributes: Gender, Ethnicity, Career Stage, University Rank and Geolocation. We explore two main types of demographic profiles: 1) Boolean profiles in which each attribute is represented using a Boolean (TRUE (1) if from the underrepresented class, FALSE (0) otherwise); and 2) continuous value profiles in which demographic attributes may be modeled using a wider range of values from 0.0..1.0. Next, each paper's quality is modeled using the impact factor of the conference in which it actually appeared. We then create fair algorithms to select papers for the conference based on both quality and diversity. We evaluate our algorithms to determine which provides the best tradeoff between increased diversity and decreased quality. To the best of our knowledge, previous work to increase fairness has used one attribute at a time with Boolean values only. Thus, our research should contribute new algorithms that consider multiples attributes with Boolean and continuous values simultaneously. Our main contributions in this work are:

- Modelling author demographics using profiles that contain multiple demographic features.
- Developing and evaluating fair recommendation algorithms for paper selections that balance quality and diversity.
- Achieving demographic parity between the accepted authors with the pool of all authors.

Related Work

Our work focuses on fair recommendation, i.e., paper recommendation that takes the authors' demographic profiles into account with papers' quality in order to provide a balanced set of recommended papers across multiple demographic groups. Because this work is based on previous work in Information Retrieval (IR), data mining, fair recommendation, and statistics, we will summarize research in those related areas. The first section presents research approaches to user profiling and how the researchers model demographic and quality profiles. In the second section, we discuss evidence of bias in academia, including the paper review process. Finally, we explore different fair recommendation approaches the authors apply in order to decrease discrimination and add more diversity to the recommended items.

2.1 User Profiles

User profiling is the process of representing the user's interests, characteristics and other personal attributes based on analyzing their interactions. It can be used to better understand the users' intentions and develop personalized services to better assist users. For example, developers can customize systems to match users' preferences or improve the accuracy of results retrieved from web searches (Gauch et al., 2007). User profiles can also be utilized by businesses to promote certain products during a campaign (Kanoje et al., 2015). User profiles can be used to infer different types of profiles such as author profiles, expert profiles, and demographic profiles. In our work, we create demographic profiles for authors, but we survey literature related to other types of profiles for context.

Many techniques have been used to build accurate user profiles and many researchers have been working in this field. (Gauch et al., 2007) surveyed multiple techniques that can be used to gather and represent information about users via their profiles. In particular, user profiles

can be represented as weighted keywords, semantic networks, or weighted concepts. They mentioned that user profiles are used many applications to provide personalized access to information such as email, electronic newspapers, and web search. The information on which to build the profiles can be collected either explicitly, i.e., directly from user input, or implicitly, i.e., by collecting information by monitoring user activities. Profiles may also be dynamic (updated over time) or static (remain unchanged once built).

User profiles have been incorporated into many applications. (Trajkova and Gauch, 2004) built ontology-based user profiles to provide personalized search that increased search accuracy by better identifying results that matched user interests. They built the profile implicitly by monitoring the user browsing activities and classifying the visited web pages with respect to an ontology of over 1000 concepts. Some scholars applied user profile on the recommendation area to enhance the recommender results. (Sugiyama and Kan, 2010) investigated building user profile when recommending papers to a user. They were working on modeling the old research papers of the users to extract their preferences and also modeling the papers that cite the work. (Labille et al., 2015) have worked addressed the same problem, but instead of using a keywords vector to represent the profiles and the documents, they used concepts vectors, a flattened representation of the user's ontological profile. The user profiles were constructed by classifying all the papers for an author (Chandrasekaran et al., 2008) and later extended to recommend papers for users of the CiteSeerX digital library (Kodakateri Pudhiyaveetil et al., 2009). In both cases, papers were recommended based on conceptual matches between the user's profile and the documents in the corpus. Recently, (Shu et al., 2019) used user profiles to detect fake news by analyzing the connection between the user profile on social media and the fake news. They used multiple features to build the profile such as age, location, profile image, etc. They studied the

sharing activities for the users on social media and compared them with the group users who were sharing fake and real news, so they can analysis the differences between their profiles features in order to distinguish between the fake news from real news.

One of the challenges that researchers face when building user profiles is that user preferences change over time. To avoid this issue, recommender systems need to take time into account and use temporal information. Since this information might not be easily available, scholars are working on inferring it from available features. For instance, (Alkan and Daly, 2020) constructed dynamic user profiles by determining product reviews to extract temporal information and profile the users. In their paper, they designed a system to infer the age category preference of users and then use it to obtain a dynamic recommendation to suggest products to the user in the future. They argue that the recommendations they have built using this dynamic aspect are more accurate and predictable.

2.1.1 Expert Profiles

Expert profiling is a method to describe the skills and interests of researchers, so it is an important step to identify the right person for a specific task. Enterprises could use this idea when they have projects to find the skilled people they already have or identify talented staff to recruit (Gauch et al., 2007). This approach could be useful for academic institutions such as universities and conferences. For example, it can help them find qualified researchers to review papers. (Balog and Rijki, 2007) presented their work in automatically finding the areas that a person is expert in and they represent this in a topical profile. They also focused on finding the social profile for the expert and this profile captures the connections of the person. For their experiments, they used a big dataset, the W3C corpus from large enterprises. For the topical profile, they implemented two different methods: The first approach applies information retrieval

techniques to get documents that are related to the researcher and measure how relevant to the user are. In this method, they depend on the name and the email for the experts to measure the relatedness. In contrast, the other method works on both the user and keywords that represent the knowledge areas for that user. They work on these two vectors to estimate the skills for the experts based on the overlap between them. Other researchers explore using more than one dataset to build expert profiles using similar ways. For example, (Deng et al., 2007) used DBLP bibliography and Google Scholar as datasets in their work to find experts for a specific task in the academic field. Because DBLP contains only experts' names and their papers titles, Google Scholar was used to address these limitations and complement the data. The expert profiles were built similarly to the previously discussed approach, but they extended their profiler to include papers citations as a measure of importance. Some researchers involve expert profiling in academia to enhance the reviewer assignment process. For instance, (Sateli et al., 2017) proposed a text-based expert finding approach to represent the expertise as a set of weighted keywords. They built the profiles based on extracting the keywords from the author's home page and their publications. Then, users can use this system to find the matching expertise by entering the keywords that related to what they are looking for.

2.1.2 Demographic Profile

Demographic attributes can be used to study human health, population variation, and statistical theory (Chappelow, 2019). Recently, many researchers are using demographic attributes to increase fairness in multiple areas. (Galhotra et al., 2017), for example, study software fairness and discrimination and they have proposed approaches to evaluate whether or not software, e.g. risk-assessment calculations, is biased based on some demographic attributes. In particular, they studied the effect of attributes such as occupation, number of work hours, education level, gender

and race. Other researchers focus on including demographic attributes in recommender systems and they argue that discrimination and unfairness can appear in these systems due to bias in the algorithms or training data. By incorporating explicit demographic profiles, researchers hope to develop recommender systems that limit unfairness and discrimination (Farnadi, 2018). Because needs vary from field to field, not all fields use the same demographic attributes. For instance, the demographic attributes such as gender, race, education and age are widely used in academic field as they are important to build the user profiles (Cochran-Smith and Zeichner, 2009) (Alhajraf and Alasfour, 2014).

One important area of research is inferring demographic attributes using users' names, web pages and other sources of information such as users' social networks. For example, (Zhong et. al., 2015) was using users' profiles on the social network to extract gender, age, education background and marital status. As a result of this research, there are some libraries available to infer gender, age, race, etc. (Santamaría and Mihaljević, 2018) have published a paper to study and compare some existed services that infer gender from names. They apply these gender services on a dataset of 7076 labeled names. The APIs they have tested are: Gender API, gender-guesser, genderize.io, NameAPI and NamSor. Ethnicity and nationality are also inferred by multiple researchers including using data mining techniques. Along these lines, (Ye et al., 2016) built a classifier based on 57 million names on contact lists collected from a big internet organization. From this training data, they can infer 39 different nationalities in a taxonomy that covers 90% of the world population. Similarly, by applying their classifier to Twitter data, they are able to infer ethnicity as well. Their classifiers, called NamePrism (Ye et al., 2017) is publically available and can be used to infer gender and ethnicity. Genderize.io, developed by (Strømgren 2016), is also publically available and it is used to infer gender. In our research, we

use a NamSor API to extract gender from the user's names. This tool was built based on 142 languages and the overall gender precision and recall of this tool are 98.41% and 99.28% respectively (blog, NamSor, 2018).

2.2 Bias in Academia

Bias means tending to unfairly support or oppose peoples or ideas. Bias is an area of concern within academic fields. Bias in research can be seen when preferring one outcome or result over others during the testing or sampling phase, and also during any research stage, i.e. design, data collection, analysis, testing and publication (Pannucci and Wilkins, 2010). Some scholars argue that bias occurs during the committee review also. (Bornmann and Daniel, 2005) discussed the bias that might appear in the committee decisions when awarding doctoral and post-doctoral research fellowships. They focused on some sources of bias including gender, nationality, major field of study, and institutional affiliation to study their influence on decision making. They concluded that there was some evidence that gender, major field of study, and institutional affiliation caused bias when selecting doctoral fellowships, but that nationality did not seem to be a source of bias.

(Gabriel, 2017) conducted a study to investigate discrimination in British academia focusing on ethnicity. The results showed that the proportion of black professors in the UK is only 0.45% of all professors compared to their populations in academic staff which is 1.45%. Furthermore, when they considered gender with ethnicity, they found that the proportion of black female professors was only 0.1% of all professors. On the other hand, they found encouraging signs that hiring strategies are improving and that bias is decreasing. As for the USA, Flaherty (Flaherty, 2019) conducted a study to investigate discrimination in the US college faculty focusing on ethnicity. The results showed that the proportion of black professors is only 6% of

all professors compared to white professors' percentages which are 76%. More recently, an article published by researchers from Stanford Graduate School of Education in 2021 showed that, in the United States, more doctoral degrees have been earned by women than men. Despite this, women are still less likely than men to receive tenured positions, have their research published, or obtain leadership roles in academia. After analyzing one million doctoral dissertations from US universities, they found that the authors whose topics are related to women or who used methodologies that refer to women have decreased career prospects versus those related to men (Andrews L., 2021).

2.3 Bias in Peer Review

Other studies discuss the lack of fairness in the peer review process that has a major impact on accepting papers in conferences. An article published by (Lerback and Hanson, 2017), shows that the bias against women of all ages is still an issue. After analysis of data between 2012 and 2015 for the journals of the American Geophysical Union (AGU), the results demonstrated that women were given fewer opportunities to be reviewers in the journals compared to their number in the community and the journals as authors. The reasons behind that are: the authors and editors propose female reviewers less frequently than male reviewers and many women refuse to do the reviews, although they found that the first reason is the main one. A similar study was published by (Murray et al., 2019) indicated that there is evidence of existing bias in peer review when they studied the submissions from 2012 to 2017 to the biosciences journal eLife. Their results showed that bias is still involved in the reviewing process and the reviewers tend to accept the papers whose authors have the same gender and are from the same region. A study by (Tomkins, 2017) provided evidence that single-blind review provided a disparate advantage to papers submitted by known authors and authors working in high rank institutions. However,

there is still evidence that bias is involved int the review process. Although using double-blind reviews might decrease bias against minorities, some researchers demonstrate that bias still exists in the reviewing process. (Cox and Montgomorie, 2019) analyzed data from an ecology journal publication for 2010-2018 with single-blind and double-blind review. They concluded that the double-blind review did not increase the proportion of females significantly compared with single-blind review.

Several researchers propose methods to improve the quality and fairness of peer review during the paper assignment process, one of the first steps in the peer review process is finding willing reviewers and assigning reviewers to papers, without addressing solutions to ensure that the reviewers are not affected by the paper author demographics that accepted papers have good quality.

2.3.1 Paper Assignment Fairness

Some researchers have explored fairness when choosing a suitable reviewer to review a paper. (Long et al., 2013) considered the goodness aspect and fairness aspect to solve issues in Paper-Reviewer Assignment (PRA). For the goodness part, they wanted to maximize the topic coverage of the assignment, so they suggested a new approach called Maximum Topic Coverage Paper-reviewer Assignment (MaxTC-PRA) to ensure assigning papers to reviewers with maximizing the total number of distinct topics of papers covered by the chosen reviewers. To measure fairness, they enumerated different types of conflicts of interest (COI) between an author and a reviewer that should be avoided when assigning a reviewer. To evaluate their work, they collected published papers in KDD 2006-2010 and the program committee of ICDM 2010 and KDD 2010 as reviewers. They conclude that their method outperformed the other algorithms based on topic coverage. Some researchers consider the problem of reviewing the most

disadvantaged papers (interdisciplinary papers) to increase the fairness and accuracy in the reviewing process. (Stelmakh et al., 2021) focused on fairness and statistical accuracy in assigning papers to reviewers in conferences during the peer review process. For the fairness aspect, they improved the quality of reviewing the interdisciplinary papers depending on an incremental max-flow procedure by applying the max-min fairness aspect. Max-min fairness is a method considering the least qualified reviewers to maximize the paper quality. For the accuracy aspect, they improve the way of selecting the best papers for publishing during the peer review process and consider the noise in the reviews and subjective reviewers' opinions. They evaluate their work by applying two experiments with synthetic data using Amazon Mechanical Turk. The results show that their algorithm has best fairness and the quality is increased when the fairness of the assignment is increased.

Most of these studies propose methods to improve the quality of the reviewer assignment process without addressing solutions to ensure that the reviewers are not affected by the paper author demographics that accepted papers have good quality. We contribute to this area by creating author profiles with multiple demographic features and using them in new fair recommendation algorithms to achieve demographic parity when selecting papers for inclusion in a conference.

2.4 Fairness

Fairness is important in making financial, scholastic, and career decisions. As we rely more and more on computational methods to make decisions, it is clear that fairness and avoidance of bias in algorithms an important area of research. Many researchers are focusing on this problem, trying to improve computationally-driven decisions to make them fairer. In this section, we will discuss some of the previous and ongoing research in this area.

2.4.1 Demographic Parity

In the United States, several protected classes are legally protected against discrimination, i.e., race, color, religion, national origin, sex, age, physical health, etc. ("Protected group", 2020). These protected groups have been targets of discrimination and it is important that people and algorithms make fair financial, scholastic, and career decisions. To avoid bias, it is not enough to just ignore protected attributes while making a decision because it is often possible to predict these attributes from other features. To achieve fairness, many approaches aim for demographic parity, which is when members of the protected groups and non-protected groups are equally likely to receive positive outcomes. However, this requirement generally causes a decrease in utility. Yang et al. (Yang & Stoyanovich, 2017) focus on developing new metrics to measure the amount of bias present in ranked outputs by measuring the lack of demographic parity in ranked outputs. Zehlike et al. (Zehlike, et al., 2017) also address the problem of improving fairness in the top-K ranking problem over a single binary type attribute when selecting a subset of candidates from a large pool. Their method was designed to pick from two queues (one for protected candidates and the other for non-protected). It maximizes utility subject to a group fairness criteria and ensuring demographic parity at the same time. We extend this work by using multiple binary attributes when picking a subset of authors from the pool to achieve demographic parity. We also incorporated the diversity and the quality of the authors during the selection process to minimize the utility loss and maximize the diversity. Recently, some authors have been working on Generalized Demographic Parity (GDP) which is a group fairness metric for continuous and discrete features, to make fairness metrics more accurate. (Jiang et. al, 2022) proposed their method by displaying the relationship between joint and product margin distributions distance. They demonstrate two methods named histogram and kernel with linear

computation complexity. Their experiment showed that GDP regularizer can reduce bias more accurately.

2.4.2 Fairness in Machine Learning

As we rely more and more on computational methods to make decisions, it is clear that fairness and avoidance of bias in algorithmic decisions are of increasing importance. Many research investigations show that machine learning approaches can lead to biased decisions and that the data itself can be a source of this bias. Sometimes, machine learning models are trained on biased data and this will cause discrimination in the results that perpetuate historical discrimination (Asudeh, 2019). Another source of bias is limitations of the features related to the protected group. There may be fewer features relevant to a minority group features or the data for features related to the minority group may be less reliable. In addition, by the very fact that the protected group is often a minority, there is likely to be less training data for the protected group relative to the majority group (Zhong, 2018). Thus, researchers are working to improve classifiers so they can achieve good utility in classification for some purpose while decreasing discrimination that can happen against the protected groups (Dwork et al., 2012).

To avoid bias, it is not enough to just ignore protected attributes while making a decision because it is often possible to predict these attributes from other features. For example, omitting race from a mortgage application decision does not guarantee a lack of bias because zip code and other features are often correlated with race. To achieve fairness, many approaches aim for *demographic parity*, which is when the results achieve statistical parity among the protected and non-protected groups. However, this requirement generally causes a decrease in utility. In (Hardt et al. 2016), authors proposed two new ways to measure fairness instead of demographic parity, *equalized odds* and *equal opportunity*. The goals for these two approaches are achieving fairness

and proposing more accurate classifiers. The researchers developed a supervised learning approach with respect to protected attributes that satisfied equal opportunity. They evaluated their work using ROC (Receiver Operator Characteristic) curve which is a way to their results to demographic parity. They found that using equal opportunity as the fairness criteria led to the creation of a more accurate classifier that exhibited a lower utility loss than when demographic parity was used as the criteria. To avoid unfairness decisions in some classifiers, (Zafar et al. 2017) reiterate two goals of fairness from (Hardt et al. 2016) and add one more type which is *disparate mistreatment* when the sensitive attribute is associated with a higher misclassification rate. They introduce a new metric to measure disparate mistreatment and develop a method to train decision boundary-based classifiers, e.g., logistic regression classifiers, to avoid disparate mistreatment. They evaluated their work by applying it on real and synthetic datasets and comparing their results with (Hardt et al. 2016), demonstrating that their proposed method can decrease disparate mistreatment with only a small accuracy loss. Some other researchers tried to enhance fairness by training machine learning models without knowing the protected group memberships. In particular, (Lahoti et. al, 2020) proposed an Adversarially Reweighted Learning (ARL) approach to improve the utility for the least represented protected groups when they train the model. During the training stage, they rely more on the non-protected features and task labels to identify unfair biases and train their model to improve fairness. Their solution outperformed state-of-the-art alternatives across a variety of datasets.

2.4.3 Fairness in Ranked Outputs

Ranking the outputs has become very common these days with online systems to display results to the user such as ranking books in libraries, jobs opportunities, products and opinions. Ranking methods use demographic, behavioral, or other features to produce a ranked output that

represents the relative quality of individuals. Today, ranking systems have responsibilities not only to the users but also to the items they ranked, specifically if they ranked people for loans or job seeking, for example (Singh & Joachims, 2018). Because unfair ranking methods can lead to unfair decisions when choosing items, (Yang & Stoyanovich 2017) focus on developing new metrics to measure the amount of bias present in ranked outputs. They proposed three metrics to measure bias, i.e., lack of statistical parity, in ranked outputs. They generated several synthetic ranked datasets that ranked people depending on their income using gender as their protected attribute. Finally, they present an optimization-based method to improve the fairness of ranked methods. Their approach learns a mapping between data and outcomes that preserves accuracy while improving fairness. Other scholars developed algorithms to train Learning To Rank (LTR) models fairly. For instance, (Bower A. et. al, 2021) proposed a method to train a model individually and ensure that the membership of minority group members and majority group members are similar. They utilize the definition of individual fairness from supervised learning in ML to design optimal transport-based regularizer. They showed that while assigning exposure fairness for the group might not lead to fair LTR models for individuals, improving individual fairness can assign exposure fairness for the group.

(Zehlike et al. 2017) also address the problem of improving fairness in the top-K ranking problem over a single binary type attribute when selecting a subset of candidates from a large pool. Their goal is to ensure that the ranking results at any cut-off *k* include a proportion of individuals from a protected group that exceeds a specific threshold. Their two conditions to maximize utility ensure that every member in the top-k is more qualified than the others outside the top-k (or the difference in qualifications is small) and, for every pair in the top-k, the higher-ranked candidate is more qualified than the one below. They consider one protected attribute

(either gender or race) in different data sets for validation of their approach. They apply normalized Discounted Cumulative Gain (nDCG) to calculate the quality of the ranking results. Another study by (Singh and Joachims 2018) worked on adding a reasonable level of fairness to algorithms that produced ranked outputs of people and items while maximizing the utility based on addressing the three fairness constraints mentioned in (Zafar et al., 2017). They used the DCG (Discounted Cumulative Gain) metric and Disparate Treatment Ratio (DTR) to evaluate their work. After applying their methods on job seeker and news recommendation datasets, they concluded that their methods improved the level of fairness with a little drop in DCG compared to the rankings without fairness. In their previous work, (Zehlike, 2017) developed a post-processing approach to increase fairness based on utility-ranked results for protected and non-protected groups. More recently, (Zehlike & Castillo, 2018) proposed an algorithm to provide fairness using an in-processing approach based on a learning-to-rank framework that addresses discrimination and inequality of opportunity. They compare their in-processing approach to other learning-to-rank methods while applying them on two datasets and the results showed that optimizing for fairer results does not necessarily decrease relevance and can, in heavily biased data sets, actually improve relevance.

Research Plan

3.1 System Overview

Our research focuses on designing algorithms to improve the author diversity of papers accepted to a conference while minimizing the drop in the overall quality of the accepted papers. Our approach is based on building a profile for each paper that reflects the paper's overall quality and also models the diversity of the paper authors. This profile is then used by the recommender module to select papers for inclusion in the conference. Figure (3.1) shows the framework of our system.

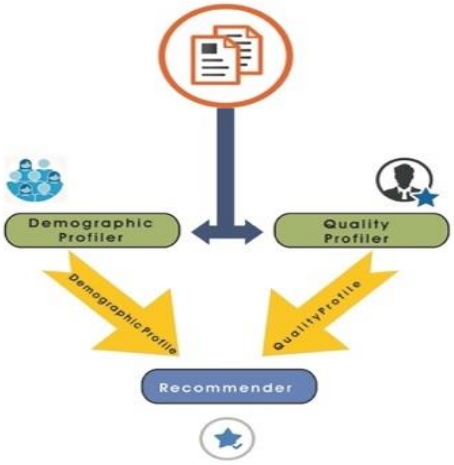

Figure (3.1) System framework

By considering author diversity as part of the selection criteria, it is our belief that acceptable papers can be chosen with little negative impact on conference quality but a possible large impact on conference diversity. We will model papers based on their quality as well as the demographic attributes of their authors. Then, we will develop algorithms that consider both aspects of a paper, its quality and the authors' ability to increase the conference diversity, when recommending papers to be selected by the conference. This system can be applied during

conferences reviews, journal paper reviews, while awarding grant proposals, and other tasks within academia.

3.2 Demographic Profile Construction based on Protected and Nonprotected groups

To build a diverse community of accepted authors for a given conference, we must first build a demographic profile for each paper by modeling the demographic features for the paper's authors so that this information is available during paper selection. Some demographic features are protected attributes, e.g., gender, race, age, nationality, that qualify for special protection from discrimination by law (Inc. US Legal, n.d.). We use these since they have been shown to be common sources of bias. In this section, we will describe how we collect the demographic features for each author in our papers pool and then how we build the paper profile. The main process to extract the data is scraping the available information using their scholar or home pages.

3.2.1 Data Extraction

For a given paper, our goal is to extract five demographic features that are Gender, Race, University Rank, Career Stage, and Geolocation for its author(s). Each feature is mapped to a Boolean value, either 1 (true) or 0 (false) based on that paper's author(s) membership in the protected group. We then extended our approach beyond current approaches by modeling demographics with continuous-valued features (each feature is mapped to a value between 0 and 1). Table (3.1) outlines the protected and non-protected categories for each of our demographic features.

Most papers have more than one author, so we first build a profile for each of the paper's authors and combine them to create a profile for the paper as a whole. First, we will discuss each demographic attribute we study.

Table (3.1) Demographic features categories

Features	Category
Gender	Female / Male
Ethnicity	Non-White / White
Geo-Location	Developing /Developed (by country)
	EPSCoR / Non-EPSCoR (by state in USA)
Career Stage	Junior / Senior
University Rank	Less than or equal mean/ more than mean

Gender: To gather information about an author's gender, we use the NamSor API v2, a data mining tool that uses a person's first and last names from different languages, alphabets, countries, and regions to infer their gender. The software processed more than 4 billion names with high precision and recall which are 98.41% and 99.28% respectively. The tool returns a value between -1 and +1 indicates that the name is male if it is close to +1 and female if it is close to -1. The accuracy of gender prediction using this tool is close to 99% (blog, NamSor, 2018).

After collecting each author's gender, we map females to 1 since they are the protected group and males to 0. To calculate the continuous value for gender, we map females and males to the complement of their participation in computer science. Women are considered a protected group since they make up only 27% of professionals in the computer science field (Khan, Robbins and Okrent, 2020).

Ethnicity: To predict ethnicity, we again use the NamSor tool, a web API that is used to predict ethnicity from the first and last names with the limitation of 500 names/month. It returns ethnicity as one of five values: {White, Black, Hispanic, Asian, other} (blog, NamSor, 2019). Non-whites

are considered a protected group since they make up less than 40% of professionals in the computer science field (Khan, Robbins and Okrent, 2020).

The continuous values for Ethnicity were calculated by mapping each category to the complement of its proportion in the population of Computer Science professionals from (Zweben, and Bizot, 2018) (Computer, engineering, & science occupations, 2020). Whites comprise 70.46% of computer science professionals, so they are mapped to 0.2954. Similarly, Black, Asian, Hispanic and others are assigned to 0.9295, 0.8237, 0.9281, and 0.7400 respectively.

Career Stage: In order to extract the academic position for each author, we utilize the researcher's Google Scholar pages (Google Scholar,2020), a publicly open web search engine that consists of scholarly literature that includes the publications, citations number, and h-index score of each researcher. For those who do not have a Google Scholar page, we extract their information from their homepages manually. Researchers whose primary appointment is within industry are omitted from our data set.

The results are then mapped to Boolean values, 0 if they are a senior researcher (non-protected) and 1 if they are a junior researcher (protected). senior researchers are defined as {Distinguished Professor, Professor, Associate Professor} and junior researchers are defined as {Assistant Professor, Postdoc, Student}. To calculate the continuous values for this feature, we will map to six values equally distributed between [0, .., 1.0] in increasing order by rank, i.e., Distinguished Professor: 0/5 = 0.0; Professor: 1/5 = 0.2; ...; Student: 5/5 = 1.0. Table (3.2) shows the values for each category.

Table 3.2 Career Stage Weight Allocation

Position	Weight
Distinguished Professor	0.17
Professor	0.33
Associated Professor	0.50
Assistant Professor or Lecturer	0.67
Post-Doctoral or Research Fellow	0.83
Graduate Student	1.0

University Rank: Collecting this feature is done by extracting the institution's name from Google Scholar home page for the author (Google Scholar,2020) or their home pages. We then use the World University Rankings obtained from Times Higher Education magazine (Times Higher Education, 2020). These values range from 1 to 1001+ and the list includes around 1400 universities from 92 countries.

To assign the Boolean University Rank value, we use the median University rank to partition authors into low-rank (1) or high-rank institutions (0). To calculate the Continuous value, we normalize the raw value to get a value between 0.0 and 1.0. We normalize the University Rank value using formula (1).

$$R_C = \frac{U_r}{L_r} \qquad (1)$$

where U_r is the value of the university rank and L_r is the lowest university rank (1001). The higher the value we get, the lower the university rank.

Geolocation: We set the researcher's geolocation (country and state if inside the US) based on information extracted from their institution's home page using the university name that was extracted. If the author is working inside the United States, we extract the state name as well. We find the category of the country (developed or developing) by mapping the country to the tables of the developed and developing economies that offered by the UN (Nations, 2020).

Thus, the Geolocation Boolean value is assigned to 0 if the researcher is working in a developed country (non-protected group) and 1 if a developing country (protected group). For those who live in the US, we use the EPSCOR (Established Program to Simulate Competitive Research) (National Science Foundation, 2019) to map the Geolocation to Boolean values. EPSCoR states which obtain less federal grant funding are the protected group with the value 1 and non-EPSCoR states values are 0. To calculate the continuous value for the Geolocation, we use the complement values of Human Development Index (HDI) ranking (Human Development Report, n.d.). The values are ranging from 0.957 to 0.394 and table (3.3) shows a sample of these values.

Table (3.3) Countries HDI sample

Country	HDI
Norway	0.957
Ireland	0.955
Iceland	0.949
Germany	0.947
Sweden	0.945
Australia	0.944

H-index: In addition to the above features, we extract the h-index for each author so we can measure the conference utility. To implement that, we extract the h-index from each author's

Google Scholar page (Google Scholar, 2020). If the author doesn't have a scholar page, we obtain their h-index using Harzing's Publish or Perish tool. This software collects the author's publications and calculates the number of citations and impact metrics. One of them is the h-index for the scholar (Harzing, 2016).

To conclude, each researcher has a demographic profile consists of five features (gender, ethnicity, career stage, university rank, and geolocation). Each feature has a Boolean weight that represents whether or not the candidate is a member of the protected group for that feature and a continuous value to represent the complement of the proportion of each feature among computer science professionals. In addition, we collect the h-index for each researcher using either their Google Scholar profile or is calculated and we use it to evaluate the utility of each accepted papers list in our evaluation. Table (3.4) illustrates the raw data of the authors.

Table (3.4) Sample of the authors' raw data

Author Name	Gender	Ethnicity	Career Stage	University Name	URank	Country	State	h-index
Jessica Hammer	female	W_NL	Assistant Professor	Carnegie Mellon University	27	United states	Pennsylvania	12
Gillian M. McCarthy	female	W_NL	Lecturer	Victoria University of Wellington	501–600	New Zealand		2
Shrikanth Narayanan	male	A	Professor	University of Southern California	62	United States	California	89
David C. Atkins	male	W_NL	Research Professor	University of Washington	26	United States	Washington	58
Jason Ellis	male	W_NL	Professor	Northumbria University	351–400	United Kingdom		23

3.3 Paper Profile Formation

We construct the demographic profile for each paper by combining the demographic profiles for all of the paper authors. Recall that each author has either a Boolean value profile or a continuous value profile.

Boolean: Each author's profile is a vector of five Boolean features where 1 means a member of that protected group, 0 otherwise. The paper profile is created by doing a bit-wise OR on the paper's author profiles. Thus, the paper profile is 1 for a given demographic feature when any author is a member of that feature's protected group as shown below:

$$Paper\ Profile\ Vector\ = < 1, 0, 0, 1, 0>$$

This vector means that the paper has:

<female, white, senior professor, high university rank, developed country>

We considered summing the author profiles, but this would give preferential treatment to papers with more authors and normalizing the summed profile would penalize papers with many authors.

Continuous: Each author in the paper has a vector of five continuous-valued features that represent the complement of the proportion of each feature in the CS community as described previously. The paper's demographic profile is created by selecting the maximum value for each feature among the paper authors' profiles.

3.4 Paper Quality Profiler

There are several ways to measure a paper's quality such as the number of citations of the paper, the reputation of the editorial committee for the publication venue, or the publication venue's quality itself, often measured by Impact Factor (IF) (Bornmann and Daniel, 2009). The IF is a widely used, objective measure, although it is not accurate for new venues that contain high quality papers with few citations merely because they are new (Zhuang, Elmacioglu, Lee, and Giles, 2007). However, since the conferences in our planned dataset are all well-established, we use the IF as the basis of the quality profile for the papers in our research.

Given a paper, our first step is finding a source to extract the impact factor for the conference in which it was published. We extract the Impact Factor (IF) for each paper's conference from a collection of 960 computer science conferences and journals and their impact factors published by the Guide2Research website in 2019 (Guide2Research, 2019). The IF was calculated by using Google Scholar Metrics to find the conferences H5-Index which is the h-index for the published papers in the last 5 years. Then, the largest number h was designated as the conference IF (Guide2Research, 2019) (Google Scholar, n.d.).

3.5 Pool Distribution (Demographic Parity)

When applying our proposed methods as described below, we rely on reaching demographic parity during accomplishing our goal. This means that we select the papers with respect to each features' distribution in the pool. This approach is based on trying to achieve demographic parity, i.e., selecting papers such that the demographics of the accepted authors match those of the pool of candidates. To achieve this, we measure the proportion of participants for each feature in the pool and store them in a vector (PoolParity).

PoolParity = <GenderWt, EthnicityWt, CareerWt, UniversityWt, GeoWt >

where each weight is the number of authors from that protected group normalized by the number of authors in the pool.

3.6 Recommending the Papers

The next goal is maximizing the diversity of the conference by applying two different methods to select papers with respect to each features' distribution in the pool. These approaches are based on trying to achieve demographic parity, i.e., selecting papers such that the demographics of the accepted authors match those of the pool of candidates. Our overarching goal is to produce a list

of papers that rank diverse authors highly while minimizing any decrease in the quality of the recommended papers.

After creating the vector that represents the paper demographic profiles with their features (Boolean or continuous), we rank the papers according to their diversity scores calculated using a diversity score based on either the Boolean or continuous profiles. We compare approaches that recommend papers using an overall diversity score with one that considers each of the multiple facets of diversity separately.

3.6.1 Overall Diversity Method

After creating paper demographic profiles using their Boolean and Continuous features as described before in section (3), paper diversity scores (PDScore) are calculated using formula (2) on the features Boolean values:

$$PDScore = \sum_{i=1}^{5} f_i \qquad (2)$$

where f_i is the value for each paper's demographic feature (i.e., five features for each paper). From this equation, we can find the paper's diversity score. Our first method to choose a diverse list of papers considers two different queues. The quality queue (Qquality) which contains the papers ranked by the Impact Factor (IF) as described in Section 3. This gives preference to the papers ranked highest by the reviewers, in our case represented by papers that appeared in the most selective conference. The demographic queue (Qdemog) which contains the ranked papers by PDScore. If there are papers with the same PDScores, then we sort them based on their quality score as the paper that is already in the conference with a higher diversity score has the priority to get in the new list of papers. Next, we pick papers from the top of (Qdemog) until satisfying the pool demographic parity for each feature. Once a paper is selected, it is removed from (Qquality)

to avoid choosing the same paper repeatedly. After achieving demographic parity for each of the demographic features, the remaining papers are added from the quality queue in order to meet the number of papers desired by the conference. Thus, as long as there are sufficient candidates in the pool, we are guaranteed to meet or exceed demographic parity for each protected group.

Algorithm 1: Overall Diversity

```
Qquality, Qdemog ← Initialize two empty priority queues
PoolParity ← Initialize an empty vector
Qq ← insert the papers and sort them based on Quality-Scores
for each feature:
    PoolParity [feature] ← compute Demographic Parity
for each paper:
    PDScore ← compute paper diversity score
    add paper to Qdemog and order them using PDScore
    If 2 or more papers have same PDScore:
        Sort papers using Quality-Score
while PoolParity Not satisfied:
    Papers ← select a paper from top of Qdemog
    delete selected paper from Qquality
while # of conference papers not satisfied:
    Papers ← select a paper from top of Qquality
```

3.6.2 Multi-Faceted Diversity Method

The previous method selects papers based on the total diversity score for each paper. However, it does not guarantee that the selected authors from the protected groups are actually diverse. It might end up selecting papers that have high diversity scores but are all females from developing countries, for example, with no minority authors at all. To correct for this possibility, we extend the previous approach to consider multiple demographic queues, each ranked on a separate feature. First, we will create five ranked queues by sorting the papers using one demographic feature at a time to create five sorted queues, one per feature, in addition to the quality-ranked queue. We now have six queues total. Based on the pool demographics, we give the highest

priority to the rarest features in the pool first, so we create the accepted papers list by selecting papers from the queues whose features have the fewest candidates in the pool. We select from that queue until the demographic parity goal for that feature is achieved, then move to the next least-represented demographic group. Once a paper is selected, it is removed from all six queues to avoid choosing the same paper repeatedly. After satisfying demographic parity for all protected groups, the remaining papers are added in order from the quality queue. Again, as long as there are sufficient candidates in the pool, we are guaranteed to meet or exceed demographic parity for each protected group.

Algorithm 2: Multi-Faceted Diversity

```
FeatureName ← List of five queue names, one per feature
for each feature in FeatureName:
    DivQueue[feature] ← Initialize empty priority queue
QualityQueue ← Initialize an empty priority queue
PoolParity ← Initialize an empty vector
QualityQueue ← insert papers and sort by Quality-Score
for each feature in FeatureName:
    PoolParity [feature] ← compute Demographic Parity
for each paper:
    PDScore ← compute paper diversity score
for each feature in FeatureName:
    DivQueue[feature] ← add paper if this feature is 1
    Sort papers based on Quality-Score
    If 2 or more papers has the same Quality-Score:
        Sort papers using PDScore
while PoolParity  NOT empty:
    LowFeature ← min (PoolParity)
    while LowFeature Not reached demographic parity
        Papers ← select top DivQueue[LowFeature]
        delete selected paper from QualityQueue
delete LowFeature from DParity
while # of conference papers not satisfied:
    Papers ← select a paper from top of QualityQueue
```

Experiment and Evaluation

In the previous section, we presented several fair methods that recommend a set of papers to be selected for publication in a conference based on their authors' demographic profiles combined with the paper review quality scores. Our approaches produce ranked lists of papers that represent the papers selected to maximize quality (the baseline), maximize diversity (overall or for each feature), or produce demographic parity. We evaluate the efficacy and effectiveness of our algorithms using *Diversity Gain* (D_G) overall and per feature, Utility Loss (UL_i), and Demographic Similarity (DemographicSimilarity) produced by each algorithm. We first introduce our datasets baseline, describe the metrics and baseline, and then we p the evaluation of our approaches.

4.1 Datasets

For our driving problem, we focus on selecting papers for a high impact computer science conference from a pool of papers that vary in quality and demographics. To create pools of candidate papers that simulate the papers submitted to a conference, we select a trio of conferences based on several criteria: 1) the conferences should publish papers on related topics; 2) the conferences should have varying levels of impact {very high, high and medium} mimicking submitted papers reviewed as high accept, accept, borderline accept; 3) the conferences should have a reasonably large number of accepted papers and authors. Based on these criteria, we selected SIGCHI (The ACM Conference on Human Factors in Computing Systems), DIS (The ACM conference on Designing Interactive Systems), and IUI (The ACM Conference where the Human-Computer Interaction (HCI) community meets the Artificial Intelligence community). The papers published in SIGCHI represent papers rated highly acceptable by SIGCHI reviewers, DIS papers represent papers rated acceptable by SIGCHI

reviewers, and IUI papers represent papers rated borderline acceptable. Excluding authors from industry, we create a dataset for each conference that contains the accepted papers and their authors (see Table 4.1). This dataset contains 592 papers with 813 authors for which we demographic profiles. We will expand this work to other conferences in the future.

Table (4.1): Composition of Our Dataset.

Dataset	Accepted Papers	Authors	Impact Factor
SIGCHI17	351	435	87
DIS17	114	231	33
IUI17	64	147	27

The demographic distribution of the authors in each conference is summarized in Figure (4.1). These clearly illustrate each of the conferences had few authors from most of the protected groups with the lowest participation in the highest impact conference, SIGCHI, with gender being an exception. As an example, SIGCHI 2017 had only 8.28% non-white authors, DIS 2017's authors were only 16.45% non-white, and IUI 2017 had 27.21% non-white. Similarly, authors from developing countries dominate with 6.44% of SIGCHI 2017 authors, 6.93% of DIS 2017 authors, and 14.29% of IUI 2017 authors being from developing countries.

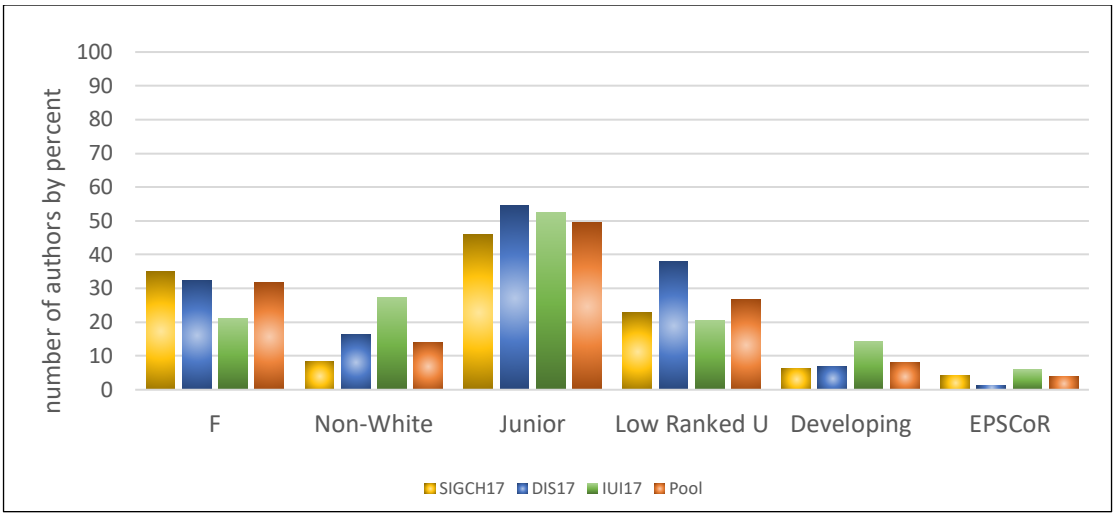

Figure (4.1): Protected Group Membership of Authors for Three Current Conferences

We define demographic parity as the participation rate for each of our demographic features in the pool created by combining the authors of all three conferences. Based on the 813 authors in our dataset, Table (4.2) presents the average participation in the pool for each feature and thus the demographic parity that is our goal.

Table (4.2): Demographic Participation from protected groups in Three Current Conferences

	Gender	Ethnicity	Career Stage	U Rank	Geolocation
SIGCHI (Baseline)	45.01%	7.69%	52.14%	25.64%	8.26%
DIS	57.89%	31.58%	72.81%	55.26%	11.40%
IUI	39.06%	56.25%	76.56%	28.13%	26.56%
Average	47.07%	18.71%	59.55%	32.33%	11.15%

4.2 Baseline and Metrics

Baseline. Our baseline is the original list of papers that were chosen by the program committee for SIGCHI 2017 and were represented in the venue. As shown in Table 2, the distribution of the

protected groups in our baseline is: 45.01% female, 7.69% non-white, 52.14% junior professors, 25.64% authors from low ranked universities and 8.26 authors from developing countries.

Metrics. Our algorithms attempt to generate a more diverse group of paper authors. We evaluate their effectiveness by calculating Diversity Gain (D_G) of our proposed set of papers versus the baseline:

$$D_G = \frac{\sum_{i=1}^{n} MIN\ (100, \rho_{G_i})}{n} \qquad (3)$$

where ρ_{G_i} is the relative percentage gain for each feature versus the baseline, divided by the total number of features n. Each feature's diversity gain is capped at a maximum value of 100 to prevent a large gain in a single feature dominating the value.

By choosing to maximize diversity, it is likely that the quality of the resulting papers will be slightly lower. To measure this drop in quality, we use the average h-index of the paper authors and compute the utility loss (UL_i) for each proposed list of papers using the following formula:

$$UL_i = \frac{U_b - U_{P_j}}{U_b} * 100 \qquad (4)$$

where U_{P_i} is the utility of the proposed papers for conference i and U_b is the utility of the baseline. We then compute the utility savings (Y_i) of papers for conference i relative to the baseline as follows:

$$Y_i = 100 - UL_i \qquad (5)$$

We compute the F measure (Jardine, 1971) to examine the ability of our algorithms to balance diversity gain and utility savings:

$$F = 2 * \frac{D_G * Y_i}{D_G + Y_i} \qquad (6)$$

In order to measure how far away from demographic parity our results are, we calculate the Euclidean Distance (Draisma, et al., 2014) between our selected papers and the pool:

$$\text{DemographicDistance} = \sqrt{\sum_{i=1}^{5}(F1_i - F2_i)^2} \qquad (7)$$

where F1 is the participation of each feature in the proposed list of papers to select and F2 is the feature's participation in the pool. Finally, we normalized the distance values to obtain the similarity percentages between our results and the pool as shown in the formula below:

$$\text{DemographicSimilarity} = 1 - \frac{\text{DemographicDistance}}{\text{MaxD}} \qquad (8)$$

where MaxD is the largest possible distance between two vectors in our feature space.

To summarize the ability of the methods to balance the competing demands of increasing demographic parity and saving utility, we again apply the F measure using formula 6 calculated using DemographicSimilarity and Y_i.

4.3 Results

Our recommender system produces ranked list(s) from which we select to form the accepted papers list with the overarching goal of increasing the diversity in the papers. Both methods reported here select papers from a quality sorted queue and one or more demographic queue(s). Whenever there are ties in a demographic queue, those papers are sorted by their quality score.

4.4 Comparison with the Baseline

We report the differences between the accepted papers in SIGCHI 2017 and the accepted papers produced by the recommender system described in Section 4 using Boolean and Continuous

profiles. Looking at Figure (4.2), we can see that all algorithms succeeded in increasing the diversity in the recommended papers for acceptance across all demographic groups when using the Boolean profiles. However, it is clear that the Overall Diversity method produced the highest diversity in all the protected groups.

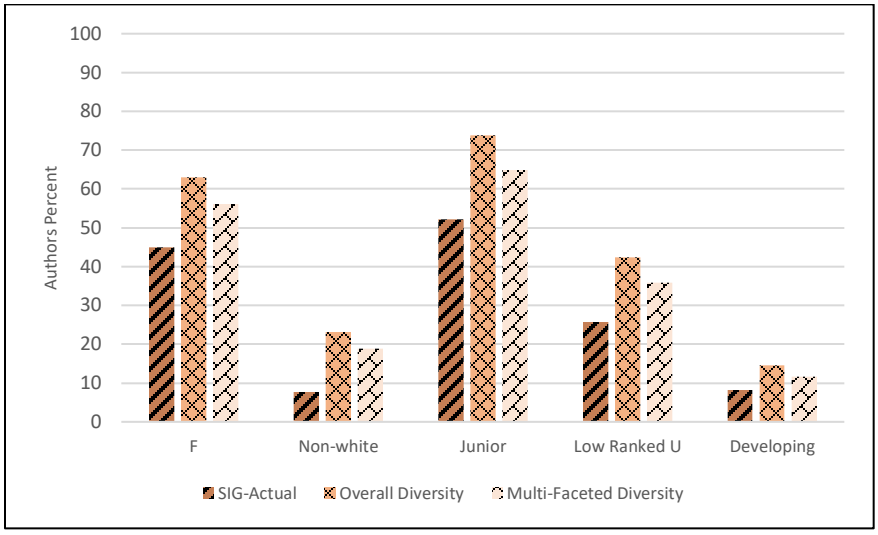

Figure (4.2): Improvement in Protected Group Participation between the SIGCHI2017 and our Paper Recommendation Algorithms when using Boolean Profiles.

Figure (4.3) represents protected group participation with the Continuous profiles to apply our proposed recommendation algorithms. We can see that all algorithms succeeded in increasing the diversity in the recommended papers for acceptance across all demographic groups. With the Continuous profile, the Overall Diversity and Multi-Faceted Diversity methods are essentially tied.

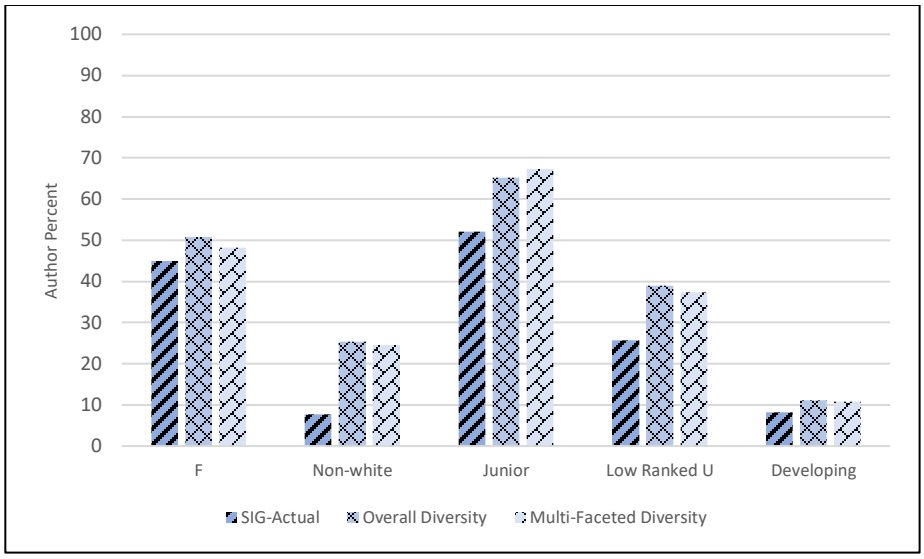

Figure (4.3): Improvement in Protected Group Participation between the SIGCHI2017 and our Paper Recommendation Algorithms when using Continuous Profiles.

Table (4.3) compares the participation of the protected groups between the actual accepted papers for SIGCHI with the accepted papers proposed by our two algorithms, and demographic parity based on the participation of the protected groups in the pool of authors in our dataset. We can see that all algorithms increase the diversity of authors across all protected groups for both the Boolean and continuous profiles. With the exception of Junior researchers with a continuous profile, the Overall Diversity algorithm increases participation among the protected groups more than the Multifaceted Diversity algorithm across all demographics. With the same exception, the Boolean profile also increases diversity more than the continuous profile. As expected, these diversity-based recommendation methods overcorrected for bias by including more authors from the protected groups proportionally than in the pool as a whole.

Table (4.3): Protected Group Participation for the recommender algorithms using Boolean and Continuous profiles.

Feature	SIGCHI	Overall Divers (Bool)	Overall Divers (Cont.)	Multi-Faceted Divers (Bool)	Multi-Faceted Divers (Cont.)	Pool
Female	45.01%	62.96%	50.71%	56.13%	48.15%	47.07%
Non-White	7.69%	23.08%	25.36%	18.80%	24.50%	18.71%
Junior	52.14%	73.79%	65.24%	64.96%	67.24%	59.55%
Low Ranked Univer.	25.64%	42.45%	39.03%	35.90%	37.32%	32.33%
Develop Country	8.26%	14.53%	11.11%	11.68%	10.83%	11.15%

The recommended papers are a mix of papers from the three conferences in our datasets in different proportions as described in Table (4.4). The Multi-Faceted Diversity method selects the highest proportion of the recommended papers, 85.8% (Boolean) and 78.06% (Continuous), from the actual SIGCHI papers, but Overall Diversity also selects the majority of its papers, 75.5% and 62.11%, from the original SIGCHI selected papers. We further observe that both algorithms selected the majority of papers from the demographic queue(s) with only a few from the quality-sorted queue. The Overall Diversity method selected 67.24% (Boolean) and 66.67% (Continuous) of its accepted papers from the demographic queue and only 32.76% (Boolean) and 33.33% (Continuous) from the quality queue. In contrast, the Multi-Faceted Diversity method selected nearly all of its accepted papers, 92.88%, from one of the five demographic queues, and only 7.12% from the quality queue.

Table (4.4): Proportion of Recommended Papers from each Conference.

	Overall Diversity (Bool)	Overall Diversity (Contin)	Multi-Faceted Diversity (Bool)	Multi-Faceted Diversity (Contin)
SIGCHI	265 (75.5%)	218 (62.11%)	301 (85.8%)	274 (78.06%)
DIS	59 (16.8%)	87 (24.79%)	47 (13.4%)	61 (17.38%)
IUI	27 (7.7%)	46 (13.11%)	3 (0.9%)	16 (4.56%)
Total # of papers	351	351	351	351

We also compare the performance of our algorithms with respect to the quality of the resulting accepted papers. Table (4.5) summarizes the diversity gain (D_G), Utility Savings (Y_i), and F scores for the accepted papers proposed by each algorithm when using the Boolean profiles. Both methods obtained Diversity Gains of over 45% for the proposed set of accepted papers, with the biggest gain occurring with the Overall Diversity algorithm. The gains in diversity occur with Utility Savings of 93.47% for the Overall Diversity algorithm versus 97.52% for the Multi-Faceted Diversity algorithm. Based on these results, we conclude that the Overall Diversity algorithm outperforms the Multi-Faceted Diversity algorithm with minimal utility loss.

Table (4.5): Diversity gain and utility savings for the overall diversity and multi-faceted diversity algorithms versus the Baseline for Boolean profiles.

Method	D_G	Y_i	F-score
Overall Diversity	64.58%	93.47%	76.39
Multi-Faceted Diversity	46.00%	97.52%	62.51

Table (4.6) summarizes the diversity gain (D_G), Utility Savings (Y_i), and F scores for the accepted papers proposed by each algorithm when we use the continuous profiles. Both methods increased the author diversity in the proposed set of accepted papers by more than 40% with a Utility saving of 102.49 % versus the actual SIGIR 2017 accepted papers. This means that by considering author demographics and aiming for demographic parity, the quality of the selected papers actually increased.

Table (4.6): diversity gain and utility savings for the overall diversity and multi-faceted diversity algorithms versus the Baseline using Continuous values.

Method	D_G	Y_i	F-score
Overall Diversity	44.90%	102.49%	62.44
Multi-Faceted Diversity	42.50%	102.45%	60.08

Diversity-based algorithms may overcorrect and result in reverse discrimination, or the diversity gains may all be in one subgroup while other underrepresented populations are ignored. Tables (4.7) and (4.8) show the results when evaluating our algorithms' ability to achieve demographic parity with Boolean and Continuous features, respectively. We observe that, based on this criteria, the Multifaceted Diversity algorithm produces results closest to Demographic Parity, with 95.01% similarity to the pool and a utility loss of just 2.48% when using Boolean profiles.

Table (4.7): Demographic parity similarity and utility savings for the overall diversity and multi-faceted diversity algorithms versus the baseline (Boolean).

Method	Demographic Similarity	Y_i	F-score
Overall Diversity	89.15%	93.47%	91.26
Multi-Faceted Diversity	95.01%	97.52%	96.24

We further observe that the Multi-Faceted method produces even better Demographic Parity of 95.12% when using continuous-valued features and actually results in a 2.45% increase in utility. This means that, by considering author diversity and aiming for demographic parity when selecting papers, the quality of the papers accepted to the conference could actually be improved.

Table (4.8): Demographic parity similarity and utility savings for the overall diversity and multi-faceted diversity algorithms versus the baseline (Continuous).

Method	Demographic Similarity	Y_i	F-score
Overall Diversity	94.80%	102.49%	98.27
Multi-Faceted Diversity	95.12%	102.45%	98.44

4.5 Discussion

In the previous sections, we discuss the evaluation for our approaches by comparing them to the baseline using Diversity Gain(D_G), Utility Saving (Y_i), F-score metrics, and DemographicSimilarity. We compared two algorithms for our recommender system using Boolean and continuous weight features in the demographic profiles.

4.5.1 Diversity Gain Comparison

Table (4.9) summarizes the results for all experiment with each algorithm on both profile weights, evaluated using Diversity Gain, Utility loss, and the F-measure. From this we can see that the Overall Diversity method with Boolean profiles maximized diversity gain and the Overall Diversity method with continuous weights minimizes the utility loss. However, the Overall Diversity method with Boolean weights produces the highest F-score that balances these.

Table (4.9): Diversity Gain and Utility Saving with Boolean and Continuous weights profiles

	Profile	D_G	Y_i	F-score
Overall Diversity	Boolean	**56.33**	93.47	**76.39**
	Continuous	44.90	**102.49**	62.44
	Average	50.61	97.98	69.41
Multi-Faceted Diversity	Boolean	**46.00**	97.52	**62.51**
	Continuous	42.50	**102.45**	60.08
	Average	44.25	99.98	61.29
Average	Boolean	51.16	95.49	69.45
	Continuous	43.70	102.47	61.26

Overall Diversity versus Multi-faceted Diversity method

Averaged over both feature weights, the Overall Diversity measure produces higher diversity gain, 50.61% versus 44.25%. The Multi-Faceted Diversity method produces a smaller drop in utility on average, 0.02% versus 2.02%. However, when balancing diversity gain with utility

drop, the Overall Diversity method still produces better results with an average F-score of 69.41versus 61.29 for the Multi-Faceted approach.

Boolean versus Continuous feature weights

Averaged over both methods, the Boolean profiles produces higher diversity gain, 51.16% versus 43.70% for Continuous. The Continuous profiles produces a little gain in utility on average, 2.47% versus 4.51% drop in utility for Boolean profiles. However, when balancing diversity gain with utility drop, the Boolean profiles still produces better results with an average F-score of 69.45 versus 61.26 for Continuous profiles.

4.5.2 Demographic Parity Comparison

Table (4.10) summarizes the results for all experiment with each algorithm on both profile weights, evaluated using Demographic Similarity, Utility loss, and the F-measure. From this we can see that the Multi-Faceted method with Continuous profiles maximized Demographic Similarity and the Overall Diversity method with continuous weights minimized the loss in utility. In fact, it actually produced a gain in utility! The Multi-Faceted method with Continuous weights also produces the highest F-score that balances these.

Table (4.10): Demographic Similarity with Boolean and Continuous weights profiles

	Profile	Demographic Similarity	Y_i	F-score
Overall Diversity	Boolean	89.15%	93.47%	91.26
	Continuous	**94.80%**	**102.49%**	**98.27**
	Average	91.97	97.98	94.76
Multi-Faceted Diversity	Boolean	95.01%	97.52%	96.24
	Continuous	**95.12%**	**102.45%**	**98.44**
	Average	95.06	99.98	97.34
Average	Boolean	92.08	95.49	93.75
	Continuous	95	102.47	98.35

Overall Diversity versus Multi-faceted Diversity method

Averaged over both feature weights, the Multi-Faceted Diversity measure produces higher Demographic Similarity, 95.06% versus 91.97%. The Multi-Faceted Diversity method produces a smaller drop in utility on average, 0.02% versus 2.02%. Also, when balancing diversity gain with utility drop, the Multi-Faceted Diversity method still produces better results with an average F-score of 97.34% versus 94.76% for the Overall Diversity approach.

Boolean versus Continuous feature weights

Averaged over both methods, the Continuous profiles produces higher Demographic Similarity, 95% versus 92.08% for Boolean profiles. The Continuous profiles produces a little gain in utility on average, 2.47% versus 4.51% drop in utility for Boolean profiles. However, when balancing diversity gain with utility drop, the Continuous profiles still produces better results with an average F-score of 98.35versus 93.75 for Boolean profiles.

Conclusion

5.1 Summary

Promoting diversity is a very important goal in many areas of life, business, and academia. There is evidence that the review process for conference papers still fosters discrimination against minority groups that leads to their exclusion from publication and networking opportunities. Increasing diversity in conference presenters could positively impact conference attendee diversity and provide new ideas and new opportunities that arise from mixing people from different regions around the world. We present new recommendation algorithms that increase diversity when recommending papers for acceptance in conferences while minimizing any decrease in quality. Our methods promote diversity by considering multidimensional demographic author profiles as well as paper quality when recommending papers for publication in a conference. We model papers using quality based on impact factors and demographics, based on five Boolean and continuous features that model the authors' demographic attributes. Most previous work focuses on algorithms that guarantee fairness based on a single, Boolean feature, e.g., race, gender, or disability. In contrast, we consider gender, ethnicity, career stage, university rank and geolocation to profile the authors.

We demonstrate our approach using a dataset that includes authors whose papers were selected for presentation at conferences in Computer Science that vary in impact factor to mimic papers rated by reviewers at different levels of acceptability. The Overall Diversity method ranks the papers based on an overall diversity score whereas the Multi-Faceted Diversity method selects papers that fill the highest-priority demographic feature first. The resulting recommended papers were compared with the baseline (the actual accepted papers for SIGCHI 2017) in terms

of diversity gain and utility savings as measured by a decrease in the average h-index of the paper authors.

Our first goal was to build paper profiles based on Boolean and Continuous author demographics. We achieve this by web scraping the demographic feature for each author in the paper and then combine them to obtain the profile for each paper. This resulted in a dataset of 592 demographics and quality profiles for 813 academic authors from three conferences authors.

Our second goal was to maximize the diversity of the recommended authors. Our best method for this, the Overall Diversity method, increased diversity by 64.58% (using Boolean-valued features) with only a 6.53% drop in utility and 44.90% (using continuous-valued features) with 2.49% increase in utility.

Our third goal was to achieve demographic parity. Our best method for this, the Multi-Faceted Diversity method, produced results closest to demographic parity with more than 95% similarity to the pool. It achieved a 46% gain in diversity with only a 2.48% drop in utility with Boolean demographic profiles and a 42.50% gain in diversity with 2.45% increase in utility with continuous-valued demographic features. This last result demonstrates that increasing diversity does not necessarily come at a cost in utility.

5.2 Future Work

For the future, we will develop new algorithms that guarantee demographic parity to avoid overcorrection. We are currently working on other algorithms to correct the demographic parity overcorrection that resulted from the previous approaches in order to publish it in a journal. We are implementing two methods, one based on considering the demographics that has higher portions in the pool first and the other based on round robin techniques for choosing paper from the pool. Additionally, we will explore dynamic hill-climbing algorithms that adjust the

recommendation criteria after each paper selection. Finally, we will build a larger dataset by incorporating other trios of conferences and investigate the effectiveness of machine learning and deep learning techniques to improve the diversity for our papers.

Reference

Alhajraf, N. M., & Alasfour, A. M. (2014). The impact of demographic and academic characteristics on academic performance. *International Business Research*, *7*(4), 92.

Alkan, O., & Daly, E. (2020, June 15). *User profiling from reviews for accurate time-based recommendations*. arXiv.org. Retrieved April 29, 2022, from https://arxiv.org/abs/2006.08805

Andrews, L. Edmund. (2021, December). Stanford research reveals a hidden obstacle for women in Academia. *Stanford News Service*. Retrieved from https://news.stanford.edu/press-releases/2021/12/16/hidden-obstacle-women-academia/

Asudeh, A., Jagadish, H. V., Stoyanovich, J., & Das, G. (2019, June). Designing fair ranking schemes. In IJCAI Barak B. (2018). On double blind reviews in theory conferences. In Windows on Theory. Retrieved in May, 21, 2006. (pp. 1259-1276).

Barak, B. (2018). On double blind reviews in theory conferences. *Windows on Theory*. https://windowsontheory.org/2018/01/11/on-double-blind-reviews-in-theory-conferences/

Balog, K., & De Rijke, M. (2007, January). Determining Expert Profiles (With an Application to Expert Finding). In IJCAI (Vol. 7, pp. 2657-2662).

blog, NamSor. (2018). Understanding NamSor API precision for Gender inference. Inferring The World's Gender and Ethnic Diversity using Personal Names. https://namesorts.com/2018/01/31/understanding-namsor-api-precision-for-gender-inference/

blog, NamSor. (2019). NamSor US 'Race' / Ethnicity model helps estimate diversity in Chicago Police. Inferring The World's Gender and Ethnic Diversity using Personal Names. https://namesorts.com/2019/07/12/namsor-us-race-ethnicity-model-helps-estimate-diversity-in-chicago-police/

Bornmann, L., & Daniel, H. D. (2005). Selection of research fellowship recipients by committee peer review. Reliability, fairness and predictive validity of Board of Trustees' decisions. *Scientometrics*, *63*(2), 297-320.

Bornmann, L., & Daniel, H. D. (2009). The state of h index research: is the h index the ideal way to measure research performance?. *EMBO reports*, *10*(1), 2-6.

Bower, A., Eftekhari, H., Yurochkin, M., & Sun, Y. (2021). *Individually fair ranking*. ICLR 2021. Retrieved from https://arxiv.org/abs/2103.11023

Bowman, N. A., & Bastedo, M. N. (2011). Anchoring effects in world university rankings: exploring biases in reputation scores. *Higher Education*, *61*(4), 431-444.

Gabriel, D. (2017). Race, racism and resistance in British academia. In Rassismuskritik and Widerstandsformen (pp. 493-505). Springer VS, Wiesbaden.

Cannon, S., Reid, D. A., McFarlane, K., King, L., MacKenzie, L., Tadaki, M., and Koppes, M. (2018). Race and gender still an issue at academic conferences. *The Conversation*. https://theconversation.com/race-and-gender-still-an-issue-at-academic-conferences-92588

Chandrasekaran, K., Gauch, S., Lakkaraju, P., & Luong, H. P. (2008, July). Concept-based document recommendations for citeseer authors. In *International conference on adaptive hypermedia and adaptive web-based systems* (pp. 83-92). Springer, Berlin, Heidelberg.

Chappelow J. (2019), Demography, In Investopedia. Retrieved from https://www.investopedia.com/terms/d/demographics.asp

Cochran-Smith, M., & Zeichner, K. M. (Eds.). (2009). Studying teacher education: The report of the AERA panel on research and teacher education. Routled

Code.org. (2020)., Women computer science graduates finally surpass record set 17 years ago, but percentages lag behind. *Medium*. https://medium.com/@codeorg/women-computer-science-graduates-finally-surpass-record-set-17-years-ago-20a79a76275

Google. (2021). *Google 2021 diversity annual report*. 2021 Diversity Annual Report. Retrieved from https://static.googleusercontent.com/media/diversity.google/en//annual-report/static/pdfs/google_2021_diversity_annual_report.pdf?cachebust=2e13d07

Computer, engineering, & science occupations. Data USA. (2020). https://datausa.io/profile/soc/computer-engineering-science-occupations#:~:text=Race%20%26%20Ethnicity&text=64.2%25%20of%20Computer%2C%20engineering%2C,or%20ethnicity%20in%20this%20occupation.

Cox, A. R., & Montgomerie, R. (2019). The cases for and against double-blind reviews. *PeerJ*, *7*, e6702.

Deng, H., King, I., & Lyu, M. R. (2008, December). Formal models for expert finding on dblp bibliography data. In 2008 Eighth IEEE International Conference on Data Mining (pp.163-172) IEEE.

Draisma, J., Horobeţ, E., Ottaviani, G., Sturmfels, B., & Thomas, R. (2014). The Euclidean distance degree. *In proceedings of the 2014 Symposium of Symbolic-Numeric Computation (SNC'14). ACM.* (pp. 9-16). http://dx.doi.org/10.1145/2631948.2631951.

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

eeoc.gov. (n.d.). US Equal Employment Opportunity Commission. https://www.eeoc.gov/eeoc/

Falkenberg, L. J., & Soranno, P. A. (2018). Reviewing reviews: An evaluation of peer reviews of journal article submissions. *Limnology and Oceanography Bulletin*, *27*(1), 1-5.

Farnadi, G., Kouki, P., Thompson, S. K., Srinivasan, S., & Getoor, L. (2018). A fairness-aware hybrid recommender system. *arXiv preprint arXiv:1809.09030*.

Flaherty, C. (2016). More Faculty Diversity, Not on Tenure Track. *Inside Higher ED*. https://www.insidehighered.com/news/2016/08/22/study-finds-gains-faculty-diversity-not-tenure-track

Flaherty, C. (2018). When Journals Play Favorites. *Inside Higher ED.* https://www.insidehighered.com/news/2018/03/02/study-finds-evidence-institutional-favoritism-academic-publishing

Flaherty, C. (2019). Professors Still More Likely Than Students to Be White. *Inside Higher Ed*. https://www.insidehighered.com/quicktakes/2019/08/01/professors-still-more-likely-students-be-white

Galhotra, S., Brun, Y., & Meliou, A. (2017, August). Fairness testing: testing software for discrimination. In Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering (pp. 498- 510). ACM.

Gauch, S., Speretta, M., Chandramouli, A., & Micarelli, A. (2007). User profiles for personalized information access. *The adaptive web*, 54-89.

Google Scholar. 2020. *Google*. https://scholar.google.com/

Google Scholar (n.d.). https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng

Guide2Research. (2020). Top Computer Science Conferences. *Guide2Research.* http://www.guide2research.com/topconf/

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*.

Harzing, A. (2016). Publish or Perish. https://harzing.com/resources/publish-or-perish

Helmer, M., Schottdorf, M., Neef, A., & Battaglia, D. (2017). Gender bias in scholarly peer review. *Elife*, *6*, e21718.

Human Development Reports. (n.d.). Human Development Data Center. *Human Development Reports*. http://hdr.undp.org/en/data.

Inc. US Legal. (n.d.). Protected Group Member Law and Legal Definition.https://definitions.uslegal.com/p/protected-group-member/

Jacob, B. A., & Lefgren, L. (2011). The impact of research grant funding on scientific productivity. *Journal of public economics*, *95*(9-10), 1168-1177.

Jaschik, S. (2016). The Gender Factor in Conference Presentations. *Inside Higher ED*. https://www.insidehighered.com/news/2016/09/07/new-study-suggests- continued-bias-academic-conference-panel-selections

Jefferson, T., Wager, E., & Davidoff, F. (2002). Measuring the quality of editorial peer review. *Jama*, *287*(21), 2786-2790.

Jiang, Z., Han, X., Fan, C., Yang, F., Mostafavi, A., & Hu, X. (2021, September). Generalized demographic parity for group fairness. In *International Conference on Learning Representations*.

Jones, T. M., Fanson, K. V., Lanfear, R., Symonds, M. R., & Higgie, M. (2014). Gender differences in conference presentations: a consequence of self-selection?. *PeerJ*, *2*, e627.

Kamishima, T., Akaho, S., & Sakuma, J. (2011, December). Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops* (pp.643-650). IEEE.

Kanoje, S., Girase, S., & Mukhopadhyay, D. (2015). User profiling trends, techniques and applications. arXiv preprint arXiv:1503.07474.

Khan, B.,Robbins, C., and Okrent, A. (2020). Science and Engineering Indicator. *NSF*. https://ncses.nsf.gov/pubs/nsb20198/demographic-trends-of-the-s-e-workforce

Kodakateri Pudhiyaveetil, A., Gauch, S., Luong, H., & Eno, J. (2009, October). Conceptual recommender system for CiteSeerX. In Proceedings of the third ACM conference on Recommender systems (pp. 241-244). ACM.

Labille, K., Gauch, S., Joseph, A. S., Bogers, T., & Koolen, M. (2015). Conceptual Impact-Based Recommender System for CiteSeerx. In *CBRecSys@ RecSys* (pp. 50-53).

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., & Chi, E. (2020). *Fairness without Demographics through Adversarially Reweighted Learning*. NeurIPS 2020. http://alexbeutel.com/papers/NeurIPS-2020-fairness-without-demographics.pdf

Lemire, A. D. (2020, December 3). *Double-blind peer review is a bad idea*. Daniel Lemire's blog. https://lemire.me/blog/2020/11/19/double-blind-peer-review-is-a-bad-idea/.

Lerback, J., & Hanson, B. (2017). Journals invite too few women to referee. *Nature News*, *541*(7638), 455.

Long, C., Wong, R. C. W., Peng, Y., & Ye, L. (2013, December). On good and fair paper-reviewer assignment. In *2013 IEEE 13th international conference on data mining* (pp. 1145-1150). IEEE.

Manjoo, F. (2014). Exposing hidden bias at Google. The New York Times, 24.

Murray, D., Siler, K., Larivière, V., Chan, W. M., Collings, A. M., Raymond, J., & Sugimoto, C. R. (2019). Gender and international diversity improves equity in peer review. *BioRxiv*, 400515.

National Science Foundation. (2019). Established Program to Stimulate Competitive Research (EPSCoR). *NSF website*. https://www.nsf.gov/od/oia/programs/epscor/nsf_oiia_epscor_EPSCoRstatewebsites.jsp

Nations, U. (2020). The World Economic Situation and Prospects. 2020. Acessado em, 20. https://www.un.org/development/desa/dpad/wp-content/uploads/sites/45/WESP2020_Annex.pdf

Palus, S. (2015). Is Double-Blind Review Better?. *American Physical Society.* https://www.aps.org/publications/apsnews/201507/double-blind.cfm

Pannucci, C. J., & Wilkins, E. G. (2010). Identifying and avoiding bias in research. *Plastic and reconstructive surgery*, *126*(2), 619.

Protected group. (2020). Wikipedia. https://en.wikipedia.org/wiki/Protected_group

Sankar, P. (2015). The pervasive bias against female computer science majors. *Fortune Magazine*. https://fortune.com/2015/04/20/the-pervasive-bias-against-female-computer-science-majors/

Santamaría, L., & Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, *4*, e156.

Strømgren C. (2016). Genderize io. Retrieved from https://genderize.io/

B. Sateli, F. Löffler, B. König-Ries and R. & Witte, "ScholarLens: Extracting competences from research publications for the automatic generation of semantic user profiles," in PeerJ Computer Science, 3, e121., 2017.

Saxena, A. (2014). Workforce diversity: A key to improve productivity. *Procedia Economics and Finance*, *11*, 76-85.

SIGCHI. (2019). Diversity of the Program Committee for CHI 2020. Retrieved from https://chi2020.acm.org/blog/diversity-of-the- program-committee-for-chi-2020/

Sikdar, S., Marsili, M., Ganguly, N., & Mukherjee, A. (2016, October). Anomalies in the peer-review system: A case study of the journal of High Energy Physics. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 2245-2250).

Singh, A., & Joachims, T. (2018, July). Fairness of exposure in rankings. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2219-2228).

Shu, K., Zhou, X., Wang, S., Zafarani, R., & Liu, H. (2019). The Role of User Profile for Fake News Detection. arXiv preprint arXiv:1904.13355.

Stelmakh, I., Shah, N., & Singh, A. (2021). PeerReview4All: Fair and accurate reviewer assignment in peer review. *Journal of Machine Learning Research*, *22*(163), 1-66.

Sugarman, D. B., Nation, M., Yuan, N. P., Kuperminc, G. P., Hassoun Ayoub, L., & Hamby, S. (2018). Hate and violence: Addressing discrimination based on race, ethnicity, religion, sexual orientation, and gender identity. *Psychology of violence*, *8*(6), 649.

Times Higher Education (THE). (2020). World University Rankings. https://www.timeshighereducation.com/world-university-rankings/2020/world-ranking.

The New York Times Company. (2021). Building a Culture That Works for All of Us Retrieved from https://www.nytco.com/company/diversity-and-inclusion/a-call-to-action/

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017). Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, *114*(48), 12708-12713.

Trajkova, J., & Gauch, S. (2004, April). Improving ontology-based user profiles. In Coupling approaches, coupling media and coupling languages for information retrieval (pp. 380-390). ACM.

Sugiyama, K., & Kan, M. Y. (2010, June). Scholarly paper recommendation via user's recent research interests. In Proceedings of the 10th annual joint conference on Digital libraries (pp. 29-38). ACM.

Williams, C. L. (1992). The glass escalator: Hidden advantages for men in the "female" professions. *Social problems*, *39*(3), 253-267.

Women in Computer Science. Women in Computer Science: Getting Involved in STEM. *ComputerScience.org*. (2021, May 5). https://www.computerscience.org/resources/women-in-computer-science/.

Wu, Y. (2009). NSF's Experimental Program to Stimulate Competitive Research (EPSCoR): Subsidizing academic research or state budgets?. *Journal of Policy Analysis and Management*, *28*(3), 479-495.

Yang, K., & Stoyanovich, J. (2017, June). Measuring fairness in ranked outputs. In Proceedings of the 29th International Conference on Scientific and Statistical Database Management (pp. 1-6).

Ye, J., Han, S., Hu, Y., Coskun, B., Liu, M., Qin, H., & Skiena, S. (2017, November). Nationality classification using name embeddings. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 1897-1906). ACM.

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

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., & Baeza-Yates, R. (2017, November). Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 1569-1578).

Zehlike, M., & Castillo, C. (2020, April). Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference 2020* (pp. 2849-2855).

Zhong, Z. (2018). A Tutorial on Fairness in Machine Learning. *Towards Data Science*. https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

Zhuang, Z., Elmacioglu, E., Lee, D., & Giles, C. L. (2007, June). Measuring conference quality by mining program committee characteristics. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (pp. 225-234).

Zippia. (2019). *Computer Science Professor Demographics and AND STATISTICS IN THE US*. Retrieved from https://www.zippia.com/computer-science-professor-jobs/demographics/

Zweben, S., & Bizot, B. (2018). 2018 CRA Taulbee survey. *Computing Research News*, *30*(5), 1-47.