12-2021

Fair and Diverse Group Formation Based on Multidimensional Features

Mohammed Saad A Alqahtani
University of Arkansas, Fayetteville

Fair and Diverse Group Formation Based on Multidimensional Features

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Engineering with
a concentration in Computer Science

by

Mohammed Saad A Alqahtani
King Khalid University
Bachelor of Science in Computer Science, 2009
Middle Tennessee State University
Master of Science in Computer Science, 2013

December 2021
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	Paul Cronan, Ph.D. Committee Member

David Andrew, Ph.D.
Committee Member

Abstract

The goal of group formation is to build a team to accomplish a specific task. Algorithms are being developed to improve the team's effectiveness so formed and the efficiency of the group selection process. However, there is concern that team formation algorithms could be biased against minorities due to the algorithms themselves or the data on which they are trained. Hence, it is essential to build fair team formation systems that incorporate demographic information into the process of building the group. Although there has been extensive work on modeling individuals’ expertise for expert recommendation and/or team formation, there has been relatively little prior work on modeling demographics and incorporating demographics into the group formation process.

We propose a novel method to represent experts’ demographic profiles based on multidimensional demographic features. Moreover, we introduce three diversity ranking algorithms that form a group by considering demographic features along with the minimum required skills. Unlike many ranking algorithms that consider one Boolean demographic feature (e.g., gender or race), our diversity ranking algorithms consider multiple demographic features simultaneously. Finally, we introduce a fair team formation algorithm that balances each candidate's demographic information and expertise. We evaluate our proposed algorithms using real datasets based on members of a computer science program committee. The result shows that our algorithms form a program committee that is more diverse with an acceptable loss in utility.

Acknowledgments

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

Firstly, thank to Allah, my lord, for guiding me and providing me all my strengths during the tough times during this research.

To my parents: all grateful words will not be enough to present my appreciation for all supports you provide for me and for being understandable and patient with me being away from you. I know I missed many events and moments that you guys want me there with you but from these moments, I will always be there. Thank you!

To my advisor, Dr. Susan Gauch: I reach this point because of your guidance and invaluable comments. Your insightful feedback, advice, and recommendations derive me to improve my thinking and bring my work to a high level. Thank you for your kindness and excellent collaboration.

To my committee members, Dr. Brajendra Panda, Dr. David Andrews, and Dr. Paul Cronan: Thank you for your all recommendations and valuable 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 giving their best effort to the university and students.

To my wife, Abeer, and my sons, Elias and Anas: +

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.

Dedication

To my Parents, Saad and Thamra

To my wife, Abeer

.

Table of Contents

1 Introduction 1

1.1 Motivation 1

1.2 Research Goals 3

1.3 Approach 4

2 Related Work 5

2.1 Profiling 5

2.1.1 Expertise Profiling 5

2.1.2 Demographic Profiling 9

2.2 Group Formation 11

2.2.1 Group Formation Algorithm 11

2.3 Fairness 13

2.3.1 Fairness in Group Formation 14

2.3.2 Fairness in Machine Learning 15

2.3.3 Fairness in Ranked Output 16

2.3.4 Bias in Academia 17

2.3.5 Diversity and Fairness in Academic Peer Review 19

3 Research Plan 21

3.1 Goal1: Creating a group Based on Expertise 21

3.1.1 Creating the Expert Profile 22

3.1.2 Forming the Group Based on Expertise 23

3.2 Goal2: Forming a group Based on Diversity 23

3.2.1 Collecting Demographic Information 23

3.2.2 Protection Classification 25

3.2.2 Creating a Demographic Profile 26

3.2.3 Forming Diverse Group 30

3.3 Goal3: A Hybrid Approach 35

4 Evaluation 38

4.1 Dataset 38

4.2 Baseline and Metrics 39

4.2.1 Metrics 39

4.3 Experiment 1: Evaluating Diversity-Based Recommendation 41

4.3.1 Boolean Based Program Committees 45

4.3.2 Continuous Based Program Committees 49

4.3.3 Boolean-Continuous Comparison 52

4.4 Experiment 2: Evaluating Our Multidimensional Fair Algorithms 55

4.4.1 Univariate Greedy Algorithm 55

4.4.2 Multivariate Greedy Algorithm 58

4.4.3 Multidimensional Similarity Algorithm 60

5 Validation 64

6 Conclusion 67

6.1 Summary 67

6.2 Future Work 68

7 Bibliography 69

List of Tables

Table 3.1 Sample of the raw data….. ……………………….………………..…….25

Table 3.2 Demographic Profile Classes ………….…………….………….….…26

Table 3.3 An Example of Boolean Weight Profile ………..…..…….………..……..28

Table 3.4 Career Stage Weight Allocation …………………….…………..…….30

Table 4.1 Conference Set of 2017……….…………………………….……...…..….39

Table 4.2 Boolean results for UGA, MGA, and MDA versus the baseline…………..47

Table 4.3 Continuous results for UGA, MGA, and MDA versus the baseline……..…51

Table 4.4 Comparison between Boolean and continuous weight representation……..54

Table 4.5 Average findings of UGA for different values of alpha…………..……….57

Table 4.6 Average findings of MGA for different values of alpha……………..…….59

Table 4.7 Average findings of MDA for different values of alpha………………..….61

Table 5.2: Comparison between our Hybrid (MDA) and the current PC's…………...65

Table 5.2: Comparison between our Hybrid (MDA) and the current PC's.…….……66

List of Figures

Fig 3.1 An Example of a Google Scholar Profile ………………………………….….…..22

Fig 4.1 Demographic Distribution of the Three Pools.……………………………...…43

Fig 4.2 Demographic Distribution of all PC’s Proposed by our Three Algorithms…….44

Fig 4.3 Comparison of the protected groups' improvement of SIGCHI 2017………….45

Fig 4.4 Comparison of the protected groups' improvement of SIGMOD 2017…......….46

Fig 4.5 Comparison of the protected groups' improvement of SIGMOD 2017…………46

Fig 4.6 Boolean Average Results of UGA, MGA, and MDA. ………………..…..….48

Fig 4.7 Comparison of the protected groups' improvement of SIGCHI 2017…………...49

Fig 4.8 Comparison of the protected groups' improvement of SIGMOD 2017…......….50

Fig 4.9 Comparison of the protected groups' improvement of SIGMOD 2017…………50

Fig 4.10 Continuous Average Results of UGA, MGA, and MDA…………………..…52

Fig 4.11 Weight Representation Assessment based on the average results of all PCs…..54

Fig 4.12 Comparison of F-score and D_S between UGA and the baseline ………………58

Fig 4.13 Comparison of F-score and D_S between MGA and the baseline ………………60

Fig 4.14 Comparison of F-score and D_S between MDA and the baseline ………………62

Fig 4.15 Comparison of F-score between all Hybrid approaches. ……………….………63

1. Introduction

1.1 Motivation

Research plays a significant role in advancing any nation's economy. Countries, universities, and companies spend millions of dollars on research in different areas of research to develop new methods in many industries. In 2015, the United States and China combinedly invested $840 billion on research and development [1]. Much of this research activity takes place in academia. According to [2], $68.8 billion was spent on research and development in 2015 by U.S. universities and colleges. Hence, faculty and Ph.D. candidates work on developing many innovations. In academia, publication plays a significant role in building their reputations and presenting their research findings [3]. Thus, they concentrate on choosing the proper conference or journal in which to publish their results. Once a paper is submitted to a conference, it undergoes a review process supervised by the program committee in order to select the best articles to present at the conference. Assigning qualified, unbiased reviewers for each paper is vital in the peer-review process. Unfortunately, some innovative papers may be rejected due to a poor review from an unqualified or biased reviewer. There are different types of bias such as gender, race, language, and location biases. As an example, a study by Casati et al. [3] shows that some journals prefer researchers who live in the same area of the journals. Moreover, a significant difference was found in the acceptance rates for medical conferences between those written by English-speaking versus non-English- speaking countries [3].

To address these issues, some work has been proposed to enhance peer review by making the process more transparent and less vulnerable to various types of bias [3]. One suggestion for reducing unfairness is considering diversity in the program committee responsible for peer-review. It was reported by [4] that diversity has a positive impact on the success of institutions and companies where they have accomplished better outcomes with more diverse teams. Having a more diverse team provides different points of view due to variety in backgrounds and experiences. Some studies have been done about the effectiveness of diversity specifically ethnicity, on scientific performance. The result shows that ethnicity improves scientific influence. Thus, having diverse individuals working together with various backgrounds and cultures leads to better outcomes [5].

It is clear that there is a lack of diversity within Computer Science as a whole. For example, fewer than 27% CS professionals are female and whites dominate more than 65% of CS professionals [6]. This lack of diversity is reflected in participation in conferences [7] and in the members of the program committees that govern academic conferences in Computer Science [8]. Addressing this, SIGCHI, one of the highest impact ACM conferences, announced an explicit goal to increase the diversity of their PC in 2020 [9]. Thus, having diverse individuals working together with various backgrounds and cultures leads to better outcomes [5].

The group formation process has been well studied in academia as a way to enhance learning and educational environments. Research has explored the best ways to form teams to improve the teaching methods in schools [10] and, more recently, work has focused on the importance of team composition to achieve diversity and optimal performance [11]. Group formation is also an important process for business, so other investigations have focused on how to form the best-qualified team to perform specific tasks [12] [13]. They concentrated in performing a team based on the qualification without considering demographic information.

In academia, group formation often overlaps work in the expert recommendation, systems that identify the best-qualified expert to achieve a particular task. Usually, the candidates' expertise is extracted from their publications. For example, early systems built an expert profile depending on the internal email network within an institution [14]. Although promising work has been done, several challenges remain. One of those challenges is that it is difficult to find a precise way to model the expertise of candidates based on their publications. Another critical challenge is that those systems tend to be biased towards well-known candidates making it difficult for new, but qualified, researchers to access opportunities. This is similar to the bias observed in machine learning systems that can perpetuate bias in their outcomes based on bias in the training data [15].

Despite its importance, few researchers have investigated building effective teams while, at the same time, selecting team members fairly with respect to demographic diversity. In other areas, new algorithms are emerging based on the notion of fairly treating candidates for jobs, loans, scholarships, and other opportunities from underrepresented groups. These efforts typically focus on members of protected groups, those groups who are protected against discrimination by US law, generally race, gender [16]. However, these algorithms maximize fairness with respect to a single protected feature. Thus, utilizing several demographic features to increase fairness is significant.

1.2 Research Goals

Our overarching goal is to develop group formation algorithms that can create a diverse group of experts. A key component of this is a fair expert recommendation system that provided opportunities to all candidates. Although our algorithms should be applicable in the team building system, we focus on the field of academic group formation. Mainly, this research could be valid to form a group of scholars to review papers in the conference or journals, or as a proposal review board, or an academic committee. In particular, we concentrate on forming a conference program committee from a set of candidate researchers. Our specific goals are as follows:

- Goal 1: Develop a group formation algorithm based on expertise only.
- Goal 2: Develop a group formation algorithm based on diversity only.
    - Subgoal 3.1: Represent diversity with multiple, Boolean features.
    - Subgoal 3.2: Represent diversity with multiple, continuous features.
- Goal 3: Develop a fair group formation algorithm that considers expertise and diversity.

1.3 Approach

In this dissertation, we develop and evaluate several algorithms to create a group of researchers based on their expertise and demographic features. We use, as our driving problem, selecting members of a conference program committee (PC). We create a group of candidate researchers from the authors of papers accepted by the conference in previous years and the conference program committee members of those years. For our first goal, we evaluate a variety of group formation algorithms, e.g., hill climbing, that select PC members based on their expertise only. For our second goal, we evaluate different ways of representing the desired diversity and selecting members based on diversity only. Unlike other approaches that used one Boolean feature to maximize the diversity, we focus on methods that take into account multiple, discrete, and continuous valued, demographic features simultaneously. Finally, we develop a hybrid method that provides a balance of the previous two approaches. We evaluate different values of the tuning parameter to select PC members. This approach tries to maximize diversity, while the utility loss is minimized.

2. Related Work

Our work focuses on forming a group of candidates based on their expertise and their demographic information. It builds upon previous work in Information Retrieval (IR), data mining, machine learning, and statistics. Hence, in the following sections, we discuss different areas of research that are related to our work. In the first section, we explore various tasks related to profile modeling, specifically, approaches to modeling expertise and demographics. Next, we discuss group formation techniques and investigate constraint satisfaction while forming a team to perform a specific task. The last section covers group fairness in which we discuss various causes of bias in group formation algorithms how this bias can occur in academia. The section concludes with a summary of different methods for forming a diverse group for peer review and why diversity is essential in that case.

2.1 Profiling

User profiles are an integral part of all work into personalization [17]; one cannot design systems that adapt to individuals without having an accurate model of the user’s capabilities and needs. A related area of research specifically addresses modeling an expert’s areas and depth of knowledge. A number of efforts have focused on presenting a comprehensible view of the knowledge of that expert. However, considering other factors such as demographic attributes to build different kinds of profiles is less well studied. In this section, we discuss both main types of profile, i.e., an expertise profile and a demographic profile.

2.1.1 Expertise Profiling

Expertise profiling focuses on modeling the skills and knowledge of an expert in a variety of areas [18]. This profile describes what an expert knows and how much he knows about a particular topic. Thus, the profile encapsulates the level of competency for an expert across a range of skills. One approach to expertise profiling is to manually assign scores to each member of an organization across a set of capabilities. In this approach, each expert can be represented as a vector of scores which can be aggregated to create an expertise matrix for the organization as a whole. However, due to the manual nature of this method, it is subject to reliability and scalability issues [19]. Thus, automated techniques have been proposed to employ the people's or organization's documents to collect skills.

There are two main language-based approaches to create an expertise profile, the *candidate-based* technique, and the *document-based* technique. The candidate-based approach builds the profile for an expert based on a set of documents that is related to that expert. "*P@noptic expert*", proposed by Craswell et al. [20], was one of the first research projects to use the candidate-based approach to build expert profiles. Based on these profiles, they developed a search service that allowed users to locate an expert based on the desired expertise. This search system is similar to a traditional search engine in which documents are retrieved based on a submitted query. They informally evaluated their system and reported that the results were promising. However, the evaluation was very general, and more precise evaluation, including comparisons to other approaches, is required to truly confirm the effectiveness of their solution. Reichling et al. [21] extended the candidate-based approach to expertise modeling by incorporating feedback from the experts to increase the profiles' accuracy. Their approach is semi-automatic since the initial profile is constructed for each expert based on documents related to them, but a refinement step requires the experts to evaluate and update their profiles to improve their correctness. Thus, this improved accuracy comes at the cost of an increased burden on the users.

Unlike the candidate-based approach that focuses on keyword extraction to model one expert at a time, the *document-based* attempts to identify broad areas of expertise by clustering texts into groups. Once the clusters are formed, it then attempts to identify experts corresponding to each cluster or area of expertise. Although the document-based approach has the advantage of identifying perhaps overlooked areas of expertise within an organization, in general, it is not able to accurately model the skills of an expert. Essentially, each expert is represented by a set of weights related to areas of expertise learned from the document collection as a whole. However, like all clustering approaches, it is difficult to explain the meanings of the various clusters and thus the areas of expertise.

Traditionally, expert retrieval, also called expertise finding, has been approached as an information retrieval problem and often addressed as an important part of a company's search system [14]. However, the importance of expertise profiling has expanded to academic communities that need to identify experts to review papers as a member of a conference committee or form a group of experts to review grant proposals or perform another research task. For example, Tang et al. [22] developed a text-based expert finding system called "Arnetmine." In their approach, the expertise is represented as a set of weighted keywords extracted from each scholar's home page and publications. To locate an appropriate expert, the users enter keywords related to the desired expertise. They used DISTINCT model proposed by Yin et al. [23]. to evaluate their work. This system uses different types of linkages associated with weights to differentiate objects' identities. The evaluation of their system is based on human judgments by comparing their system to the DISTINCT. Their results showed that their model outperformed the DISTINCT model with respect to precision and recall [22].

Chandrasekaran et al. [24] also model an expert's knowledge as a set of weighted areas of expertise. However, rather than using a set of learning areas in which the semantics are unclear, they based their profile on ACM's Computing Classification System (CCS) taxonomy. The primary purpose of creating this profile was to build a recommender system to suggest papers from the CiteSeer^x database to authors of CiteSeer^x papers. Chandrasekaran et al. [24] classified an author's published papers by ACM CCS taxonomy categories to create an expertise profile of that author. The outcome of this process is a profile represented as a weighted ontology rather than a list of keywords or learned areas. This work was expanded by Kodakateri et al. [25] to build users' profiles based on the papers they viewed rather than their published articles. Based on a user's profile, the system can suggest CiteSeer^x publications to end-users. Although this approach has not previously been used to recommend experts, the weighted ontology can be viewed as an expertise profile.

More recently, this approach was used by the Indiana Database of University Research Expertise (INDURE) project. In this work, the experts' information was collected from four sources: personal homepages, publications and PhD dissertations, profiles filled by candidates, and NSF funded projects. The keywords used to represent the candidates' skills were extracted from this text and then classified into ontologies developed by the National Research Council to create expert profiles. Hence, Users were able clients to browse expertise information and employ queries to search for an expert based on those ontologies. They modeled their query using query expansion to enrich the original client's query and thus, more specific details be provided [14].

Sateli et al. [26] developed a system called "ScholarLens" that also created a semantic expertise profile based on a scholar's publications. They compared two methods to extract the expertise of a scholar. The first method considers the full documents of the user's publications, and the other method uses the words "claim" and "contribution" to extract subsets of the text. In order to make these profiles applicable in different applications, they represented candidates' profiles using the Resource Description Framework (RDF) and competence records using Linked Open Data (LOD). They represented areas related to a specific domain by utilizing entity recognition and a linked open data (LOD) cloud. For instance, "information retrieval" and "algorithms" are both relevant to the computer science domain. They evaluated their model using human studies which showed that the two approaches performed with high precision for the top ten outcome expertise of each scholar.

One of the respectful measures to show a researcher's work's quality and productivity is the h-index metric, introduced by [27]. It produces a score that is computed based on the total number of publications and number of citations. It implicates the scholar's publications' quality and is utilized for employing, awarding committees, and funding determination [28] [29]. Moreover, different scholarly databases such as Google Scholar, Web of Science, and Scopus use this metric to present a researcher's output quality. Hence, in our research, we use this score to determine each candidate's expertise since it represents a precise score and provides an excellent analysis of research productivity, unlike others [30].

2.1.2 Demographic Profiling

Many studies [31] have shown the importance of considering demographic attributes when using automated systems to make decisions about people. However, online profiles typically exclude this information since users are often concerned about how such information might be used. So, when an organization wants to use demographic features to ensure fairness and anti-discrimination in their algorithms, they must often develop procedures to predict some attributes such as gender, nationality, and ethnicity based on the actual provided features such as the user's name [24]. Besides, the specific demographic attributes that are important to consider varies from one environment to another. In academia, for example, demographic profiling generally consists of attributes such as ethnicity, age, gender, race, and socioeconomic background [32].

Several suggestions exist for extracting gender and ethnicity. One of the earliest methods is the web scraping approach [33], in which demographic information is predicted from a personal home page. After that, the user profile consists of this and other information to be used by recommender systems. Recently, many methods used machine learning to classify ethnicity and gender based on a given name. However, because those systems mostly depend on a trained data set, they provide a strong result for some names and fail with others, which leads to significant accuracy issues. For example, Michael [34] investigated a list of more than 4,000 names to analyze the popularity of a name by country and used that to determine gender. This work was extended by [35] to improve accuracy and coverage. The work by [36] employed an Open Gender Tracking Global Name Data project that consists of names from four countries and then determined a person's gender based on the first name.

Similarly, Knowles et al. [37] applied the same approach, but they also operated an SVM classifier. When utilizing a Twitter data set, their model performed better than other models. Lately, a new model named "NamePrism" was developed by [38] to extract a person's name, ethnicity, and nationality. They analyzed the data of 57m clients of primary Internet service organizations using homophily in correlation patterns. In this study, we use a demographic profile technique used by [24] in which we use genderize.io and NamePrism to extract gender and ethnicity.

2.2 Group Formation

Different research areas have investigated the process of team formation with the goal of forming an innovative team. One key to a successful organization is having a good leader and collaborative group who work as a team to achieve all tasks [13]. In their study considering this issue, Anagnostopoulos, Aris, et al. [39] developed a model to build a group of experts to work on a specific task, taking into account the required skills of that task, the cost, and the fact that each member should have the same workload. Additionally, Brocco et al. [40] used human resource databases to develop a recommender model for companies. Their results show that to make a team successful, an organization should find a way to make all members collaborate. To this end, Deibel [10] developed two methods to form a group of students who are productive and participate appropriately as a team. She argued that focusing on the quality of collaboration is vital to create a productive team. Lappas et al. [41] were the first to develop an algorithm for building a group based on a social network. Their method focused on the expertise and the strength of the experts' relationship. Likewise, Owens et al. [42] claimed that the quality of the relationships between team members is crucial to the team's success. However, our research does not consider the social relationships that form within a group because considering this kind of connection leads to bias in the recommending process.

2.2.1 Group Formation Algorithms

Researchers have proposed many team formation algorithms with respect to the different factors they feel create the best possible team. For example, Juang et al. [43] developed BCPruning and SSPruning algorithms to select the best leader and correlated members of a team with the minimum communication cost. The BCPruning algorithm initially determines the leading experts able to obtain the lowest possible cost of communication and allows early termination of the candidate, while the SSPruning algorithm allows experts to present their expertise to limited skill distance, which leads to the selection of the best leader and corresponding members to form the team. Giancarlo Fortino et al. [44] delivers a method to create an effective team considering the trust-based procedure in virtual communities. Their technique is based on a voting system that, as real communities, takes reference recommendations to build a local reputation for each user. Unlike other methods that utilize a social relationship (global reputation), they focus on more reliable, trusted information to build the local reputation. This reliable information is provided by a user’s friend, a friend of a friend, and so on. In [45], the following algorithms were evaluated to find the best one for building an effective team: Bayes classifier, decision tree with ID3 and C4.5, CART, and Fuzzy K-means. Their results show that the Bayes classifier outperformed the other algorithms. However, they claimed that K-means and Fuzzy C-means could not be applied due to the type of data those researchers used. Chen et al. [46] proposed a genetic grouping algorithm for the reviewer group construction problem (RGCP). They tested their approach at the National Natural Science Foundation, China (NSFC) using a particular case in which they tried to automate the process of building a group of reviewers with regard to three features: age, region, and professional title. However, in our study, we create a group of experts considering five demographic features.

In academia, the grouping procedure is a well-known problem that has been studied intensely. For instance, Wang et al. [47] introduced DIANA algorithms to form a group of students based on multiple and continuous parameters. They applied their model with the constraint that no student should be ignored. Although DIANA performed better than a random selection approach and the results show a small difference within the group in terms of performance, the algorithm was limited by having only a small sample of students and needs to be tested with general size samples. The work by [48] proposed a genetic algorithm for heterogeneous grouping (GAHG) for building a team of computer science students to work on a software development project. The main criteria for selecting a student for the team were the student's programming and educational skills. Tabo et al. [49] derived a software tool to form a group for academic tasks. The tool creates a group based on people's predefined characteristics. However, this tool was built to resolve only the timing issue of forming a group. In a different perspective, Stepanova et al. [50] developed a method to create a team by focusing mainly on teaching students how they work as a team and what consequences might occur when they work improperly.

2.3 Fairness

Many studies have agreed with the need for teamwork to improve the performance and the speed of completing a task. Therefore, many efforts have been made to investigate the best way to form a group of talented people for a specific task. As discussed in the previous section, there are several algorithms to automate the process of recommending members to join the group; however, these methods can involve bias. There are several causes of this bias, such as the people involved in the selection, the models built from past data that could be biased, or the algorithms used to select candidates. Hence, in this section, we explore bias from different viewpoints. The first two sections provide a discussion about bias in group formation strategies and machine learning. In the third section, we discuss the importance of fairness in ranking. Then, we present literature related to bias in academia with real-world cases and facts that highlight this issue's importance. Finally, we focus on the task of peer review and how diversity and fairness can be achieved.

2.3.1 Fairness in Group Formation

Fairness requires that a protected population should be treated similarly as a whole [16]. Feldman et al. [51] investigated the problem of unintentional bias and how it impacts various populations that should be similarly treated. Additionally, they studied the disparate impacts by examining the process of predicting the protected population based on a variety of features. For example, the protected population can be defined based on gender, ethnicity, nationality, or other features.

Much of the research into fairness has focused on decision-making models, mainly supervised machine learning algorithms. These models can be viewed as classifiers that partition candidates into two classes, where the members of one class get a benefit (e.g., approved for a mortgage) and members of the other class do not. Group formation is an instance of this problem in which one class contains candidates selected to join the group and the other class contains excluded candidates. Several studies have shown that, even when the classifications algorithms themselves are not biased, the bias in training data may influence the outcome of the classifiers [51] [52]. Zamel et al. [53] derived a learning algorithm for fair classification by providing suitable data representation and at the same time obfuscating any data about membership in a protected group. Kamishima et al. [52] proposed a regulator that decreases the unfavorable result of a classifier by controlling the trade-off between the classification's accuracy and fairness. Similarly, Luong et al. [54] focused on achieving group fairness by proposing a fairness regulator or relabeling training data.

Although there is active research into incorporating fairness in decision-making algorithms, all of the approaches so far focus on approaches that consider a single protected feature at a time, e.g., race. The algorithms also assume that the candidates can be partitioned into two groups, protected or not, based on this feature, which limits the feature to have a Boolean value. We contribute to research in this area in two ways by considering multiple features simultaneously and also by extending the approaches to consider non-Boolean feature values.

2.3.2 Fairness in Machine Learning

The idea of protecting specific subsets of the population within overall communities against discrimination became law in the United States. These protected classes were identified based on gender, color, religion, age, national origin, ethnicity, genetic information, and citizenship. As a result, researchers have put significant effort into providing models that guaranteed fairness across these protected classes in decision making systems. Specifically, scholars developed machine learning approaches that focused on solving the issue of bias by identifying the main sources of bias in their results so that they could be eliminated. One of the main reasons for bias in machine learning outcomes is that future results often reflect bias present in training data. Thus, future outcomes may produce discriminatory outcomes that propagate historical disparity Asudeh et al. [15]. Even when the historical outcomes are not explicitly biased, training data can still create biased outcomes in machine learning due to a lack of training data for minority groups. Typically, the protected group is the minority group that means features related to this minority be less present in the data set. Thus, features related to the majority be reliably predicted by machine learning Zhong [55]. To address these issues, many efforts focus on enhancing the process of classification to prevent the occurrence of demographic disparity Dwork et al. [56].

Merely avoiding training on features that explicitly identify members of protected groups may not solve the issue of bias as some features may be related to protected features in some way. For example, if we avoid including race in a machine learning system, another feature such as neighborhood may be correlated to race [55]. Many authors have proposed solutions to achieve fairness by guaranteeing demographic parity, i.e., guaranteeing that members of the protected group appear in the result set with the same statistical representation as within the training data as a whole. However, this approach may result in utility loss of the overall solution. Thus, researchers attempt to use other ways to present fairness in their models. Hardt et al. [57] used *equalized odds* and *equal opportunity* to measure fairness instead of demographic parity. Those two criterions mainly focusing on achieving fairness and at the same time perpetuate high accurate classifiers. They used the ROC (Receiver Operator Characteristic) curve to evaluate their work and compare it to demographic parity. Their results showed that their proposed achieved lower utility loss compared to approaches focusing on demographic parity. Zafar et al. [58] introduced a new notion of unfairness that is *disparate mistreatment*, a negative situation that can arise when there is a higher misclassification rate for members of the protected reflect this type of bias. Thus, they proposed a new metric to measure the prevalence of disparate mistreatment for decision boundary-based classifiers. They developed a new approach to minimize disparate mistreatment and compared their work with that of Hardt et al. [57]. Their results showed that their proposed approach more effectively reduces *disparate mistreatment*.

2.3.3 Fairness in Ranked Output

Ranking is omnipresent in online systems since so many sites provide a list of items to the users in decreasing order of expected utility such as books in libraries, job opportunities, and opinions [59]. Whereas the previous work has focused on fairness in binary classification, Zehlike et al. [16] developed a fair ranking algorithm that incorporates protected features within ranked results. Their approach ensures that a proportion of the protected group within the results at any threshold is statistically at least equal to a predefined value. To guarantee such a condition, they utilized constraints to make sure a selected member was not less qualified than anyone not included. They achieved that by introducing two limitations: All group candidates should be more skilled than others left behind, and for every two candidates, the more qualified one should be ranked above the less qualified one. They considered one protected attribute (either gender or race) in different data sets for validation of their approach.

Singh et al. [59] proposed a model that maximizes ranking utility while minimizing unfairness based on three fairness constraints presented in Zafar et al. [58]. They investigated these constraints' effectiveness while employing probabilistic rankings and linear programming to find maximizing ranking utility. They utilized job seekers and news recommendation datasets in their experiment. The DCG (Dicounted Comulative Gain) metric and Disparate Treatment Ratio (DTR) were used to evaluate their system. They concluded that their model could effectively provide a high level of fairness while maximizing the overall utility. Singh et al. [60] propose the Fair-PG-Rank algorithm that maximizes utility, identifies bias sources, and follows application requirements of individual and group fairness constraints for ranking.

2.3.4 Bias in Academia

The problem of bias in academia has been investigated several times. A recent study by Gabriel [61] shows that ethnic racism exists in British academia. For example, in the UK, black professors constitute only 0.1% of all professors although they make up 1.45% of the UK population. Bornmann et al [62] studied the influence bias on the selection of doctoral and post-doctoral candidates. They investigated the influence of the following kinds of bias: gender, nationality, area of study, and organizational affiliation. Their results show some evidence of gender, area of study, and affiliation bias, but not nationality bias. Chaves et al. [63] claim that students give low scores evaluation for instructors who are female or from minorities. However, those who are white males get higher scores in the instructors' evaluations.

One of the significant areas of research into bias is the peer review process for scholarly publications. Lee et al. [64] studied different kinds of bias in peer review and how it impacts the review process of accepting or rejecting submitted articles. They concluded that fairness guarantees meritocracy and stability. However, a study by Holman et al. [65] provided some evidence that females are persistently underrepresented in publications from computer science, math, physics, and surgery. They employed the PubMed and arXiv databases that contain work by more than 36 million authors from over 100 countries published in more than 6000 journals. Their outcomes show a clear gender gap, especially in senior authorship positions. Moreover, prestigious journals have significantly fewer women authors. The rate of men submitting papers to journals is double that of women. One surprising result from their study provides evidence that there are proportionally fewer female authors from wealthy countries such as Japan, Germany, and Switzerland than from developing countries.

Lerback et al. [66] confirmed that females underrepresentation in peer review is still an issue. They claimed that the women authors’ lack of seniority is the main reason for such a problem. However, Murray et al. [67] conducted an analysis of bias in the peer review process for the biosciences journal *eLife* between 2012 and 2017 that indicates that there is still bias involved. They found evidence that a reviewer is more likely to accept publications by authors of the same gender and from the same country as themselves. To avoid this type of bias, many publications and conferences have adopted a double-blind review. However, several studies show that 25–40% of the time, reviewers can recognize authors [68] [69], which can lead to bias. Lane [70] suggested that within specific fields, these numbers be higher.

2.3.5 Diversity and Fairness algorithms in Academic Peer Review

The previous section shows show that there is work needed to address bias that exists in academia. Research by Rodriguez et al. [71] proposed an algorithm based on co-authorship networks to select paper reviewers for a submitted paper. Although their approach is efficient, and the algorithm itself is unbiased, focusing on the co-authorship network can itself lead to bias. Yin et al. [12] studied the relationships between bias and three features: the reviewer's reputation, the co-authorship connection, and the coverage. They found that the probability of the occurrence of bias increased as a factor of the strength of the social connection among authors. They suggested that to avoid biased results; one should ensure diversity in the peer review committee itself. Haffar et al. [72] investigate different peer review processes and provide evidence that shows various types of bias that may occur due to how those systems are applied. Wang et al. [73] investigate all accepted papers' average quality as impacted by editorial behaviors. Their results showed that when even only 10% of the editors were biased, the quality of accepted papers was reduced by 11%. They proposed an agent-based model to enhance the process of peer review and suggested that increasing the number of reviewers can improve the quality and fairness of the peer review. However, this approach creates a higher overall review burden. The importance of a diverse peer review committee was also supported by Lane et al's recent study [74] that showed that a diverse group of reviewers leads to higher quality peer review and a reduction in bias.

To address the issues above, Chen et al. [46] developed a model to form a diverse peer review committee by utilizing a hybrid genetic algorithm. Although they were able to improve the diversity of the resulting committee, they did not address the problem of possible expertise loss in the resulting group.

To summarize, bias exists within academia even though the research community has taken strides to avoid it. Several studies indicate that increasing diversity when forming a group can enhance its quality of work and produce fairer results. However, no current algorithms focus on guaranteeing diversity in peer review groups while also balancing the effort expended and the overall group expertise.

3. Research Approach

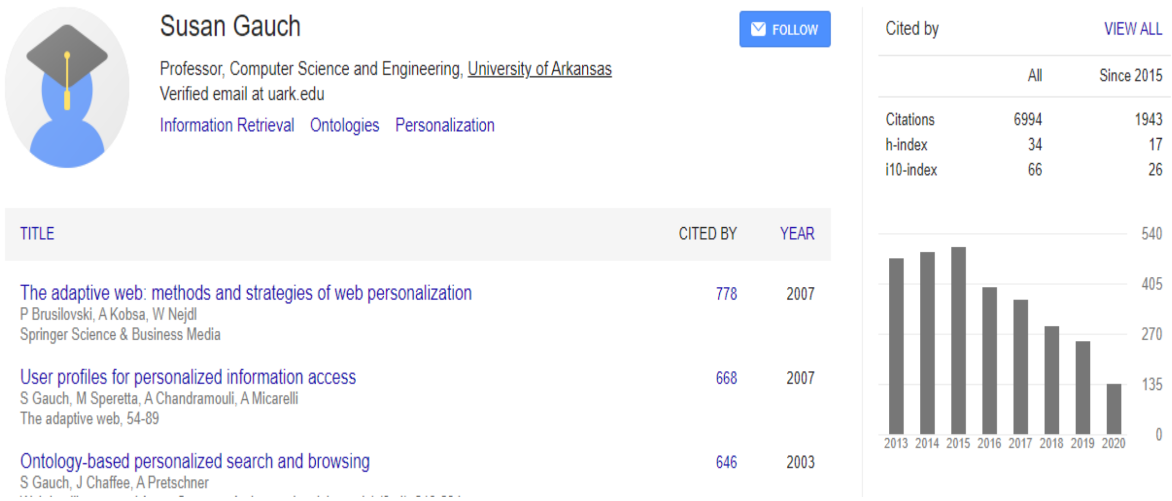
The main goal of this research is to develop algorithms that create a diverse group while minimizing utility loss. These algorithms provide ways to form a wide variety of groups that balance demographic variety and expertise tasks such as a group of students and employee people in companies. In our research, we focus on the task of creating a program committee (PC) for a specific conference. This is used as a driving problem to demonstrate and evaluate our group formation process. We discuss our first approach to build a PC based on expertise alone that solely considers the researchers' knowledge in a specific field. In our second approach, we focus on building a group solely based on their demographics, maximizing diversity. Most previous research concentrates on promoting a diverse group depending on a single parameter such as age, or gender. However, we present novel algorithms that consider several features simultaneously. Our third approach is a hybrid of the previous two that attempts to form a group that balances the need for experienced scholars with the desire for diversity among the group members.

3.1 Goal 1: Creating A Group Based on Expertise

The Information Retrieval research community has been studying the problem of expertise retrieval for quite some time. They have focused on two main issues, i.e., 1) finding appropriate expertise by searching for people with expertise in a particular topic; and 2) creating expertise profiles that summarize individuals' depth of expertise across a range of topics [75]. To address our first goal, creating a group of researchers based on their expertise, we must address both problems. To accomplish that, we need to represent an expertise profile for each author. Then, we must develop a selection algorithm to choose the experts for the PC based on their expertise.

3.1.1 Creating the Expert Profile:

To build the profile for a scholar, we need to know her or his name and have some source of information about their skills. In our research, we use the *h-index* score of each researcher as evidence of their expertise. This metric was introduced by Jorge E. Hirsch in 2005 [27] to compute the value of the researcher's works. We use Google Scholar [76], a publicly open web search engine that consists of scholarly literature that includes the publications, citations number, and h-index score of each researcher. As we mentioned earlier, we utilize the scholar's name to extract the h-index of that researcher from his or her Google Scholar profile. An example of the h-index score provided by the Google Scholar profile is shown in Fig 3.1.



Susan Gauch

Professor, Computer Science and Engineering, University of Arkansas
Verified email at uark.edu

Information Retrieval Ontologies Personalization

Cited by VIEW ALL

	All	Since 2015
Citations	6994	1943
h-index	34	17
i10-index	66	26

TITLE	CITED BY	YEAR
The adaptive web: methods and strategies of web personalization P Brusilovski, A Kobsa, W Nejdl Springer Science & Business Media	778	2007
User profiles for personalized information access S Gauch, M Speretta, A Chandramouli, A Micarelli The adaptive web, 54-89	668	2007
Ontology-based personalized search and browsing S Gauch, J Chaffee, A Pretschner	646	2003

Fig 3.1 An Example of a Google Scholar Profile

Once we extract the *h-index* for all candidates, we utilize it as that candidate's expertise profile. Since it is based on their track record of citations, this score allows us to rank the candidates by the depth of their research productivity and reputation.

3.1.2 Forming the Group Based on Expertise

Once we have acquired expertise profiles for the researchers, we are ready to start creating a program committee (PC) for that conference based on only expertise. In general, we have a pool of qualified researchers consisting of N members, and we need to select M of them to build the group. The expertise group formation method uses a greedy algorithm that iteratively adds the expert with the highest h-index to the PC until all M members have been selected.

3.2 Goal 2: Creating a Group Based on Diversity

Our second goal focuses on algorithms to form groups that maximize diversity. When creating a group of people to work on a specific task, we typically focus on selecting the best-qualified ones to achieve our goal. Indeed, people always aim for the highest level of achievement for each mission. Thus, they select the same people every time they have a task similar to previous ones they have already accomplished. However, a concern arises regarding fairness and providing opportunities to all people. Certainly, there is a minimum level of expertise required to be considered to join a group but, beyond that, we want to guarantee that groups are composed of representatives from all demographic groups. To achieve groups that guarantee demographic diversity, we create a demographic profile similar to [2]. In this work, we concentrate on five demographic features that have been found to be sources of potential bias in academia: gender [77] [8], ethnicity [61], geolocation [67], career stage [7] [8], and university rank [77]. Then, we develop and evaluate algorithms that form a group based on the demographic profiles

3.2.1 Collecting Demographic Information

We model each scholar using the five demographic features mentioned previously. For each, we collect their demographic information from publicly available information. We write web-scraping scripts to help automate the process as much as possible. The following describes our approach to extracting the demographic features:

- **Gender**: We determine the gender of each scholar using NamSor [78], a tool that predicts gender based on an individual's full name. It also returns the degree of confidence in the prediction in a range between 0 and 1. NamSor gender API has an overall 98.41% precision and 99.28% recall [79]. Additionally, the NamSor inventor used official directory of the European union [80] to assess the NamSor API gender for European names. The gender error rate was only less than 1%. It returns the following set of values: {Male, Female}.
- **Ethnicity**: We determine the ethnicity of each scholar using NamSor [78], a tool used to extract the ethnicity of an individual based on that individual's full name. For each name, they return the most likely ethnicity from a set of 5 possibilities, e.g., {W_NL (for White), B_NL (for Black), HI (for Hispanic), A (for Asian), or O (for others)}. NamSor is widely used to predict ethnicity and gender in many studies.
- **Geo-Location:** We determine the location at which the researcher works by extracting the University name from their Google Scholar profile (scholars without profile pages be omitted from the data set). From the university home page, we extract the country in which it is located, and, in the case of US universities, we determine the state and store the two-letter abbreviation for the state. We obtain a set of values that consists of the countries name, e.g., {USA. UK, India). For those in the United States, the state abbreviation represents the Geo-Location such as AR, NY, and CA.
- **Career Stage:** We extract each scholar's academic position from their Google Scholar profile. We consider the following sets of ranks: {Student, Post-doc, Assistant Professor,

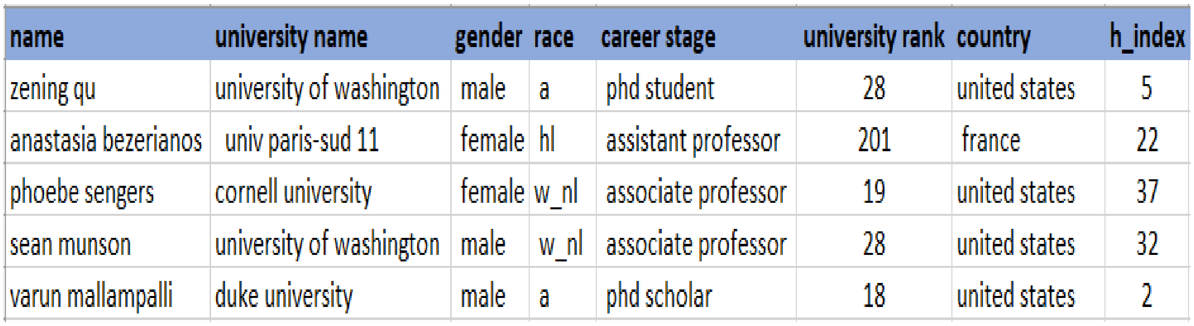
Associate Professor, Professor, Distinguished Professor. We omit from our data set researchers whose primary appointment is within the industry.

- **University Rank:** We extract each scholar's affiliated university from their Google Scholar page. Then, we collect the rank of that university using Times Higher Education [81]. This site ranks institutions from 1 to 1001+.

Note: Candidate PC members without a Google Scholar profile, and those without academic positions, are omitted from our dataset.

After we collect all data, we have the raw data that we utilize in this research to represent the desire profiles. Table 3.2 shows an example of raw data for some researchers from SIGCHI 2017 Program Committee.

Table 3.1 Sample of the raw data



name	university name	gender	race	career stage	university rank	country	h_index
zening qu	university of washington	male	a	phd student	28	united states	5
anastasia bezerianos	univ paris-sud 11	female	hl	assistant professor	201	france	22
phoebe sengers	cornell university	female	w_nl	associate professor	19	united states	37
sean munson	university of washington	male	w_nl	associate professor	28	united states	32
varun mallampalli	duke university	male	a	phd scholar	18	united states	2

3.2.1.1 Protected Groups

The definition of the protected group depends on several factors such as environment, culture, and organization [2]. We base our definition of the protected parameters on which group is typically underrepresented in the population being studied, i.e., researchers in Computer Science. Table 3.3 outlines the protected and non-protected classes for each of our demographic features.

Table 3.2 Demographic Profile Classes

Features	Classes (Protected / non-Protected)
Gender	Female / Male
Race	Non-White / White
Geo-Location	Developing /Developed (by country)
	EPSCoR / Not-EPSCoR (by state in USA)
Career Stage	Junior / Senior
University Rank	More than mean / Less than or equal mean

3.2.1.2 Creating a Demographic Profile

Once we collect the raw attribute values for each scholar, we investigate two methods of representing an individual's demographic profile, one based on discrete weights for the features and another based on continuous weights.

Mapping Boolean Weights

Because most previous work on fairness focuses on a single Boolean value, we begin by mapping each feature to a Boolean value, either 1 (true) or 0 (false), based on that author's membership (or not) in a protected group. Thus, each feature in a profile is represented using a Boolean value, generally, 1 (True) if the individual is a member of the protected group and 0 (False) otherwise.

The following describes our approach to interpret those values to Boolean values:

- **Gender**: In our case, the gender protected parameter is Female. Thus, we assign this to 1, and Male to 0.
- **Ethnicity**: For ethnicity parameters, we partition them into two values; 0 for White and 1 for Non-White.
- **Geo-Location:** We determine the Boolean value of Geolocation by country GDP per person rank. If the GDP of a country is above the world average rate, we consider that country as developed (Boolean value = 0), otherwise developing (Boolean value = 0). For those who are from the United States, we assign the Boolean value by state EPSCoR; 1 for EPSCoR state and 0 otherwise. Those ranges were selected based on their scale in our sources.
- **Career Stage:** We determine the Boolean value of Career Stage by considering the scholar as senior or junior; Student, Post-doc and, Assistant Professor be junior (value = 1), and Associate Professor, Professor and, Distinguished Professor be senior (value = 0).
- **University Rank:** For university rank, if the value of university rank is above the average, we rate it as low (Boolean value = 1), otherwise high (Boolean value = 0). Unranked University is determined as a protected class which have a value = 1.

Note: For any attribute for which we cannot find a value, we assign the value to None and, in most cases, exclude that scholar from our future experimental dataset.

Once we have created a demographic profile for a researcher, we can calculate their diversity score using the following formula:

$$\text{Score(D)} = \sum_{i=1}^{n} d_i \tag{3.1}$$

where d_i represents the diversity value for each demographic feature. Thus, diversity scores for an individual user are in the range from 0...5.

To illustrate what we have described, we provide a sample of a researcher's profile presented using just three attributes in our example: Gender, Geolocation, and University Rank (U. Rank). Table 3.5 shows a simplified demographic profile for a female researcher working at the University of Arkansas which is located in the U.S., a developed nation but ranked > Avg.

Table 3.3 An Example of Boolean Weight Profile

Class	Gender	Geo-Location	U. Rank
Variable	female	U.S. (Developed)	ranked (>Avg)
Value	1	0	1

Based on this demographic profile, her diversity score would be 2.

Mapping Continuous Weights

A Boolean value is a very imprecise way to represent the attributes. For features that have a range of raw values, partitioning them into two classes may introduce avoidable errors. For example, a university ranked exactly 200 would be in the protected class whereas its near neighbor, ranked 201, would be in the non-protected class. Thus, we explore using a range of values for each feature using a continuous weight in the range of 0.0 to 1.0, as follows:

- **Gender**: In this feature, 27% of computer science professionals are females [82]. Hence, we assigned them a value of 1-0.27. However, males are assigned a value of 1-0.73.

- **Ethnicity**: We have a set of five ethnicities: White, Black, Asian, Hispanic, and others. In the United States, whites dominate 65% of computer science and engineering employment. In the same fields, Asian, Black, Hispanic, and Others represent 19.8%, 5.6%, 7.5%, and 0.5%, respectively. Hence, whites are assigned a value of 1-0.65, Asian, Black, Hispanic, and Others are assigned values of 1-0.198, 1-0.056, 1-0.075, and 1-0.005, respectively.
- **Geo-Location:** We consider the GDP (Gross Domestic Production) to represent the values of this feature. Those data are retrieved from the World Development Indicators database [83]. We assign the weight of each GDP value using the following equation:

$$W_{Geo} = 1 - \frac{C_{GDP}}{MAX_{GDP}} \tag{3.2}$$

where MAX_{GDP} is the maximum country GDP and C_{GDP} denotes each country's GDP. Thus, we have a weight between 0.0 and 1.0 for each country. For those who are from the United States, a state is assigned a weight based on the fund a state received from NSF.

- **Career Stage:** In this feature, we differentiate the academic career stage into six positions. Each of them is associated with a weight in the range of 0.0 to 1.0, described in the following table:

Table 3.4 Career Stage Weight Allocation

Position	Weight
Full Professor or above	0.16
Senior or Principal	0.32
Associated Professor	0.48
Assistant Professor or Lecturer	0.64
Post-Doctoral or Research Fellow	0.80
Graduate Student	1.0

- **University Rank:** In this feature, we have a range of university ranks between 1 to +1001. For unranked universities, we assigned them a value of 1001. Once we have all universities' ranks, we use the following equation to compute the weight of each affiliation:

$$W_{UR} = \frac{U_R}{L_R} \tag{3.3}$$

Where U_R is a rank of each university and L_R is the lowest university rank in our collection. For each university rank, the weight is in the range of 0.0 to 1.0.

3.2.2 Forming a Diverse Group

Once we have expertise and diversity profiles available for a pool of candidates, we need to develop techniques to form the group. The previous section focused on creating the group based only on expertise; this section details our approaches to building a group based on only demographic diversity. We implement and evaluate several approaches to forming a group that maximizes diversity. Maximizing fairness with respect to a single protected variable has been well-studied recently. However, maximizing fairness along multiple dimensions is relatively uncommon, requiring us to develop and evaluate new algorithms. These approaches are described in the following sections.

3.2.2.1 Approach 1: Univariate Greedy Algorithm

The first approach makes use of only the demographic profiles of the candidates in the pool. We compute the diversity score for each researcher based on their demographic profile based on the sum of their feature values as described in equation 3.1. These profiles are ranked based on the diversity score. Then, the candidates are selected to join the group using a greedy algorithm. To achieve this, the candidates are placed in a priority queue based on their scores. Then, the top candidate is iteratively removed from the priority queue until the desired group size is achieved. For example, if forming a program committee for a conference, the desired PC size is set based on the size of the current, true PC for that conference. If we have two or more researchers with the same diversity score, and we get to choose only one or some of them, we randomly select the candidate that gets added to the group. We implement this approach using both the discrete- and continuous-weight demographic profiles.

Algorithm 1 Univariate Greedy

```
priority_queue ← Initialize an empty queue
For each profile:
    Diversity score ← compute profile score
    Add profile to priority_queue using diversity score as priority order
candidates ← Select $N$ profiles from top of priority_queue
```

3.2.2.2 Approach 2: Multivariate Greedy Algorithm

The previous methods maximize the resulting group's diversity score, but it does not guarantee multidimensional diversity among the resulting group members. It could result in a high diversity score by selecting an entirely female group, for example, while accidentally excluding any members from minority groups. To address this shortcoming, we develop a multivariate greedy algorithm that employs multi-faceted ranking. It specifically creates one priority queue per demographic feature and selects from each queue in a round-robin fashion until the group is formed. Since the proposed demographic profiles model five features, we create five priority queues, each of which orders all the candidates by one specific feature. For example, one queue orders the candidates based on gender, whereas another would order the candidates based on academic position. Once all queues have been sorted, we apply round-robin selection by picking the highest sorted profile from each queue. Accordingly, the selected profile is then eliminated from all queues to endure that the same profile will not be selected again. This process continues iteratively until the group reaches the desired size. We use this approach with both discrete- and continuous-weight diversity profiles.

Algorithm 2 Multivariate Greedy

```
feature_names ← List of all queue names, one per features
For each feature in feature_names:
        priority_queue[feature] ← Initialize an empty queue
For each profile:
        For each feature in feature_names:
                score[feature] ← compute profile score for each feature
                Add profile to [feature] using score[feature] as priority order
candidates ← empty list
While number of candidates < N:
        feature ← feature_names[0]
        Repeat:
                candidate ← Get and remove profile from priority_queue[feature]
```

```
13.     Until $candidate$ is not in $candidates$
14.     Add $candidate$ to $candidates$
15.     Rotate $feature$ to end of $feature\_names$.
16. Now we have $N$ candidates selected.
```

3.2.2.3 Approach 3: Multidimensional Similarity Algorithm

This approach treats the demographic profiles as a vector of features, and it attempts to maximize the diversity of the resulting group across multiple dimensions simultaneously. This hill-climbing method is based on calculating the similarity between a gap profile, $\overrightarrow{GP}$, and the available candidate profiles, $\vec{C_i}$, using equation 3.4. We create a gap profile, $\overrightarrow{GP}$, that summarizes the difference between our goal and current group composition using formula 3.5. Initially, the gap profile is just the goal profile that is set based on the sum of all $\vec{C_i}$ in the pool of candidates normalized by the number of candidates. Thus, the goal profile represents the demographic variation in the pool of candidates, and we use that as our goal in order to strive for demographic parity. We also create a group profile that summarizes the demographic variation in the group so far. The group profile is initialized to contain zero weights for all features. As candidates are added to the group, their demographic profiles are used to update the group profile, also using formula 3.5.

$$\text{Cos (A,B)} = \frac{\sum_{j=1}^{m} a_j b_{ij}}{\sqrt{\sum_{j=1}^{m} a_j^2 * \sum_{j=1}^{m} b_{ij}^2}} \qquad (3.4)$$

The algorithm begins by initializing the four types of profiles initialized, the candidate profiles, the goal profile, the group profile, and the gap profile. Then, we develop an iterative algorithm that, at each step, calculates the similarity between each remaining candidate profile and the gap profile using equation 0.0. Based on the results, the most similar candidate is selected to join the group. The group profile is updated by adding the selected candidate's profile and the new group profile is used to update the gap profile. This process is repeated until the group reaches the desired size.

$$\overrightarrow{GP_T} = \overrightarrow{GL_{T-1}} - \overrightarrow{GR_{T-1}} \tag{3.5}$$

$\overrightarrow{GL}$ = Goal profile
$\overrightarrow{GR}$ = Group profile
$\overrightarrow{GP}$ = Gap profile

Similar to the previous approach, we apply this technique with both discrete- and continuous-weight diversity profiles.

Algorithm 3 Multidimensional Similarity

```
profiles ← empty list
for each profile in Candidates file:
    add profile to profiles
stats ← {
        mean_country_gdp ← average(gdp of all countries in Geo file)
        mean_u_rank ← average(u_rank of all profiles in candidates file)
        max_u_rank ← max(u_rank of profiles in candidates file)
        }
w ← {}
for each p in profiles:
    w[p] ← weight_function (p, stats)
GL ← {0, for each feature}
for each p in profiles:
    GL ← GL + w[p]
GL ← GL / length(profiles)
GP ← GL
GR ←{0, for each feature}
selected_profiles ← empty list
N ← number of candidates to select
loop N times:
    GP ← GP - GR
    most_similar_profile ← max (cosine_similarity(w[p], GP) for each p in profiles)
    add most_similar_profile to selected_profiles
    GR ← {0, for each feature}
    for each p in selected_profiles:
        GR ← GR + w[p]
    GR ← GR / length(selected_profiles)
return selected_profiles
```

3.3 Goal 3: A Hybrid Approach

Our third goal focuses on algorithms to form groups that attain *demographic parity*, i.e., the participation rate for each demographic feature in the PC matches the participation rate in the pool from which the group members are drawn. For the preceding two goals, we apply two main methods to accomplish two main outcomes. The first approach creates a program committee in which each member is selected to maximize the expertise required by a conference. However, this approach does not consider the resulting group's diversity. In contrast, our second approach builds a diverse program committee in which each member has at least a minimal amount of expertise related to the conference, but it has no further focus on the expertise level of the group as a whole. Hence, each approach has advantages and disadvantages. To attempt to get the best of both worlds, we introduce a hybrid approach to incorporate both of these goals by including a tuning parameter (α) which we use as in the following equation:

$$Score(H) = (1 - \alpha) * Score\ (E)' + \alpha * Score(D)' \tag{3.6}$$

where Score (E)' is the normalized version of Score (E) and Score (D)' is the normalized version of Score(D).

In order to combine the scores as described, they must be in the same range. To achieve this, we use z-value (Standard Score) to normalize both *Score(E)* and *Score(D)*. The equation of this process is introduced as follow:

$$z(Score) = \frac{Score - \mu}{\sigma} \tag{3.8}$$

Z(Score): The standard score

μ: The mean of the score

σ: The standard deviation of the score

The main challenging here is to find the most effective method of achieving demographic parity with a minimum loss in overall expertise. Hence, we test different values of α from 0.0 to 1.0 in steps of 0.0 and compare the results to the previous approaches. Note that when α is 0 the result is just the Expertise Score, i.e., the best method for goal 1. Similarly, when α is 1.0, the result is just the Diversity Score, i.e., the best method for goal 2.

As a part of our evaluation process, we investigate the influence of the tuning parameter accompanied by the normalization by test the results versus the previous approaches. Thus, we explore the best way to combine the two approaches to achieve a balance between expertise and diversity.

Algorithm 4 A Hybrid Approach

```
select_candidates (profiles, score_fn, group_size)
    scores ← {score_fn(profile) for profile in profiles}
    queue <- sort profiles by descending order of scores
    return queue[0 : group_size]
hybrid_select_candidates (profiles_by_scoreE, profiles_by_scoreD, α, group_size)
    candidates ← profiles_by_scoreE union profiles_by_scoreD
    score(E)’ ← {normalize profile’s score(E) for profile in candidates}
    score(D)’ ← {normalize profile’s score(D) for profile in candidates}
    score_fn: (profile) → $(1-\gamma) * Score\ (E)' + \gamma * \ Score(D)$’
    return select_candidates (candidates, score_fn, group_size)
```

4. Evaluation

In section 3, we present several approaches to form a group of candidates with respect to both ontology-based expertise and ontology-based diversity recommendations. The outcomes of our algorithms are sets of ranked lists that produce the top-ranked candidates. Thus, we form our groups based on those rank-ordered lists. Our algorithms attempt to generate a more diverse PC. Accordingly, we need to evaluate the effectiveness of our algorithms using $Diversity\ Gain\ (D_G)$ of our proposed PCs and the utility loss caused by each algorithm. In this section, we first introduce our datasets, then we illustrate the evaluation of our algorithms.

4.1 Dataset

We evaluate algorithms based on the process of adding candidates to join the conference program committee for a set of three conferences. For our experiment, we concentrate on ACM conferences because they cover a wide range of topics with which we are familiar. We select the ACM conferences to study based on several criteria: 1) the conferences should have high impact; 2) the conferences should have little or no overlap in topics; 3) the conferences should cover major topics of computer science; and 4) the conferences should have a reasonably large number of PC members and accepted papers. Based on these criteria, we reviewed all ACM SIG conferences and selected three of them: SIG-CHI (The ACM Conference on Human Factors in Computing Systems), SIG-MOD (International Conference on Management of Data), and SIGCOMM (The ACM Conference on Data Communication). We exclude candidates who: 1) do not have a Google Scholar profile; 2) are missing at least one feature's value; 3) primarily worked in the industry. Based on these criteria, we create a pool for each conference that contains both PC members and authors of accepted papers; The details of these conferences are shown in Table 4.1.

Table 4.1 Conference Set of 2017

Conference Name	Rank- Aminer	PC	Pool
SIGCHI	10	213	649
SIGMOD	39	130	420
SIGCOMM	52	23	148

4.2 Baseline and Metrics

We generate a ranked list of experts from each algorithm and use each list to build the PC. The main goal is to create a diverse and fair group while maximizing utility. Hence, we consider our expertise-based algorithm as the **baseline** in the evaluation process. Thus, we calculate this algorithm's utility and compare it to the other algorithms described in section 3. We evaluate each algorithm by comparing the ranked list of experts it recommends to the baseline. We assume that the PC size is the same as the current PC size and perform the comparison based on several metrics.

4.2.1 Metrics

We have five metrics that measure the resulting group's expertise:

1. the Diversity Gain (D_G) of the proposed group (see formula 4.1);
2. the utility loss (UL_i) of the proposed group (see formula 4.2);
3. the utility savings (Υ_i) of PCi relative to the baseline (see formula 4.3);

4. the F-measure [84] to examine our algorithms' ability to balance diversity gain and utility savings for goal 1, and to balance the distance similarity and utility saving for goal 2: (see formula 4.4);
5. the distance similarity (D_S) to measure the demographic parity (see formula 4.5).

For the first metric, we compute the Diversity Gain (D_G) of each proposed PC using the relative percentage gain (ρ_{G_i}) 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.

$$D_G = \text{MIN}\ (100,\ \frac{\sum_{i=1}^{n} \rho_{G_i}}{n}) \qquad (4.1)$$

By choosing to maximize diversity, it is likely that the utility of the resulting PC has slightly lower expertise. To measure this, drop in utility, we use the average h-index of the PC members and compute the utility loss (UL_i) for each proposed PC using the following formula:

$$UL_i = \frac{U_b - U_{P_j}}{U_b} * 100 \qquad (4.2)$$

where $U_{P_j}$is the utility of PC*i* and U_b is the utility of the baseline.

For the third metric, we compute the utility savings of PC*i* by subtracting the utility loss of each PC from 100 as follows:

$$\Upsilon_i = 100 - UL_i \qquad (4.3)$$

The fourth metric computes the F-score [84] to examine the tradeoff between two metrics using the following equation:

$$\text{F} = 2 * \frac{A * \text{B}}{A + \text{B}} \qquad (4.4)$$

An algorithm may overcorrect and thus, derive a bias issue in favor of some protected groups. Therefore, we introduce the Euclidean distance to evaluate our algorithms' ability to achieve demographic parity. To achieve that, we compute the similarity between a pool demographic distribution and each proposed PC derived from that pool using the following formula:

$$\mathrm{d(p,q)} = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \tag{4.5}$$

Where p_n is the percentage of each demographic group in the pool (such as females dominate 0.45 of the pool) and q_n is the percentage of each parameter in the proposed PC.

4.3 Experiment 1: Evaluating Diversity-Based Recommendation

In this section, we consider the process of recommending candidates to build the PC based only on their diversity scores. As described in section 3.2, we do not use researchers' expertise in our diversity-based recommendation algorithms. For each of the three conferences, we evaluate the following algorithms for diversity recommendation:

1. Univariate Greedy Algorithm (UGA)
2. Multivariate Greedy Algorithm (MGA)
3. Multidimensional Similarity Algorithm (MDA)

Each of these algorithms generates two groups of candidates: one based on Boolean weights and the other based on continuous weights described in section 3.2. For each diversity-based group so formed, we compute the diversity score for each candidate, and we evaluate the algorithms using the following metrics:

1. Diversity Gain (Formula 4.1)
2. Utility Loss (Formula 4.2)

3. Utility Savings (Formula 4.3)
4. F-measure (Formula 4.4)

These results are used to identify the algorithm that best maximizes the diversity of the proposed group. Furthermore, we evaluate the best weights representation by comparing the Boolean-based outcomes to continuous based outcomes. Once we determine the best weight representation, we use it in our Hybrid approach.

Each of our algorithms generates a ranked list that we utilize to select a desired number of candidates to form the PC. We have a pool of candidates for each conference consisting of PC members and authors of all accepted papers. Hence, before investigating our algorithms, we first show the data distribution of each pool. We present those pools based on the protection classification described in section 3.2 (see Fig 4.1). These clearly illustrate that all of the three pools had a low participation rate from the most protected groups. As an example, SIGCOMM 2017 had only 16.22% female pool members and, SIGMOD 2017's pool was only 16.43% female. Similarly, seniors dominate with 59.63% of SIGCHI 2017, 58.81% of SIGMOD 2017, and 55.41% of SIGCOMM 2017 pools. However, SIGMOD and SIGCOMM pools had more non-white than white members.

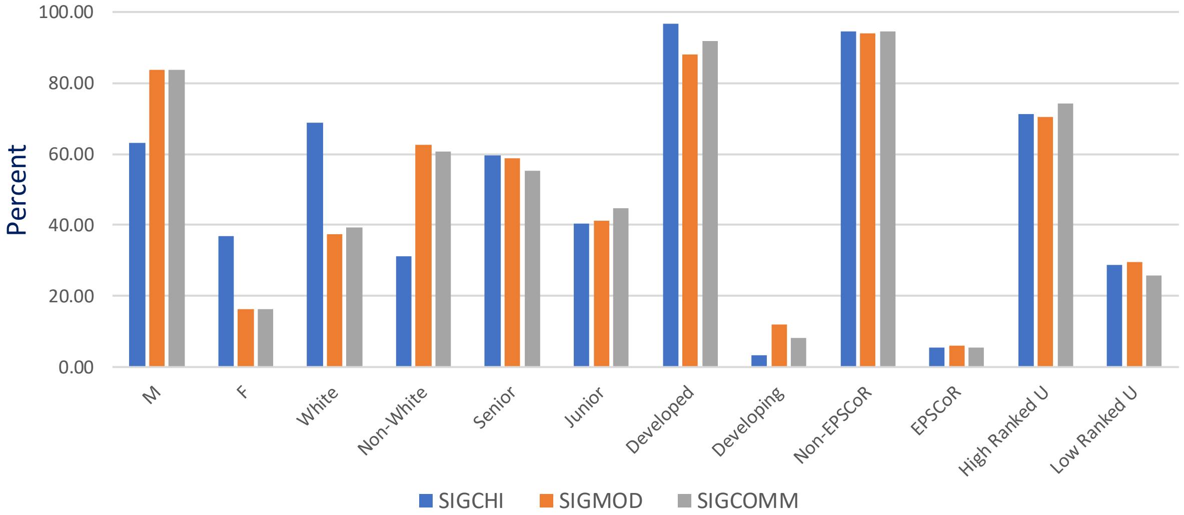


Figure 4.1: Demographic Distribution of the Three Pools.

In the previous paragraph, we show the data distribution of all pools being utilized in our research. Hence, we implement our baseline to form the three PC's based on the expertise only. We use these PCs to evaluate our algorithms. The demographic distribution of the three produced PC's by the baseline is shown in Fig 4.2. We can demonstrate that all of the three PCs had a considerable gap between protected and non-protected groups. For instance, senior dominates with 91.08% of SIGCHI PC, 99.23% of SIGMOD 2017, and 91.30% of SIGCOMM 2017 PC. Likewise, SIGCHI PC had only 27.70% of females, and SIGMOD PC was only 16.92%. Noticeably, the baseline does not select any member from developing countries or EPSCoR states for SIGCHI 2017. Although it selected members from developing countries and EPSCoR states for SIGMOD and SIGCOMM, the gap is still huge as there are more than 89% members from developing countries and more than 97% from EPSCoR states.

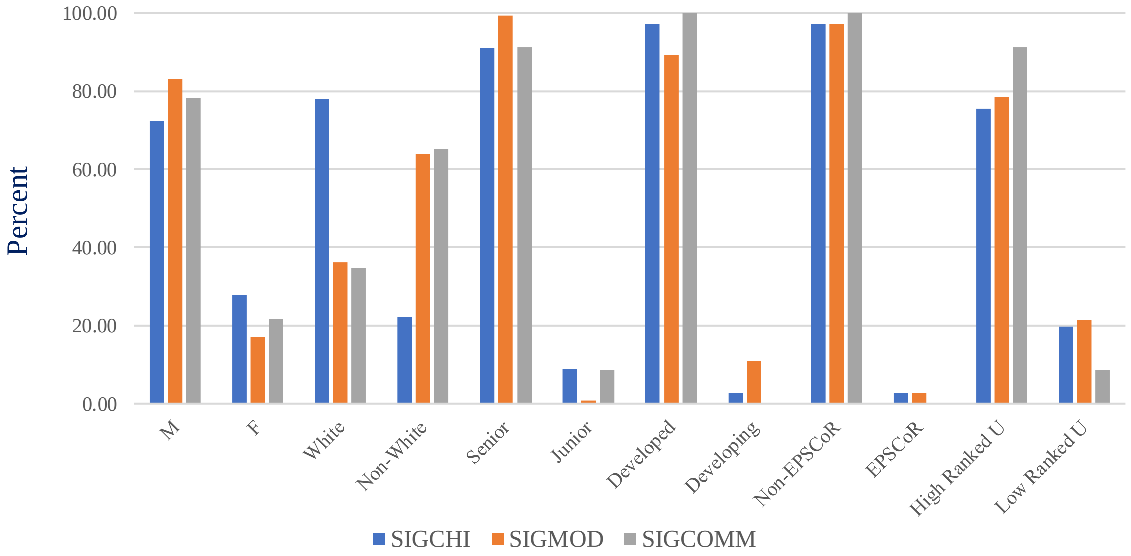


Figure 4.2: Demographic Distribution of all PC's Proposed by our Expertise Algorithm.

At the beginning of this section, we present the dataset's composition and the baseline data distribution to implement and evaluate our three algorithms. In the following subsections, we assess the effectiveness of the three methods being proposed in this work. The assessment is processed using the first four metrics described in section 4.2. Each algorithm produces two PCs: one is based on Boolean wight, and the other is based on continuous weight. Hence, we evaluate both by comparing the results of all algorithms using the mentioned metrics. We first explore the performance of our approaches when they derive PC's based on Boolean weight. Then, we apply the same process when they propose PCs based on continuous weight. Finally, we compare the Boolean and continuous weights representations, and the best one is promoted to the Hybrid approach.

4.3.1 Boolean Based Program Committees

In this section, we evaluate all PC's proposed by our algorithms based on Boolean weight. Our algorithms produce ranked list(s) from which we select to form the PCs with the overarching goal of increasing the program committee's diversity. Hence, we report the differences between the PC produced by the baseline, utility selection, and the PCs proposed by the algorithms described in Section 3. The main goal is to maximize diversity in the proposed PCs, and thus we evaluate our algorithms with regard to increasing diversity gain and minimizing utility loss. Looking at Figures 4.3, 4.4, and 4.5, we can see that all algorithms succeeded in increasing the recommended PCs' diversity in all demographic groups except race in some cases. Specifically, our MDA mostly produces the same number of non-white candidates as the baseline for SIGCOMM and SIGMOD 2017. Also, the MGA algorithm selects fewer non-white candidates than those chosen by the baseline for SIGCOMM 2017. In some cases, the UGA overcorrects, and in its efforts to choose diverse members, we end up with a demographically biased PC in favor of some protected groups, e.g., female.

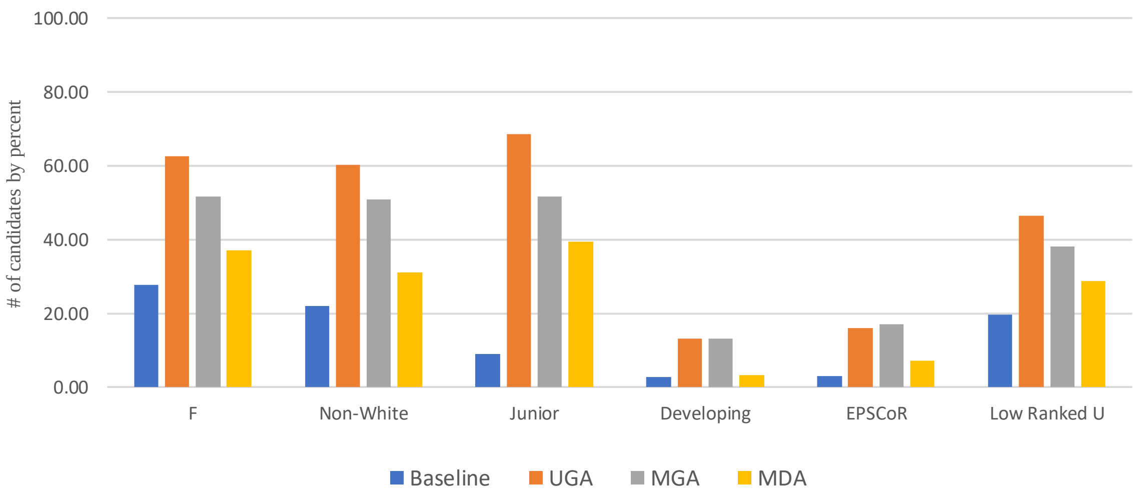


Figure 4.3: Comparison of the protected groups' improvement between the baseline PC and proposed PCs of SIGCHI 2017 produced by UGA, MGA, and MDA.

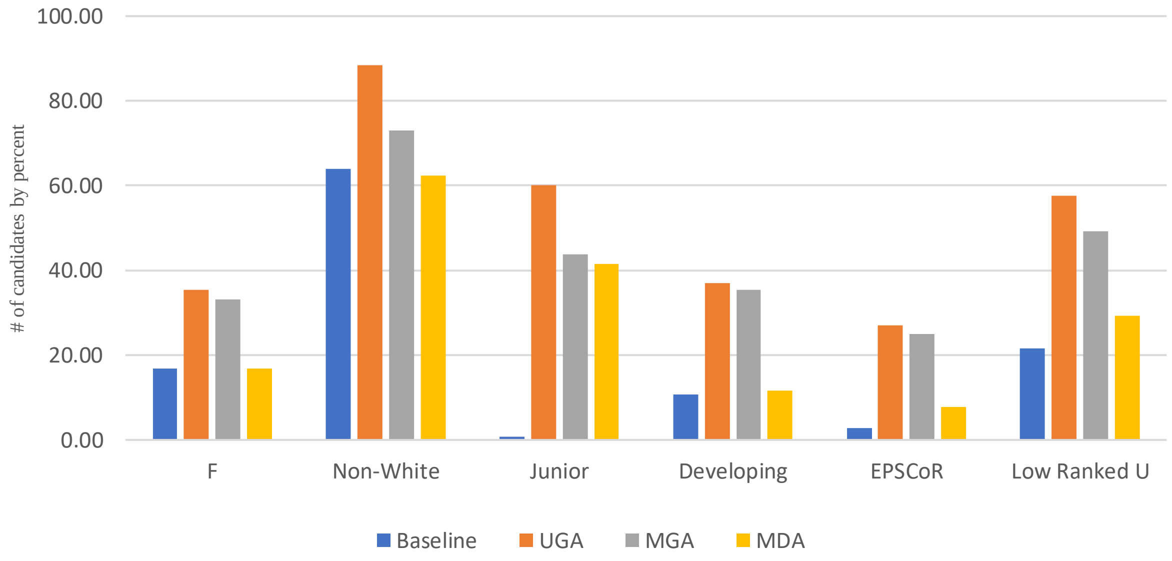


Figure 4.4: Comparison of the protected groups' improvement between the baseline PC and proposed PCs of SIGMOD 2017 produced by UGA, MGA, and MDA.

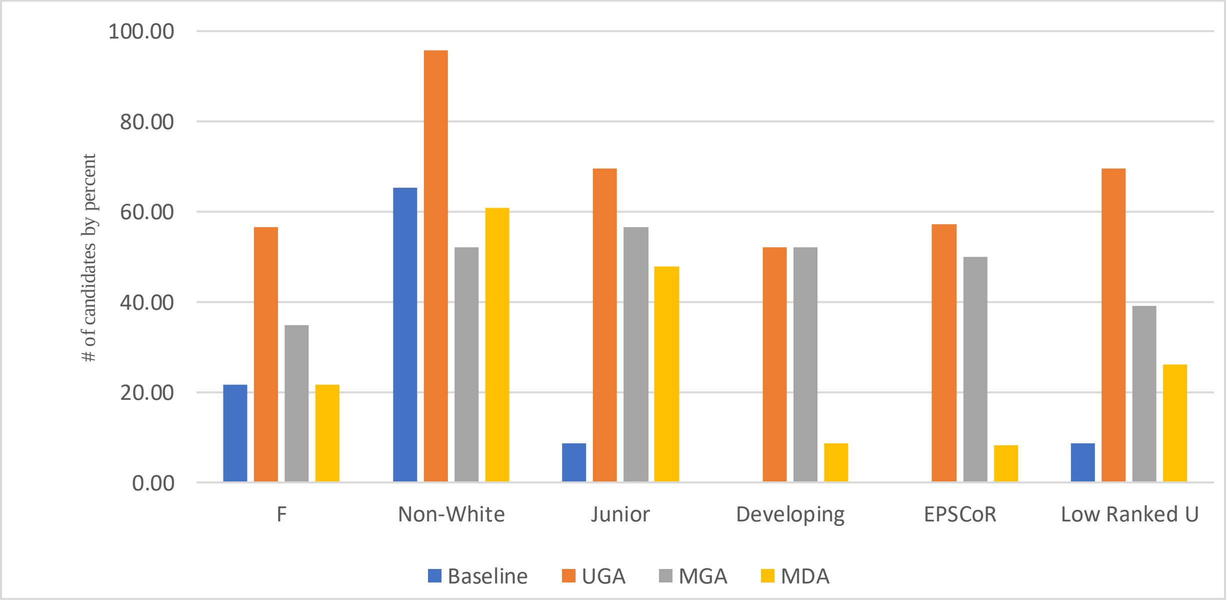


Figure 4.5: Comparison of the protected groups' improvement between the baseline PC and proposed PCs of SIGCOMM 2017 produced by our UGA, MGA, and MDA.

Table 4.2: Experimental results for the UGA, MGA, and MDA algorithms versus the baseline.

Table	D_G	UL_i	Υ_i	F
SIGCHI				
UGA	76.9	63.04	36.96	49.92
MGA	**83.86**	56.32	43.68	**57.44**
MDA	53.56	46.87	53.13	53.34
SIGMOD				
UGA	57.68	63.58	36.42	44.65
MGA	67.17	54.8	45.20	54.04
MDA	37.57	**44.09**	**55.91**	44.94
SIGCOMM				
UGA	34.41	80.29	19.71	25.06
MGA	54.59	68.19	31.81	40.20
MDA	66.49	56.66	43.34	52.48
Average				
UGA	56.33	68.97	31.03	39.88
MGA	**68.54**	59.77	40.23	**50.56**
MDA	52.54	**49.21**	**50.79**	50.25

We must also compare the effect of the algorithms with respect to the expertise of the resulting PCs. Table 4.2 summarizes the diversity gain (D_G), utility loss (UL), utility savings (Υ_i), and F-scores (F) for the PCs proposed by each algorithm. The F-score is measured by computing the trade-off between diversity gain and utility saving. MGA, UGA, and MDA obtained diversity gains of over 34.41% for all three proposed PCs, with the largest growth of 83.86% obtained by MGA for SIGCHI 2017. Those gains in diversity occur with an average utility loss of 68.97% for UGA, 59.77% for MGA, and 49.21% for MDA. We can see that MGA achieves the best balance between increasing diversity and utility loss with F scores of 57.44% for SIGCHI.

Fig 4.6 demonstrates the average results of our algorithms for our conference set. We can see that all algorithms successfully maximize diversity but lead to utility loss. Hence, we compute the average F score to determine which algorithm excellently balances the diversity increases and the utility loss. Fig 4.6 shows that MGA and MDA achieved F scores of 50.56% and 50.25%, respectively. Finally, for the three conferences, MDA produces PCs that mostly mirror the corresponding pools with an F score of only 4.34%.

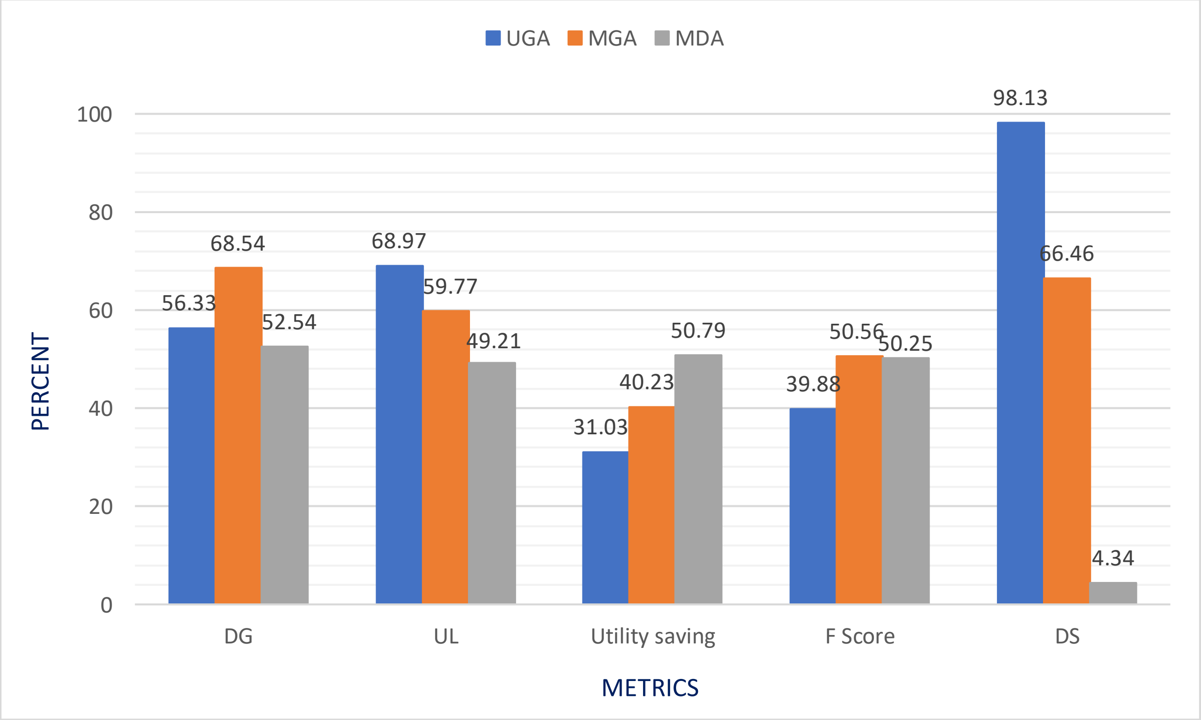


Figure 4.6: Boolean Average Results of UGA, MGA, and MDA.

4.3.2 Continuous Based Program Committees

In this section, we evaluate all PCs formed based on continuous weight. Similar to the previous section, we report the differences between the PC produced by the baseline and the PCs proposed by our algorithms. Figures 4.7, 4.8, 4.9 show that all algorithms maximized the diversity in all proposed PCs for SIGCHI except geolocation. For SIGMOD, the proposed PCs have a higher number of candidates from all protected groups except gender and race, in which MDA produced slightly fewer representatives than the baseline. Similarly, the proposed PCs by MDA have fewer female and non-white candidates than the baseline, as shown in Fig 4.9 for SIGCOMM.

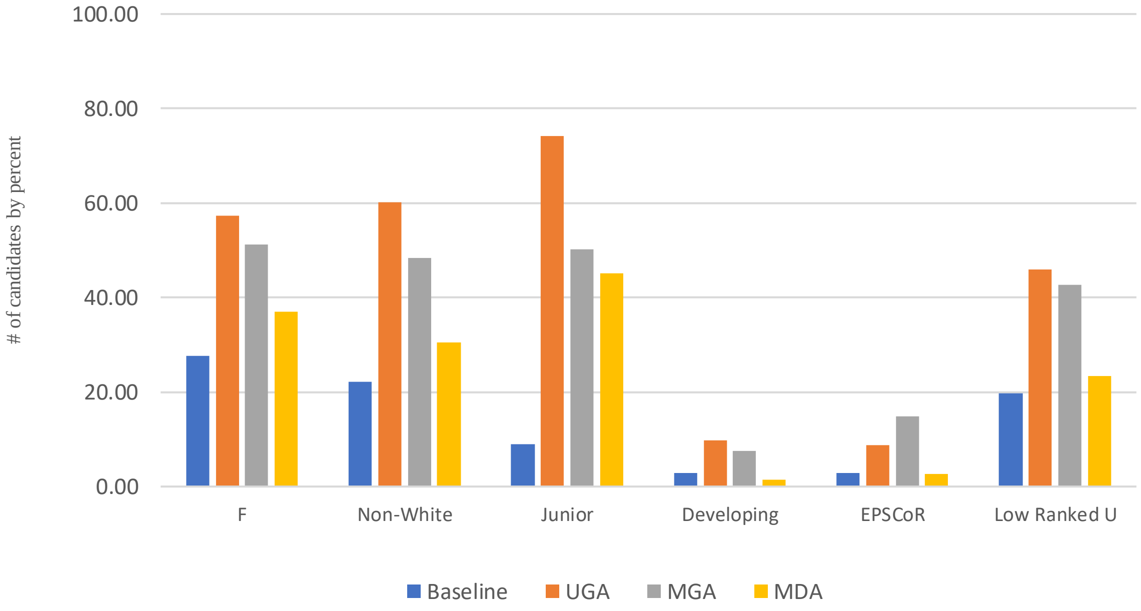


Figure 4.7: Comparison of the protected groups' improvement between the baseline PC and proposed PCs of SIGCHI 2017 produced by our UGA, MGA, and MDA.

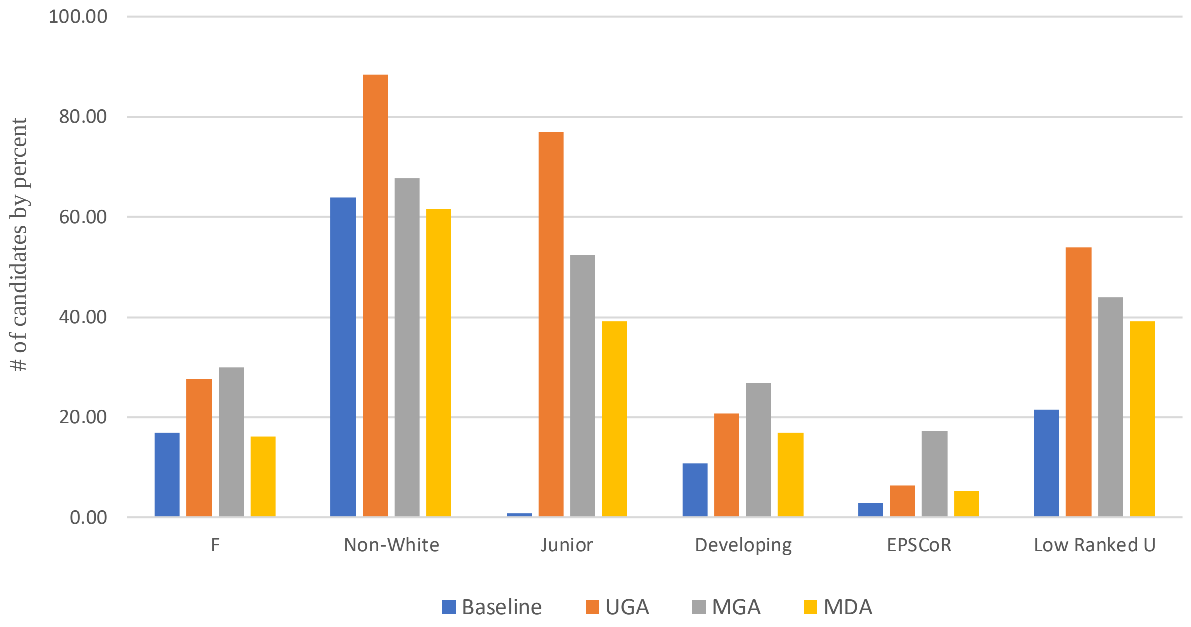


Figure 4.8: Comparison of the protected groups' improvement between the baseline and proposed PCs of SIGMOD 2017 produced by our UGA, MGA, and MDA.

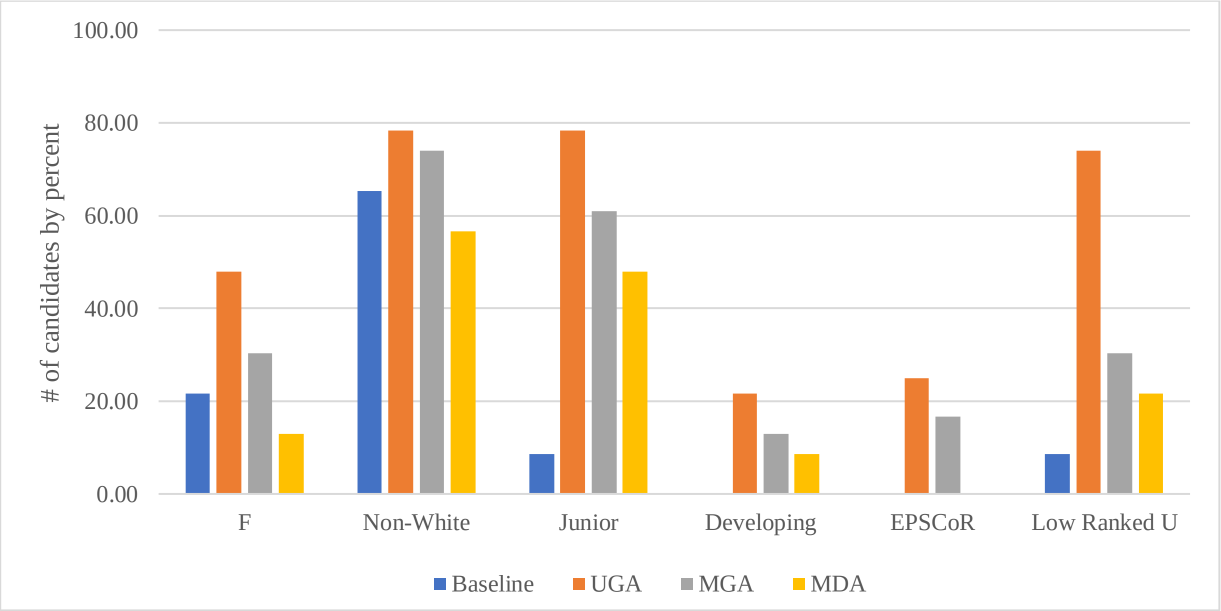


Figure 4.9: Comparison of the protected groups' improvement between the baseline PC and proposed PCs of SIGCOMM 2017 produced by our UGA, MGA, and MDA.

Table 4.3 summarizes the four metrics we use to evaluate all proposed PCs. MGA obtains the highest diversity gain with 85.03% for SIGCHI, and the lowest is 14.64% by MDA for the same conference. However, MDA's utility loss is the lowest for the same conference with 38.2%. This means MDA gets the highest utility saving. To illustrate the best algorithm that produces an excellent balance between utility saving and diversity gain, we investigate each algorithm's F score for the three conferences. Thus, MGA obtains the highest F-score with 57.49% for SIGCHI2017.

Table 4.3: Experimental results for UGA, MGA, and MDA algorithms versus the baseline.

	D_G	UL_i	Υ_i	F
SIGCHI				
UGA	69.33	68.27	31.73	43.54
MGA	**85.03**	56.57	43.43	**57.49**
MDA	14.64	**38.27**	**61.73**	23.67
SIGMOD				
UGA	56.24	71.06	28.94	38.22
MGA	68.71	53.07	46.93	55.77
MDA	49.78	50.89	49.11	49.44
SIGCOMM				
UGA	43.79	75.5	24.5	31.42
MGA	51.44	60.56	39.44	44.65
MDA	53.82	66.39	33.61	41.38
Average				
UGA	56.45	71.61	28.39	37.72
MGA	**68.39**	56.73	43.27	**52.64**
MDA	39.41	**51.85**	**48.15**	38.16

Fig 4.10 shows the three algorithms' average results of the three conferences to illustrate our algorithms' overall performance. As we mentioned earlier, the primary goal is to increase diversity while minimizing utility loss. Therefore, we focus on diversity gain (D_G), utility loss (UL), utility savings (Υ_i), and F-scores (F). We see that MGA maximizes the diversity in all PCs it produces with a diversity gain of 68.39%. The same algorithm provides the best balance between losing the utility and increasing the diversity with an F score of 52.64%. However, both UGA and MDA have an F score of 37.72% and 38.16%, respectively.

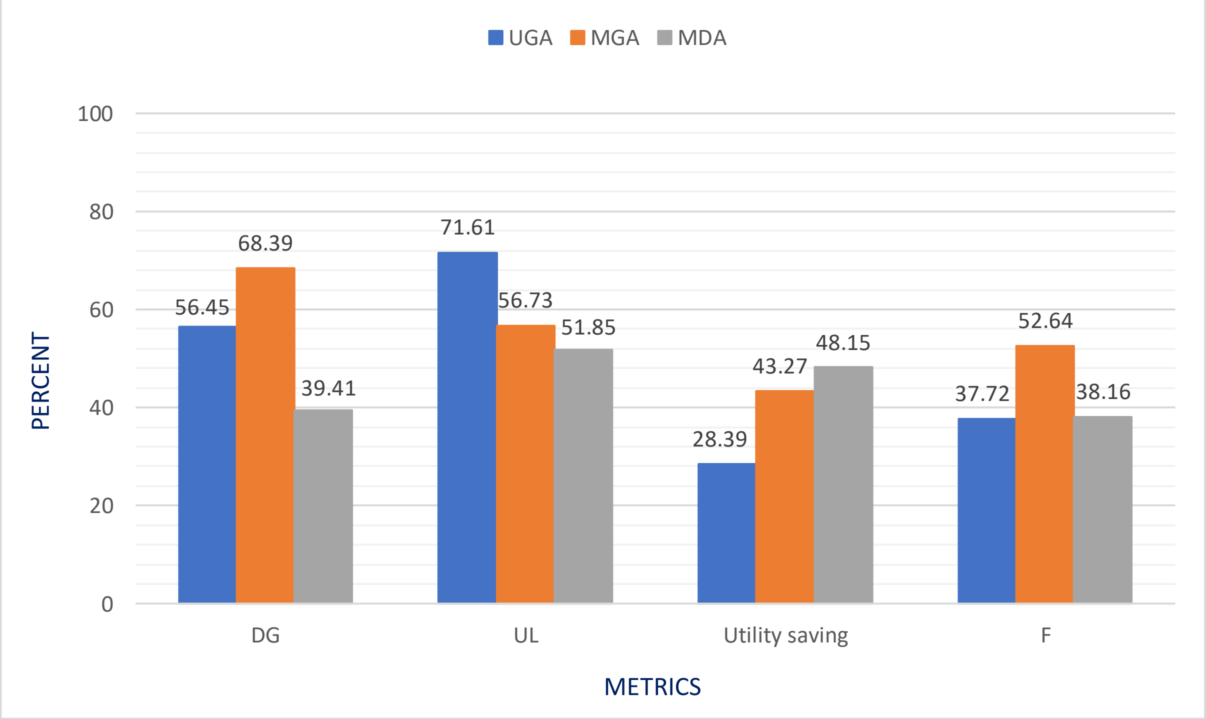


Figure 4.10: Continuous Average Results of UGA, MGA, and MDA.

4.3.3 Boolean-Continuous Comparison

In the previous two sections, we illustrate our algorithms' evaluation by comparing them to the baseline using diversity gain (D_G), utility loss (UL), utility savings (Υ_i) and F-scores (F) metrics described in section 4.2. Precisely, we assess all PCs proposed by those methods using Boolean weight, and then we apply the same process for those PCs produced using continuous weight. This section employs diversity gain, utility saving, and F-score to compare Boolean and continuous weight representation. Table 4.4 provides the average results of each algorithm with Boolean and continuous weights for the set of the three conferences. Out of the three approaches using Boolean weight, MGA gains the highest diversity percentage of 68.54% versus 56.45% for UGA, and 52.24% for MDA. However, those algorithms obtain lower diversity gain when they use continuous weight with 68.39% for MGA, 56.45% for UGA, and 39.41% for MDA. To conclude, we compute the three algorithms' average diversity gain using Boolean and continuous weights with 59.14% and 54.75%, respectively. This illustrates that our algorithms maximize diversity better when they utilize Boolean weight representation. For the utility saving metric (ϒ), we can see that our algorithms perform better when we employ Boolean weight with an average of 50.79 versus 48.15% when using continuous weight. Lastly, using Boolean weight leads to a better trade-off between diversity gain and utility loss. For instance, UGA obtains an F score of 39.88% when using Boolean weight versus 37.72% with continuous weight.

In conclusion, we can draw an overall evaluation between Boolean weight and continuous weight by computing the average results of all metrics when we use each weight representation. Therefore, Fig 4.11 shows that our algorithms perform better with Boolean weight for all metrics. For example, the three algorithms obtained an average F score of 46.9% with a Boolean weight versus 42.84% with continuous weight. Accordingly, we use Boolean weight representation to implement the Hybrid approach.

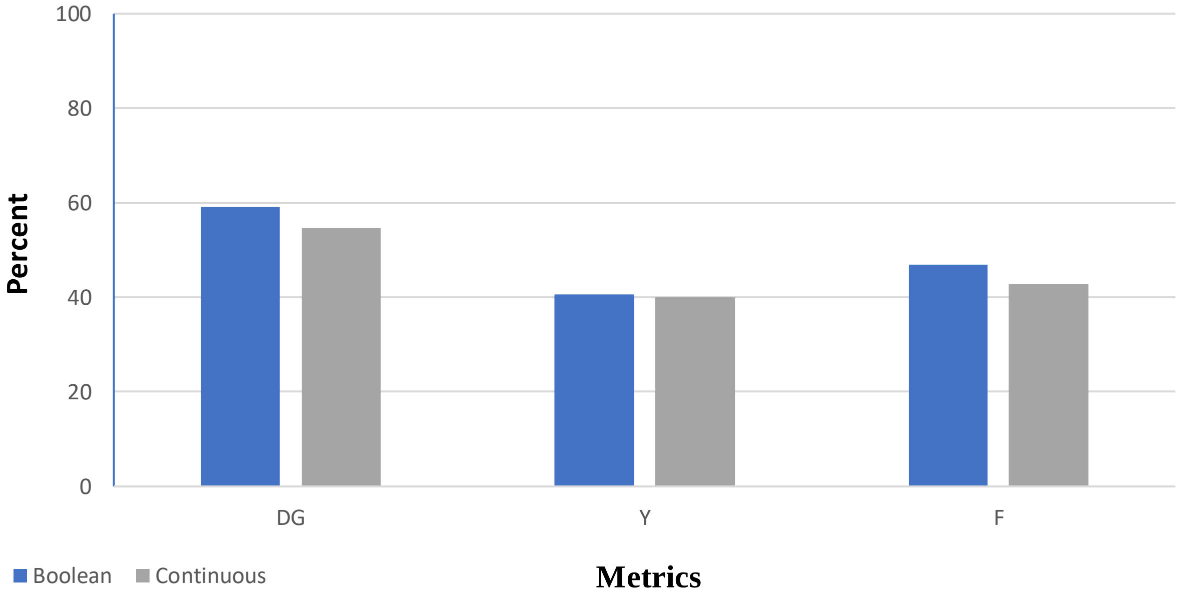


Figure 4.11: Weight representation assessment based on the average results of all PCs generated based on Boolean weight versus others developed using continuous weight.

Table 4.4: Comparison between Boolean weight and continuous weight representation.

	Weight	D_G	ϒ	F
UGA	Boolean	56.33	31.03	39.88
	Continuous	56.45	28.39	37.72
MGA	Boolean	68.54	40.23	50.56
	Continuous	68.39	43.27	52.64
MDA	Boolean	52.24	50.79	50.25
	Continuous	39.41	48.15	38.16
Average	Boolean	**59.14**	**40.68**	**46.9**
	Continuous	54.75	39.94	42.84

4.4 Experiment 2: Evaluating our Multidimensional Fair Algorithms

As described in section 3.3, our multidimensional fair algorithms combine the outputs of the expertise weights and the diversity vectors for each candidate when selecting group members using the following formula:

$$Score(H) = (1 - \alpha) * Score\ (E)' + \alpha * Score(D)' \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.4)$$

The parameter α governs the relative contributions of each candidate's level of expertise and diversity contributions. We test alpha values from 0.0 (diversity only) to 1.0 (expertise only) in steps of 0.1. The diversity gain, utility loss, distance similarity, and F-score are calculated as previously defined to determine our fair algorithm's best performance.

In the Hybrid approach, we incorporate each diversity algorithm with the expertise-based algorithm, and we employ the Boolean weight to represent the demographic features. Our main goal is to determine with alpha value best achieve demographic parity. Thus, we examine different alpha values for each algorithm to get the proper one that provides high distance similarity with a minimal utility loss. We evaluate which value obtains the best achievement of distance similarity and utility saving tradeoff using the F score metric. In the following three sections, we illustrate our Hybrid approach's performance for each algorithm we utilize to process it.

4.4.1 Univariate Greedy Algorithm with Expertise

Our UGA derives a queue contains selected candidates who are ordered based on the diversity score. In contrast, the expertise algorithm produces a queue that sorts candidates based on the expertise score only. Our Multidimensional fair algorithm unites both queues to a new queue that sorts the candidates based on the Hybrid score described in equation (3.6). We then use this score to select a desired number of candidates to form a PC. The metrics defined in section 4.2 are used to compare this PC to the baseline. For the three conferences, Table 4.5 summarizes the average of the diversity gain (D_G), utility loss (UL), utility savings (Υ_i), distance similarity (D_S), and F-scores (F) for all PCs proposed by UGA. The F-score presents the balance between the utility saving and the distance similarity saving. Accordingly, Table 4.5 shows that the highest diversity gain is obtained when alpha is 0.4 with 77.35%.

Our main goal is to obtain the best distance similarity while reducing utility loss. Thus, we graphically present the F-score for each alpha value (see Fig 4.12). In addition, we investigate which alpha value provides the best similarity distance (D_S) to achieve demographic parity. When we set the tuning parameter to 0.1, the figure shows that we get the best balance between the distance similarity and utility loss with an F-score of 62.57%. That means the Hybrid (UGA) with alpha 0.1 provides the best outcomes that achieve demographic parity with the minimal loss of expertise.

Table 4.5: Average findings of the three PCs produced by UGA versus the baseline for different values of alpha.

α	D_G	UL	ϒ	D_S	F-score
0.0	0.00	0.00	100.00	55.94	60.97
0.1	29.96	0.20	99.80	53.90	**62.57**
0.2	52.35	1.46	98.54	61.90	48.97
0.3	67.72	5.91	94.09	56.49	59.48
0.4	77.35	14.10	85.90	53.56	59.71
0.5	74.62	22.47	77.53	61.45	51.11
0.6	68.64	34.38	65.62	75.82	32.77
0.7	63.26	48.27	51.73	86.94	26.51
0.8	62	57.01	42.99	85.69	21.87
0.9	60.65	64.03	35.97	88.20	20.43
1.0	54.51	65.39	34.61	94.28	12.09

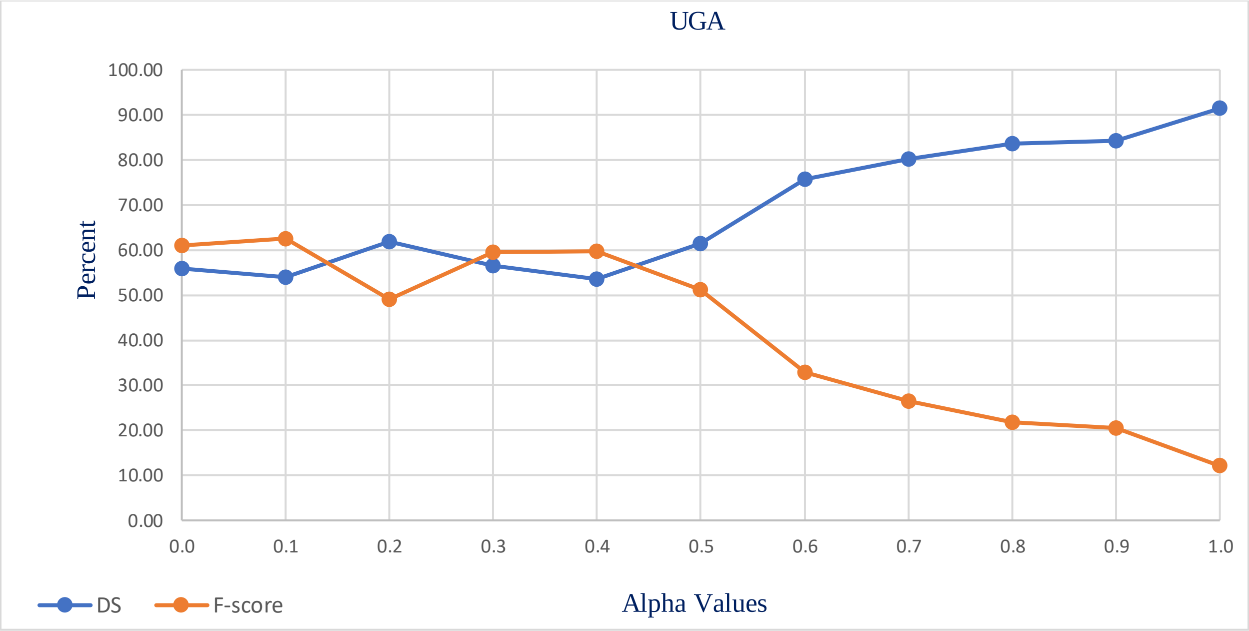


Figure 4.12: Comparison of F-score and D_S between UGA and the baseline.

4.4.2 Multivariate Greedy Algorithm with Expertise

In this section, we incorporate our MGA and expertise algorithms to implement the Hybrid approach. MGA selects candidates based on a multi-faceted ranking. On the other hand, the expertise algorithm selects candidates based on utility scores. Thus, we have two queues that our Hybrid approach combines into one queue and then compute the Hybrid score for all candidates. Once we get this score, we apply our fair algorithm for different alpha in a step of 0.1. Table 4.6 presents our findings for different values, which shows that the best performance of maximizing demographic parity with the minimal loss of expertise occurs when alpha is 0.2, with an F score of 66.4%. The alpha value that derives the best achievement of demographic parity is 0.4 because it obtains the shortest distance to the corresponding pools with a distance similarity of 45.78%, see Fig 4.13.

Table 4.6: Average findings of the three PCs produced by MGA versus the baseline for different values of alpha.

α	D_G	UL	ϒ	D_S	F-score
0.0	8.87	0	100	52.27	64.52
0.1	36.4	0.56	99.44	52.4	64.14
0.2	59.28	1.53	98.47	49.4	**66.4**
0.3	62.41	5.04	94.96	49.2	65.43
0.4	76.08	13.57	86.43	45.78	66.35
0.5	77.66	20	80	53.31	58.69
0.6	75.15	27.28	72.72	50.53	58.77
0.7	74.7	33.79	66.21	59.77	49.69
0.8	72.96	41.69	58.31	54.33	51.16
0.9	71.34	43.99	56.01	56.95	48.46
1.0	72.66	54.32	45.68	53.23	44.88

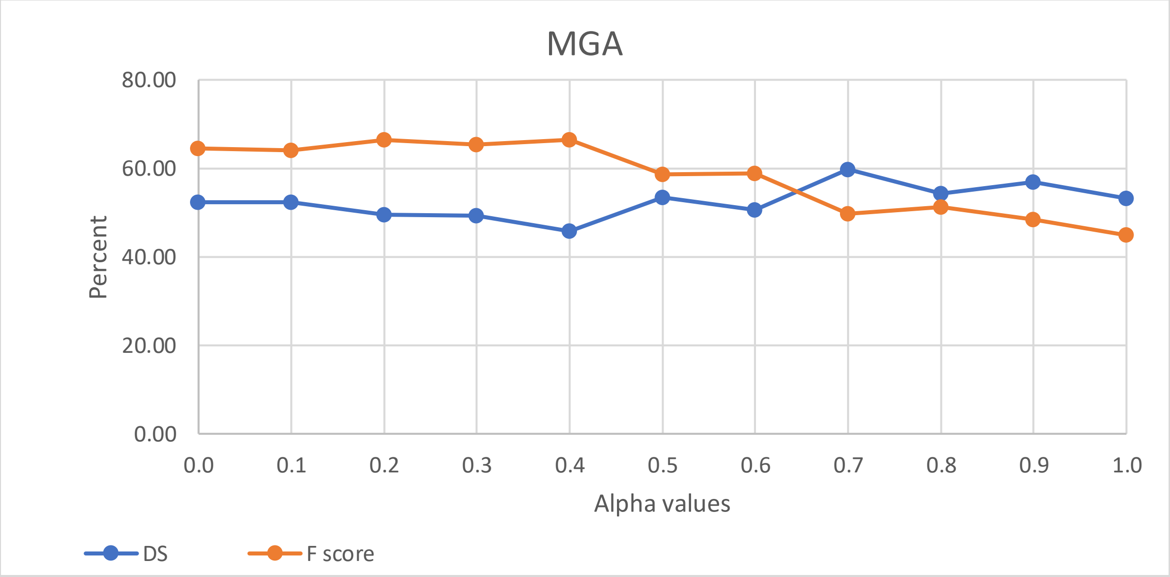


Figure 4.13: Comparison of F-score and D_S between MGA and the baseline.

4.4.3 Multidimensional Similarity Algorithm with Expertise

The previous two sections talk about how we employ our UGA and MGA to process the Hybrid approach. We consider the average F score for three conferences to represent the best tradeoff between the distance similarity and the utility loss. In addition, we investigate the diversity gain to determine which alpha provides the highest growth. This section applies the same method by combining the expertise and MDA algorithms for different alpha values. Table 4.6 shows that 0.5 alpha delivers the best balance between distance similarity and utility loss, with an F score of 73.19%. The alpha value that derives the best achievement of demographic parity is 1.0 because it obtains the shortest distance to the corresponding pools with a distance similarity of 27.88%, see Fig 4.13.

Table 4.7: Average findings of the three PCs produced by MDA versus the baseline for different values of alpha.

α	D_G	UL	ϒ	D_S	F-score
0.0	0.56	0.00	100.00	55.31	61.57
0.1	25.15	-20.45	99.88	56.55	60.08
0.2	34.30	-19.58	99.14	54.61	61.57
0.3	52.60	-15.75	96.16	50.18	65.06
0.4	59.76	-9.21	90.95	42.55	70.03
0.5	62.96	-2.57	85.34	35.32	**73.19**
0.6	65.67	1.90	81.19	33.19	73.06
0.7	68.83	29.36	54.81	44.31	43.74
0.8	69.03	7.87	76.38	44.8	62.88
0.9	65.84	7.87	76.38	35.33	69.80
1.0	63.88	22.64	64.38	**27.88**	67.78

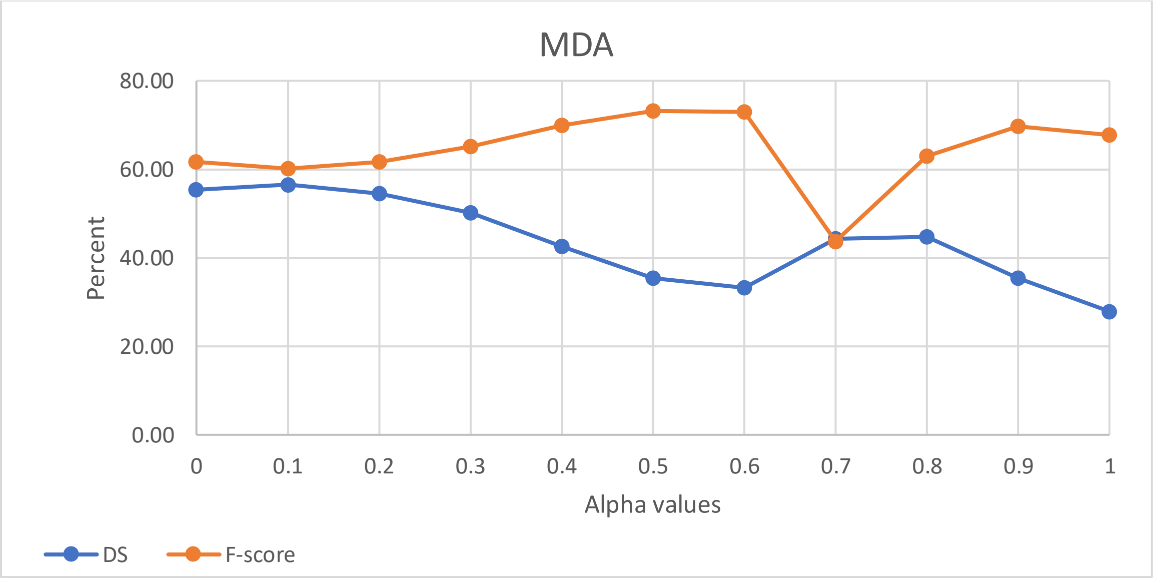


Figure 4.14: Comparison of F-score and D_S between MDA and the baseline.

To summarize our findings, our primary goal in this section is to bring off demographic parity. We focus here on the F score as it shows an accurate overview of which approach provides the best balance between the utility loss and distance similarity. We pick an alpha with the highest F score for each previous method to determine which one outperforms others. Fig 4.15 shows that the highest F score of 73.19% is associated with the Hybrid approach that incorporates baseline with MDA versus 66.4% of the Hybrid that uses MGA versus 62.57% of the one employed UGA. Thus, we use the third approach in the next section to compare it to the best algorithm of goal 1.

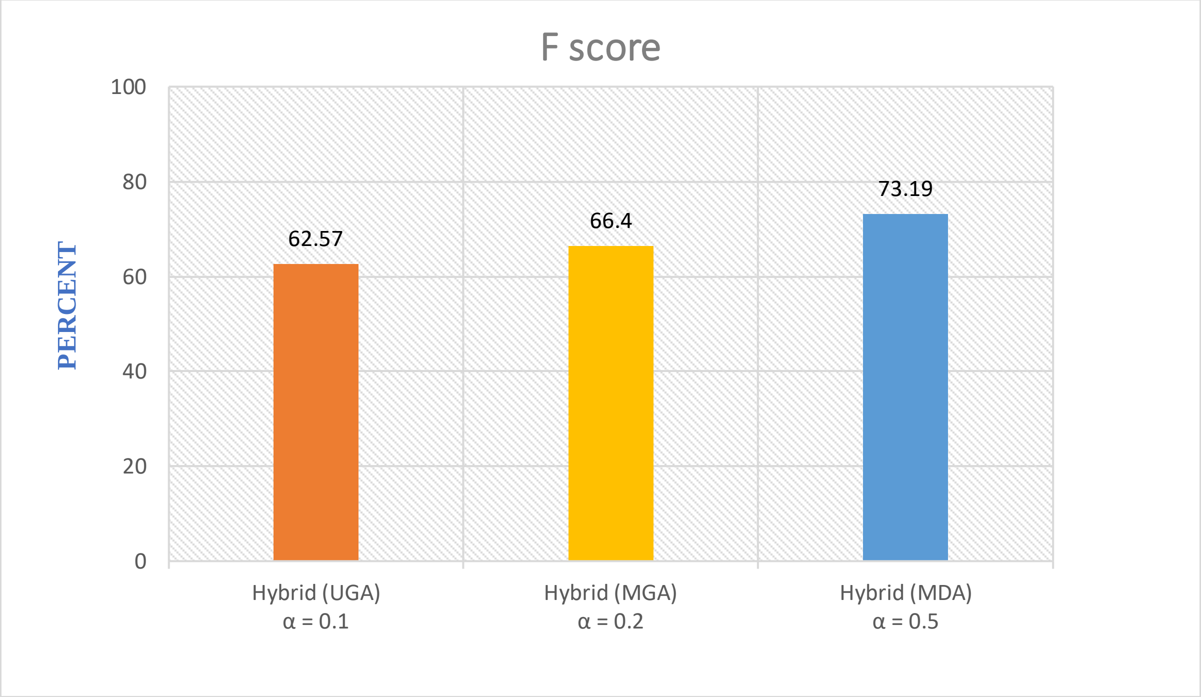


Figure 4.15: Comparison of F-score between all Hybrid approaches.

5 Validation

In this section, we validate our proposed methods by comparing their results to the actual PCs for the three conferences. Certainly, we compare the actual PCs for the three conferences to the PCs proposed by our best algorithm that maximizes diversity while preserving expertise, the MGA (see Table 5.1). In addition, we compare those actual PCs to those proposed by our best algorithm that best achieves demographic parity while preserving expertise, the Hybrid (MDA) (see Table 5.2).

For the first algorithm, MGA, the number of PC members from the protected groups was increased across all demographic features for all conferences. In most cases, the algorithm did not over-correct by including more than 50% of any protected demographic group, except for the participation of non-white that was increased to over 66.7%. The participation of females and junior researchers all increased by about 50%. Researchers from the developing world and EPSCoR states increased many-fold, although this was achieved by selecting all candidates from EPSCoR states and most candidates from developing countries. The avg utility for the proposed PCs across all conferences dropped 28.36%.

Table 5.1: Comparison between our MGA and the current PC's.

Feature	Method	SIGCHI 2017	SIGMOD 2017	SIGCOMM 2017	Average	% Increase	Avg Utility	Utility Increase
Female	Current	40.85	17.69	8.7	22.41	61.35	Current 28.95	-28.36
	MGA	48.83	29.23	30.43	36.16			
Non-White	Current	21.6	44.62	30.43	32.22	107.08		
	MGA	52.11	78.46	69.57	66.71			
Junior	Current	34.74	31.54	34.78	33.69	48.65		
	MGA	48.83	49.23	52.17	50.08			
Developing	Current	2.81	6.15	4.35	4.44	657.85	MGA 20.74	
	MGA	14.08	34.62	52.17	33.62			
EPSCoR	Current	5.71	6.78	0.00	4.16	185.59		
	MGA	7.51	10.77	17.39	11.89			
Low Rank University	Current	28.17	23.85	26.09	26.04	87.38		
	MGA	49.30	49.23	47.83	48.79			

Table 5.2 shows our comparison between the actual PCs and our Hybrid (MDA) algorithm. The number of PC members from protected groups was increased across all conferences, except for the junior group where our algorithm selects fewer junior candidates than those of the current PCs with a drop of 20.05%. In most cases, the algorithm did not over-correct by including more than 50% of any protected demographic group, except the participation of non-white that was increased to over 65.1%. In most cases, our algorithm selects almost the same total number of candidates as the current PCs which means it significantly achieves the demographic parity and

remarkably increases the utility in the proposed PCs comparing to the current PCs with an increase of 54.44%.

Table 5.2: Comparison between our Hybrid (MDA) and the current PC's.

Feature	Method	SIGCHI 2017	SIGMOD 2017	SIGCOMM 2017	Average	% Increase	Avg Utility	Utility Increase
Female	Current	40.85	17.69	8.7	22.41	34.62	Current 28.95	54.44
	MDA	46.01	22.31	26.09	30.17			
Non-White	Current	21.6	44.62	30.43	32.22	102.08		
	MDA	37.56	80.77	78.26	65.1			
Junior	Current	34.74	31.54	34.78	33.69	-20.05		
	MDA	31.46	20	39.13	26.93			
Developing	Current	2.81	6.15	4.35	4.44	95.49	MDA 44.71	
	MDA	4.23	15.38	8.7	8.67			
EPSCoR	Current	5.71	6.78	0	4.16	55.4		
	MDA	6.86	7.35	6.67	6.47			
Low Rank University	Current	28.17	23.85	26.09	26.04	17.89		
	MDA	39.44	33.85	21.74	30.69			

6 Conclusion

6.1 Summary

Groups of experts are formed in many situations within industry and academia. However, there may be bias in the traditional group formation process. This can lead to inferior results and also block members of underrepresented populations from access to valuable opportunities. To address this, researchers are developing new recommendation algorithms that try to provide demographic parity so that the resulting group mirrors the composition of the candidates from which it is formed. To this end, we investigate the issue of bias in academia, particularly the formation of conference program committees, and develop algorithms to form groups of experts that balance diversity and expertise. Our approach is based on representing candidate experts with a profile that models their expertise and demographic information. The expertise profile consists of a weighted ontology, and our diversity profile consists of a vector of five features that might be sources of bias, i.e., gender, race, career stage, geolocation, and university rank. Most previous work focuses on algorithms that guarantee fairness based on a single, Boolean feature, e.g., race, or gender, or disability. However, we consider the five demographic features at once. Those features are represented using Boolean and continuous weights.

We propose two main goals in this research: (1) the first goal presents our approaches that maximize diversity in the proposed groups; (2) the second goal shows how we achieve demographic parity in the resulted PCs. In the first one, we investigate three different approaches (UGA, MGA, and MDA) that incorporate the five demographic features simultaneously to build desire groups. Our methods increase diversity while minimizing utility loss with the best performance provides by our MGA with an F score of 50.56%. However, our fair algorithm integrates the expertise and demographic features to form the groups in the second goal. We combine the expertise approach with each diversity algorithm of goal one and then evaluate which one outperforms others in accomplishing demographic parity. As a result, our Hybrid (MDA) perform better than others, with an F score of 78.49%. In conclusion, our proposed research provides new ways to create inclusive, diverse groups to provide better opportunities and better outcomes for all.

6.2 Future Work

In upcoming work, we will explore new ways to form groups based on multi-valued demographic features. One possible way is to specify the range values of each feature applicable to be divided into different ranges, i.e., very high, high, middle, low, and very low. We may define only three parts of each diversity feature. Moreover, we will investigate and evaluate our approaches' performance when they exploit Boolean, Continuous, and multi-valued demographic features.

Bibliography

[1] V. C. Jeff Desjardins, ""Here are the countries and companies that spend the most on R&D,"," Business Insider, 16 10 2017. [Online]. Available: https://www.businessinsider.com/countries-and-companies-that-spend-the-most-on-rd-2017-10. [Accessed 06 03 2019].

[2] E. Comen, ""10 universities spending billions on R&D,"," MSN, 04 04 2017. [Online]. Available: https://www.msn.com/en-us/money/careersandeducation/10-universities-spending-billions-on-randd/ar-BBzjbN7..

[3] Casati, F., Giunchiglia, F., & Marchese, M., "Publish and perish: why the current publication and review model is killing research and wasting your money.," University of Trento., 2006.

[4] Hunt, V., Layton, D., & Prince, S., "Diversity matters," in *McKinsey & Company, 1, 15-29.*, 2015.

[5] AlShebli, B. K., Rahwan, T., & Woon, W. L., "Ethnic diversity increases scientific impact.," in *arXiv preprint arXiv:1803.02282.*, 2018.

[6] Khan, Beethika; Robbins, Carol; Okrent, Abigail, "Science and Engineering Indicator," 15 Jan 2020. [Online]. Available: https://ncses.nsf.gov/pubs/nsb20198/demographic-trends-of-the-s-e-workforce.

[7] Holman, L.; Stuart-Fox, D.; Hauser, C. E., "The gender gap in science: How long until women are equally represented?.," in *PLoS biology, 16(4), e2004956.*, 2018.

[8] Lerback, J.; Hanson, B., "Journals invite too few women to referee.," in *Nature News, 541(7638), 455.*, 2017.

[9] SIGCHI, "Diversity of the Program Committee for CHI 2020," 2019. [Online]. Available: https://chi2020.acm.org/blog/diversity-of-the-program-committee-for-chi-2020/.

[10] Deibel, K., "Team formation methods for increasing interaction during in-class group work.," in *In ACM SIGCSE Bulletin (Vol. 37, No. 3, pp. 291-295). ACM.*, (2005, June).

[11] Pociask, S., Gross, D., & Shih, M. Y., "Does Team Formation Impact Student Performance, Effort and Attitudes in a College Course Employing Collaborative Learning?.," in *Journal of the Scholarship of Teaching and Learning, 17(3), 19-33.*, (2017)..

[12] Yin, H., Cui, B., & Huang, Y., "Finding a wise group of experts in social networks.," in *In International Conference on Advanced Data Mining and Applications (pp. 381-394).*, Springer, Berlin, Heidelberg., (2011, December)..

[13] Wi, H., Oh, S., Mun, J., & Jung, M., "A team formation model based on knowledge and collaboration. Expert Systems with Applications, 36(5), 9121-9134.," 2009.

[14] Balog, Krisztian, et al., ""Expertise retrieval." Foundations and Trends®," in *in Information Retrieval 6.2–3 127-256*, 2012.

[15] Asudeh, A., Jagadish, H. V., Stoyanovich, J., & Das, G., ""Designing fair ranking schemes."," in *In Proceedings of the 2019 International Conference on Management of Data (pp. 1259-1276). ACM.*, (2019, June)..

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

[17] Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A., Brusilovsky, P., Kobsa, A., & Nejdl, W., ""The adaptive Web: methods and strategies of Web personalization."," (2007)..

[18] Balog, Krisztian, and Maarten De Rijke., "Determining Expert Profiles (With an Application to Expert Finding)," in *IJCAI*, 2007.

[19] Becerra-Fernandez, I., "Searching for experts on the web: A review of contemporary," in *ACM Transactions on Internet Technology (TOIT), 6(4), 333-*, 2006.

[20] Craswell, N., Hawking, D., Vercoustre, A., & Wilkins, P., "P@ noptic expert: Searching for experts not just for documents.," in *the Ausweb Poster Proceedings*, Queensland, Australia, 2001.

[21] Reichling, T., & Wulf, V., "Expert recommender systems in practice: Evaluating semi-automatic profile generation.," in *the SIGCHI Conference on Human Factors in Computing Systems, 59-68.*, 2009.

[22] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z., "Arnetminer: Extraction and mining of academic social networks.," in *Paper presented at the Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 990-998.*, 2008.

[23] Yin, X., Han, J., & Philip, S. Y., ""Object distinction: Distinguishing objects with identical names."," in *In 2007 IEEE 23rd International Conference on Data Engineering (pp. 1242-1246). IEEE.*, (2007, April)..

[24] Chandrasekaran, K., Gauch, S., Lakkaraju, P., & Luong, H. P., "Concept-based document recommendations for citeseer authors," in *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, Springer, Berlin, 2008.

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

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

[27] Hirsch, J. E., "An index to quantify an individual's scientific research output.," in *Proceedings of the National academy of Sciences, 102(46), 16569-16572.*, 2005.

[28] Success:, Measuring Research, "H-index Scores in Science Retrieved from," [Online]. Available: https://conductscience.com/measuring-research-success-h-index-scores-in-science/.

[29] Tang, L., & Hu, G., "Evaluation woes: Metrics can help beat bias.," in *Nature, 559(7714), 331-331.*, 2018.

[30] Bar-Ilan, J., "Which h-index?A comparison of WoS, Scopus and Google Scholar.," in *Scientometrics, 74(2), 257-271.*, 2008.

[31] Khalid, K., Salim, H. M., Loke, S. P., & Khalid, K., "Demographic profiling on job satisfaction in Malaysian utility sector.," in *International Journal of Academic Research, 3(4), 192-198.*, 2011.

[32] Cochran-Smith, M., & Zeichner, K. M., "Studying teacher education: The report of the AERA panel on research and teacher education Routledge.," 2009.

[33] Dias, T. G., & Borges, J., "A new algorithm to create balanced teams promoting more Diversity.," in *European Journal of Engineering Education, 42(6), 1365-1377. doi:10.1080/03043797.2017.1296411*, 2017.

[34] Michael, J., "40000 namen, anredebestimmung anhand des vornamens. C'T, 182-183.," 2007.

[35] Perez, I. S., "Gender-guesser. Retrieved from https://pypi.python.org/pypi/gender-guesser," 2019.

[36] Vanetta, M. (2019)., "Gender Detector Retrieved from: https://github.com/malev/gender-detector," 2019.

[37] Knowles, R., Carroll, J., & Dredze, M., "Demographer: Extremely simple name demographics.," in *Paper presented at the Proceedings of the First Workshop on NLP and Computational Social Science, 108-113.*, 2016.

[38] Ye, J., Han, S., Hu, Y., Coskun, B., Liu, M., Qin, H., & Skiena, S., "Nationality classification using name embeddings.," in *Paper presented at the Proceedings of the 2017*

ACM on Conference on Information and Knowledge Management, 1897-1906., 2017.

[39] Anagnostopoulos, A., Becchetti, L., Castillo, C., Gionis, A., & Leonardi, S., "Online team formation in social networks.," in *In Proceedings of the 21st international conference on World Wide Web (pp. 839-848). ACM.*, (2012, April).

[40] Brocco, M., Hauptmann, C., & Andergassen-Soelva, E., "Recommender system augmentation of HR databases for team recommendation.," in *Paper presented at the Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop On, 554-558.*, 2011.

[41] Lappas, T., Liu, K., & Terzi, E. (2009). (2009)., "Finding a team of experts in social networks.," in *Paper presented at the Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 467-476.*, 2009.

[42] Owens, D. A., Mannix, E. A., & Neale, M. A., "Strategic formation of groups: Issues in task performance and team member selection. Research on managing groups and teams, 1(1998), 149-165.," 1998.

[43] Juang, M. C., Huang, C. C., & Huang, J. L, "Efficient algorithms for team formation with a leader in social networks.," in *The Journal of Supercomputing, 66(2), 721-737.*, 2013.

[44] Giancarlo Fortino, Antonio Liotta, Fabrizio Messina, Domenico Rosaci, and Giuseppe M. L. Sarnè, "Evaluating group formation in virtual communities.," in *IEEE/CAA Journal of Automatica Sinica 7.4: 1003-1015.*, 2020.

[45] Ivanovska, S., Ivanoska, I., & Kalajdziski, S., "Algorithms for effective team building.," 2013.

[46] Chen, Y., Fan, Z. P., Ma, J., & Zeng, S., "A hybrid grouping genetic algorithm for reviewer group construction problem. Expert Systems with Applications, 38(3), 2401-2411.," 2011.

[47] Wang, D. Y., Lin, S. S., & Sun, C. T., "DIANA: A computer-supported heterogeneous grouping system for teachers to conduct successful small learning groups. Computers in Human Behavior, 23(4), 1997-2010.," 2007.

[48] Sukstrienwong, A., "Genetic algorithm for forming student groups based on heterogeneous grouping. In Recent advances in information science," in *Proceedings of the 3rd European conference of computer science (pp. 92-97).*, 2012.

[49] Tobar, C. M., & de Freitas, R. L., "A support tool for student group definition.," in *In 2007 37th Annual Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports (pp. T3J-7). IEEE.*, (2007, October).

[50] Stepanova, Elina, A. V. Rozhkova, and I. I. Grishina., "Team Building as a Method of Teaching Students and Group Cohesion.," in *20th European Conference on Research*

Methodology for Business and Management Studies: ECRM 2020. Academic Conferences and publishing, 2020.

[51] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S., "Certifying and removing disparate impact.," in *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 259-2,* (2015, August)..

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

[53] Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; Dwork, C., "Learning fair representations.," in *In International Conference on Machine Learning*, 2013.

[54] Luong, B. T., Ruggieri, S., & Turini, F., "k-NN as an implementation of situation testing for discrimination discovery and prevention.," in *In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 502-510). ACM.*, (2011, August)..

[55] Zhong, Ziyuan., "Toward Data Science.," Meduim., 21 10 2018.. [Online]. Available: https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb..

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

[57] Hardt, M., Price, E., & Srebro, N. ., ""Equality of opportunity in supervised learning."," in *In Advances in neural information processing systems (pp. 3315-3323).*, (2016)..

[58] Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P., ""Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. ."," in *In Proceedings of the 26th International Conference on World Wide.*, (2017, April)..

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

[60] Singh, Ashudeep, and Thorsten Joachims., "Policy learning for fairness in ranking.," in *Advances in Neural Information Processing Systems.*, 2019.

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

[62] Bornmann, L., & Daniel, H. D., "Selection of research fellowship recipients by committee

peer review. Reliability, fairness and predictive validity of Board of Trustees' decisions. Scientometrics, 63(2), 297-320.," (2005)..

[63] Chávez, Kerry, and Kristina MW Mitchell., "Exploring Bias in Student Evaluations: Gender, Race, and Ethnicity.," in *PS: Political Science & Politics 53.2 (2020): 270-274.*, 2020.

[64] Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B., "Bias in peer review.," in *Journal of the American Society for Information Science and Technology, 64(1), 2-17.*, (2013)..

[65] Holman, L., Stuart-Fox, D., & Hauser, C. E., "The gender gap in science: How long until women are equally represented?.," in *PLoS biology, 16(4), e2004956.*, (2018)..

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

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

[68] Baggs, H.G., Broome, M.E., Dougherty, M.C., Freda, M.C., & Kearney,M.H., "Blinding in peer review: the preferences of reviewers for nursing journals.," in *Journal of Advanced Nursing, 64(2), 131–138.*, (2008)..

[69] Justice, A.C., Cho, M.K., Winker, M.A., & Berlin, J.A., "Does masking author identity improve peer review quality? A randomized controlled trial.," in *Journal of the American Medical Association, 280(3), 240–242.*, (1998)..

[70] Lane, D., "Double-blind review: Easy to guess in specialist fields.," in *Nature, 452, 28.*, (Lane, 2008)..

[71] Rodriguez, M. A., & Bollen, J., "An algorithm to determine peer-reviewers.," in *In Proceedings of the 17th ACM conference on Information and knowledge management (pp. 319-328). ACM.*, (2008, October)..

[72] Haffar, Samir, Fateh Bazerbachi, and M. Hassan Murad., "Peer review bias: a critical review.," in *Mayo Clinic Proceedings. Vol. 94. No. 4. Elsevier.*, 2019.

[73] Wang, W., Kong, X., Zhang, J., Chen, Z., Xia, F., & Wang, X., "Editorial behaviors in peer review.," in *SpringerPlus, 5(1), 903.*, (2016)..

[74] T. Lane, "Diversity in Peer Review: Survey Results," COPE, 12 10 2018. [Online]. Available: https://publicationethics.org/news/diversity-peer-review-survey-results.

[75] Rybak, Jan, Krisztian Balog, and Kjetil Nørvåg, ""Temporal expertise profiling."," in *"*

European Conference on Information Retrieval., Springer, Cham, 2014.

[76] Google Scholar, "Google," . [Online]. Available: https://scholar.google.com/.

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

[78] NamSor, "NamSor," 2020. [Online]. Available: https://www.namsor.com/. [Accessed 2020].

[79] blog, NamSor, "Inferring The World's Gender and Ethnic Diversity using Personal Names.," 31 Jan 2018. [Online]. Available: https://namesorts.com/2018/01/31/understanding-namsor-api-precision-for-gender-inference/. [Accessed 2020.].

[80] Union, European, "The official directory of the European Union," 2020. [Online]. Available: https://op.europa.eu/en/web/who-is-who. [Accessed 2020].

[81] Times, . [Online]. Available: https://www.timeshighereducation.com/.

[82] Khan, Beethika; Robbins, Carol; Okrent, Abigail, "Science and Engineering Indicator," 15 Jan 2020. [Online]. Available: https://ncses.nsf.gov/pubs/nsb20198/demographic-trends-of-the-s-e-workforce.

[83] Worldbank, "gdp-ranking," 2018. [Online]. Available: https://datacatalog.worldbank.org/dataset/gdp-ranking. [Accessed 2020].

[84] Jardine, N., & van Rijsbergen, C. J., "The use of hierarchic clustering in information retrieval.," in *Information storage and retrieval*, 1971.