7-2021

Promoting Diversity in Academic Research Communities Through Multivariate Expert Recommendation

Omar Salman

University of Arkansas, Fayetteville

Promoting Diversity in Academic Research Communities
Through Multivariate Expert Recommendation

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

by

Omar Salman
Nahrain University
Bachelor of Science in Information Engineering, 2005
Nahrain University
Master of Science in Information Engineering, 2011

July 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

Mark Arnold, Ph.D.
Committee Member

David Andrews, Ph.D.
Committee Member

Abstract

Expert recommendation is the process of identifying individuals who have the appropriate knowledge and skills to achieve a specific task. It has been widely used in the educational environment mainly in the hiring process, paper-reviewer assignment, and assembling conference program committees. In this research, we highlight the problem of diversity and fair representation of underrepresented groups in expertise recommendation, factors that current expertise recommendation systems rarely consider. We introduce a novel way to model experts in academia by considering demographic attributes in addition to skills. We use the *h*-index score to quantify skills for a researcher and we identify five demographic features with which to represent a researcher's demographic profile. We highlight the importance of these features and their role in bias within the academic environment.

We utilize these demographic features within an expert recommender system in academia to achieve demographic diversity and increase the exposure of the underrepresented groups using two approaches. In the first approach, we present three different algorithms for scholar recommendation: expertise-based, diversity-based, and a hybrid algorithm that uses a tuning parameter to calibrate the balance between expertise loss and diversity gain. To evaluate the ranking produced by these algorithms, we introduce a modified normalized Discounted Cumulative Gain (nDCG) version that supports multi-dimensional features, and we report diversity gain from each method. Our results show that we can achieve the best possible balance between diversity gain and expertise loss when the tuning parameter value is set around 0.4, giving nearly equal weight to both expertise and diversity.

Finally, we explore diversity from the lens of the demographic parity and develop two algorithms to achieve a representative group that reflects the demographics of the recommendation

pool. One is inspired by *Hill Climbing*, a mathematical optimization technique, wherein a solution is built gradually to the problem, and the other one is inspired by the problem of *seat allocation* in electoral voting systems. We evaluated these algorithms by comparing them to the hybrid algorithm from the previous approach. Our evaluation shows that both approaches provide a better diversity gain as compared to the hybrid algorithm. However, Hill Climbing Diversity is more effective when it comes to expertise savings with a statistically significant result, making it the preferred algorithm to achieve the goal of promoting diversity while maintaining expertise in an expert recommendation process.

Acknowledgments

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

First and always, I want to thank Allah for giving me all my strengths and guiding me through the tough times during this research.

To my family: Thank you for supporting me and being there whenever I need you. I am grateful for your understanding and patience because I have been away for several years and missed many events.

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

To my committee members, Dr. Brajendra Panda, Dr. David Andrews, and Dr. Mark Arnold: Thank you for your important and useful comments. It has been an honor to present my research and discuss it with you.

Dedication

To my grandmother, Wajeeha

.

Table of Contents

Contents

List of Tables

List of Figures

1 Introduction

We are witnessing a great change in the amount of created content. [1]. The introduction of social media, blogs, the internet of things, and knowledge-sharing communities have increased the amount of available knowledge online. According to a study carried by IBM, it is estimated that the amount of knowledge doubles twice each day [2]. This has led modern economies to shift to knowledge-based economies where the intellectual capabilities and expertise of the people determine their values in their enterprise and society [3]. However, determining the level of expertise of a person is a major challenge because it is quite difficult to assess the amount of knowledge that people carry in their minds [4]. Hence, enterprises and companies start relying on documenting people's expertise thru centralized databases that they often called "yellow pages, expert locator systems, or expertise management systems" [5]. A major drawback of these systems is that they are prone to provide inaccurate information because it relies on the input from the employees themselves without enough judgment on the accuracy of the provided skills. Also, it requires manual efforts to update the profile of the person which might not always take place. This has made it necessary to develop an intelligent tool that helps automate the process of creating and updating expert profile and finally led to the introduction of expert recommendation systems.

Expert recommendation is the process of recommending an expert that has the required knowledge to achieve a predefined task. Different expert recommendation systems have been developed in the last decades. These systems were mainly depending on the written artifacts of the experts to determine their expertise. For example, early systems consider internal email communications inside an organization to build profiles for their experts. Another approach considers all the documents of the enterprise to represent the skills of employees. Within the last approach, the process looks similar to the document retrieval system, and thus the process received great attention from the information retrieval community. Based on that, The Text REtrieval Conference (TREC) has identified two models for expert recommendation systems that are: *profile model* and *language model*. In the *profile model*, a profile is built based on the text that is authored by the expert or a text that contains the name of the expert. However, *in the language model*, it first gets the relevant documents that contain the required expertise and searches for the experts that are associated with these documents [4]. A later approach developed by Chandrasekaran et. al [6] proposed a weighted ontology-based approach that developed a profile based on a limited set of keywords that clearly describe the skills of that domain. In our research, however, we consider a different approach that is using a bibliometric measure to quantify a set of expertise of an expert.

Expert Recommendation Systems are not only limited to industry. People in academia have used this in the hiring process or finding reviewers or assembling a conference program committee. In the literature review, we present different expert recommendation systems that have been used in academia. These systems have provided great efforts to automate the process of expertise finding. However, different challenges have led people in academia to rely on manual efforts to find experts or reviewers to join their team. A list of key challenges is:

- The accuracy is always an issue since these systems rely on textual data of the papers to build the profile for a researcher and not on keywords that precisely describe the expertise of a researcher.
- The recommendation algorithms have always had strong preferences toward well-known researchers because most of these algorithms only consider citation count to recommend an expert [7], hence young and junior researchers will always get a lower rank and have low chances to be recommended by this algorithm.

- These systems do not address the issue of gender and race gap and the need to have a diverse team. Even some efforts have been done by NSF in this regard [8]. The gap is still available.
- The recommendation algorithms do not address the issue of the demographic parity and how to produce a diverse recommendation that reflects the representation of the population of items in a recommendation pool.
- Some of these algorithms produce biased results as they have been developed using machine learning techniques where if there is a bias in the training data, we will end up having bias in the final recommendation.

To address these challenges, we present research that aims to investigate these issues and develop algorithms that provide an accurate and fair recommendation. We test different demographic and expertise attributes to recommend experts by combining Machine Learning, Expertise Retrieval, and Information Retrieval Techniques. We set three goals that aim to answer the following questions:

- How to develop an expert profile that reflects both her or his expertise and demographics?
- How can we maximize the participation of the underrepresented groups in the recommended items?
- How to produce a set of representative items in any recommendation rank that reflects the demographics of the population of items.

To achieve goal 1, we present our method to build an expert profile in academia. We propose two main components of this profile: expertise and demographic. We use *h*-index [9] to quantify the amount of expertise based on *h*-index score provided by Google Scholar (GS). The demographics of a researcher have been modeled based on five parameters that we found that they are major sources of bias in academia. We use the expert profiles from Goal 1 to develop algorithms in Goal 2. We propose three algorithms in this goal: the first is the baseline that only considers the expertise score from the expert profile to build a recommendation list of experts. The second considers the diversity that is the result of the combination of different demographics to recommend a researcher. The last is a hybrid approach that tries to achieve a balance between the first two approaches by maximizing the team diversity and minimizing the utility loss that comes from recommending a diverse group member over a homogenous team member.

In Goal 3, we explore another definition for diversity that is *diversity by proportionality*, where we develop two algorithms that address the issue of demographic parity and how to achieve a representative recommendation that reflects the demographics of a recommendation pool. With this design concept, we assume that we can achieve group fairness by ensuring a representation of a demographic segment matches the one in the recommendation pool. Also, we will ensure individual fairness within the same demographic segment by ensuring that no individual with less expertise would be ranked higher than the one with a greater amount of expertise. We conclude by comparing these two algorithms to the hybrid algorithm from Goal 2 to find the best algorithm that achieves this goal of forming a diverse and fair team with the best expertise savings.

2 Related Work

2.1 Expert Modeling

The process of recommending an expert starts with modeling that expert and hence this topic has been well studied in the IR community. Expert modeling or expert profiling usually refers to measuring the expertise of an expert without paying attention to other profile types such as social profile. In this research, we will consider two main profiles to model an expert that are: expertise profiling and demographic profiling. The following sections provide some insight into some of the work that has been done in this regard.

2.1.1 Expertise Profiling

Expertise profiling has many definitions. One accepted definition by Balog and de Rijke in [10] that is “a record of the types and areas of skills and knowledge of that individual, together with an identification of levels of “competency” in each area". Such a profile can be represented as a vector of scores that show the competency of each skill for that expert. Moreover, in an organization, where many people have different skills, a skill matrix is used. This matrix can be filled manually; however, such method is prone to reliability and scalability problems [11]. Hence, automated methods that collect skills based on user's or organization's documents have been suggested to handle such challenges.

An early attempt to automate expert profiling was proposed by Craswell et al. who developed a system called “*P@noptic expert*” [12]. In their approach, the profile consists of keywords determined from documents that contain or associated with that expert in an organization. Also, their system provides a search facility that allows users to find experts based on needed expertise. The methodology adopted by this system is similar to that of a traditional search engine in which a set of documents (experts in this case) is retrieved based on a submitted query. This illustrates the interest by the IR community in expertise profiling and recommendation. This interest has been translated into two further approaches to modeling the expertise of an expert. The two models have been built through generative language approaches and given a name of a *candidate-based* approach and *document* approach. The *candidate-based* approach is similar to *P@noptic expert* system where the profile is created based on the textual data of documents associated with a specific expert while the *document* approach is to cluster documents around a topic and then get all the experts that are associated with these documents. Different studies show that the *document-based* approach is more effective than the first approach [4]. However, these two models still do not precisely describe the skills of experts as it lacks the standard terms to describe the expertise of an expert.

Reichling and Wulf in [13] describe a system inspired by TREC models that aims to provide accurate expertise profiling through getting expert feedback and update the system if needed. The system is considered a semi-automatic because it depends on getting documents selected by the user and requires a user's evaluation and update. Although such approach enhances the accuracy of the expert profile, a major drawback is that it still requires manual efforts and update by the expert which is not always taken place. Chandrasekaran et. al in [6] develop a profile that is built by depending on ACM's Computing Classification System (or CCS) taxonomy. This profile is used to recommend papers from CiteSeer^x database to those authors of Citeseer^x papers. Although this was not explicitly focused on expertise profiling, we argue that this work has addressed the problems of accuracy in the expertise profiling by depending on a set of controlled vocabulary that describes the expertise of such an expert.

The interest in expertise profiling is not only limited to industry but it is also relevant to academia due to the demand for experts to review papers, join a program conference committee, or finding talented resources to join research teams. For example, Zhang et. el. [14] developed an expert profiling, finding, and document retrieval system called "Arnetminer". In this system, an expert profile is created by crawling personal home page of and publications of researchers. They evaluated their system by comparing it to other models and used human judgment to evaluate the result. Based on their evaluation, the proposed system outperformed other systems in terms of precision and recall. Another example is INDURE (INdiana Database of University Research Expertise) project, which is an academic expertise profiling and finding system which aims to provide an expertise database for four universities in the State of Indiana. Their approach was to build a profile based on the ontology of the National Research Council and it allows users to search and browse the expert with respect to that ontology [4]. Sateli et al. [15] created "ScholarLens" which is a system that generates a semantic expert profile based on scholar publications. The difference from Arnetminer is that they incorporated semantic representation for the user expertise by using some NLP and Open Linked Data techniques. In their approach, they developed two methods to extract the expertise of a researcher: by considering the full text of their publications, or by extracting text within the words "claim" or "contribution". They used name entity recognition and LOD (Linked Open Data) cloud to represent the relative topics to a specific domain. For example, " Data Structure", and "Algorithms" are both under computer science field. To evaluate their model, they use human judgment and both approaches provide a high precision with 10 retrieved expertise; however, the performance of their second approach degrades as the number of the retrieved competences increased.

One of the attempts to measure researcher's work's quality and productivity is the *h*-index metric, introduced by Hirsch in [9]. This metric computes a score based on a total number of publications and number of citations. It implicates the scholar's publications' quality, and it is utilized by research and higher education institutions in recruiting, awarding committees, and funding decisions [16][17]. Furthermore, many scholarly databases such as Google Scholar, Web of Science, and Scopus employ this metric to determine the quality of researcher's work. Hence, in our research, we use *h*-index as provided by Google Scholar to quantify scholar expertise since it represents a precise score and offers more excellent coverage for computer scientists [18].

2.1.2 Demographic Profiling

The definition of demographic profiling varies from system to system [19]. One widely accepted definition within academic papers is that a demographic profiling means the status of an individual from gender, race, ethnicity, socioeconomic background, and age [20]. It is important to note that is not always the case that a profile includes demographic information due to fear that such data could be used to discriminate against people. However, different organizations and researchers understand the importance of having this data in an expert profile. In fact, such data can be used to provide fairness and anti-discrimination measurements when recommending an expert in these organizations. In this review, we will focus on how to predict demographic parameters like gender, ethnicity, and nationality from a given name and how can we incorporate this information in an expert profile.

Different models have been suggested to predict gender and ethnicity. Earlier models used web scraping techniques as in [21] where user home pages are crawled and analyzed to get the demographic information and then incorporate them in a user profile that is used by demographic recommender systems. The current models use machine learning techniques to classify gender and ethnicity based on a provided name. A limitation can be drawn on these models is that its accuracy is strongly correlated with the trained dataset and hence it might perform well on some names and poorly with the others. For instance, Michael [22] uses a list that contains 45,513 names and how popular is that name by country to determine the gender based on name. Perez [23] uses the same data with a different approach to enhance the accuracy and the coverage. Vanetta [24] uses data as collected by Open Gender Tracking's Global Name Data project that covers the names from four different countries to determine the gender from the first name. Knowles et al. [25] use a similar method, but with an SVM classifier and the entire model outperformed other tools when tested on twitter dataset.

Recently, Ye et al. [26] developed an alternative approach by building a classifier system called "*NamePrism*" that predicts name, ethnicity, and nationality using homophily in the communication pattern by analyzing the data of 57M customers of a major internet service provider. Strømgren [27] predicts gender by using different collected profiles from several social networks. The tool is accessed via a web API, and results include gender, probability, and confidence expressed as a count. According to the website distinct names from 79 countries and 89 languages. Carsenat proposed a tool named *NamSor* in [28], which is a name recognition software that uses onomastics and some machine learning techniques to classify gender and race based on a given name using a database of 4 billion names. Due to greater coverage and accuracy as compared with other models [29], we consider *NamSor* to predict gender and ethnicity in our research.

2.2 Bias in Expert Recommendation

It is common that we need an expert whenever a missing skill is required. This gap comes either from an individual or a team has new roles or a new task that demands new skill to be developed.

Hence, a process to recommend an expert is required. Traditionally, this process has been done manually through human judgment and only more recently has research focused on developing fully automated expert recommendation. Today, different systems, algorithms, and practices have been developed to recommend or choose an expert. However, we still see a bias in the recommendation and selection process. Such kind of bias can be due to humans, systems, or algorithms that are supposed to select the best experts to join a team. To address this issue, we see it is important to provide some insight from the literature in this section. We explore the literature from three main perspectives: first, we will present some research that discusses the issue of group formation and how it has been developed and will see how it fails to address the issue of fixing the diversity in the team formation process. Moreover, we will discuss this from an algorithmic preceptive and see the attempts to fix bias and how algorithms can be developed to provide some bias countermeasures. Then, we will end up exploring some studies that discuss the problem of bias in academia providing some statistics and real-world cases that show the highlight the importance of this issue in our society.

2.2.1 Group Formation

The problem of group formation has been studied in management, sociology, economics, and computer science literature. For example, in [30] Brocco et al. propose a team recommender system for an enterprise by utilizing human resource databases. Mathieu et al. [31] discuss how team composition can affect the outcomes of the team. They identified four general types for team composition. They argued that the performance of the team can be affected by the strongest or the weakest person in that team. Rabanca [32] investigated team formation from three aspects that are: team expertise, team cohesion, and team size and how this can impact team performance. He suggests an approximation algorithm to optimize each aspect in a different environment. His algorithm has been tested on DLPB dataset and it turns out that the connection to the team key influencers is the most important factor that determines team success. Anagnostopoulos et al. [33] presented a model to build a team of experts tasked to achieve a predefined goal by considering three main perspectives: the amount of expertise to achieve that task, the cost, and the load balance among team members. Chhabra et. al [34] propose an algorithm to recommend a team of experts for a set of tasks based on their skills and how strong they are socially connected. Likewise, Owens et al. [35] argue that the quality of the relationships between team members can increase to the team's productivity. However, our research does not consider the social relationships that form within a group because considering this kind of connection could lead to bias in the recommending process.

Neshati et al. [36] discuss the problem of forming a group of experts for a multi-aspects task. They proposed a framework that optimizes the coverage of the required skills for a given task and they address how a group can be formed while having a limited number of skills for each expert in the team. They argued that focusing on the quality of collaboration is vital to create a productive team. Lappas et al. [37] were the first who discussed the problem of group formation in social networks from a computational perspective. They presented an algorithm to recommend a team of experts based on how their expertise and how strong their social connection. In contrast to their work, our work will not consider the social connection among the experts in recommending a person to a task as we assume that such kind of connection could lead to a bias in the selection process. Also, in our work, we focus on one aspect of group formation that is adding a member to an existing group. We believe that focusing on this side is important because in most cases we do not need to form a complete group from scratch, but we may have an existing team and we need to recruit an expert to fill the gap in the needed expertise.

Another important reason is that our aim is to provide fairness and promote team diversity in an existing team, hence we proposed to change the team structure and culture by introducing new experts that help us to achieve that goal. Finally, we see from the literature review on the group formation is that research pays no or little attention to the process of adding a new team member to an existing team formation process and hence new opportunity to explore and investigate this part make such research appealing.

2.2.2 Diverse Group Formation

Creating a group with maximum diversity has been well studied in the literature. It is usually referred to as the “Maximally diverse grouping problem". Several algorithms have been proposed to achieve the maximum diversity among group members. For example, Palubeckis et. al [38] proposed an iterated Tabu search algorithm to get the best solutions among a set of diversified groups. Galego et al. [39] evaluated and proposed a modified version of the same algorithm to solve the problem of maximum diversity in group formation. They indicated the relevancy of such a problem to student teams in business schools where creating diverse student workgroups or training teams is highly required. Lai and Hao [40] studied this from a graph theory perspective and apply a modified iterated search algorithm to get the best candidate solutions to the maximally diverse grouping problem. Dias and Borges [41] studied how a team of students can be created such that it has the maximum diversity among its members while keeping the difference among the minimum. Their algorithm is different from the previous as they considered the variance among the individual profile instead of considering the pairwise distance among the team members. Chen et al. [42] proposed a hybrid genetic algorithm to solve the problem of creating a group of reviewers with maximum diversity. However, their work has not addressed the problem of the utility loss and how maximizing the diversity could not lead to the best result.

Maass et al. [43] proposed a new approach to handle the Maximum Diversity Group Problem by Minimizing Similar Attributes (MSA) within a group. Their approach requires the user to set a desired threshold for each attribute and a penalty score would be assigned due to deviation from the targeted values, and hence maximizing the diversity within a group. MSA approach is different from other approaches that handle the same problem because it requires a target value to be set for each attribute for a certain group. MSA has been evaluated with other baseline algorithm that handles the MDGP to assign University of Michigan Engineering Global Leadership (EGL) Honors Program students to cultural families. The result shows the MSA has created a more diversified family as compared to previous approaches.

2.2.3 Group Fairness

Group fairness as a concept requires protected group members to have the same opportunities and advantages as the proportion of the population as a whole [44]. The protected group can be classified based on gender, race, ethnicity, or any other demographic feature. For example, U.S. federal law protects individuals from discrimination or harassment based on the following nine protected classes: gender, race, age, disability, color, faith or religion, national origin, parental status, or genetic information [45]. Having this importance, several algorithms have been proposed to accomplish such fairness requirements, mostly in the data mining field. Feldman et al. [46] studied the disparate impact, where unintended bias impacts different groups that should be treated similarly according to a fairness requirement. Their approach includes a test to predict the protected groups and rectify discrimination among them.

Group fairness has also been investigated from machine learning algorithms and especially in supervised learning approaches because such approaches are widely used in decision-making systems. Thus, any bias in the training data could lead to discrimination in the outcome of the classifier. For example, Kamishima et.al [47] suggested a regularization model that penalizes the prejudicial outcomes of the classifier which helps maintain an accurate classification. Kamiran et al. [48] argued that there are some explainable discriminations and proposed an algorithm that keeps such discrimination and removes illegal discrimination before getting this data processed by a classifier. Zafar et. al [49] proposed a method to have a fair classifier by having a trade-off between the accuracy of the classifier and the fairness constraints. Dwork et. al studied fairness in the classification with the goal to prevent discrimination against protected groups and maintain the utility. They proposed an approach that similar people should be treated similarly, and they suggest an approach to find similar people and achieve statistical parity. They have also investigated the relation between privacy and achieving fairness [50]. Zehlike et al. proposed [44] a new algorithm for fair ranking where that the proportion of protected members in a group remains statistically above or equal to a predetermined value. Ensuring that a proportion from a protected group is represented in the selected candidates might lead to having unqualified persons selected and leaving qualified persons behind. To overcome this, they proposed two conditions that are: Every selected candidate should be more qualified than every other candidate that is not chosen and for every pair of candidates, the more qualified candidate should be ranked above the less qualified candidates. They validated their solution by considering one protected attribute in different datasets.

The work by Zehlike et al. [44] is the most similar to our research; however, they consider only one aspect in every protected group (e.g., gender only or race only) while in our work we will consider multiple attributes when determining the protected group for each individual by introducing a combined score to achieve a higher degree of fairness among the group. For example, an African American woman will have a higher protected score than an African male American since it has two protected scores in her profile (gender and race). Another difference is that their algorithm focuses on ranking a group of individuals in the same environment while our work addresses adding a member to a group to rectify the level of diversity in that group.

2.2.4 Bias in Academia

The claims of bias in academia have been investigated in the research community. A study by Bormann and Daniel [51] has investigated whether there are some sources of bias behind the decision that is made by a committee to select doctoral and post-doctoral fellowships. They investigated this from the perspectives of gender, nationality, major field of study, and institutional affiliation. The results show that there is evidence of gender and institutional bias in the case of doctoral fellowship selection. Perna et al. [52] investigated the proportion of blacks among faculty and administrators at public higher education institutions in the South of the USA. They used data from the Integrated Postsecondary Education Data System in their study. Their result shows some enhancements to enhance the diversity has been done; however, there is still a gap of fairness for Blacks in the hiring and tenure process that is greater among higher than lower ranking faculty and among tenured than tenure track faculty. McConner [53] examined the impact of gender and race to determine if there is any kind of discrimination and bias for the minority women leaders in the higher education of the USA. The research investigated three leadership positions that are: president, dean, and faculty. The result of this research provides enough evidence to show that gender inequality and race discrimination still exist and needs to be addressed by the higher education system. Also, the study shows that women experience an inequitable work environment where work stress and difficulty in job tenure are expected. The study concludes by giving some recommendations about how to change processes that affect the recruitment, hiring, and retention of minorities in the higher education to provide fairness and equal opportunity in a working environment. Similarly, Gabriel [54] shows that minorities are underrepresented in the process of higher education recruitment in the UK. For example, black professors represent only 0.1% of all professors although they make up 1.45% of the total UK population.

Bias of expertise recommendation in academia has got a fair amount of attention by many researchers where the majority of their research in this regard has focused on the bias in the peer reviewing process. One study published by Nature magazine [55] shows that women are usually underrepresented in the peer review panel for scientific journals. The study provided some evidence by analyzing genders and ages of authors and reviewers from 2012 to 2015 for the journals of the American Geophysical Union (AGU). They found that only 20% of the reviewers are women. They provided two main reasons for that that are: editors nominated fewer women than men and the refusal of some women scientists due to engaging in other workloads. They concluded that the first reason was the major one based on the statistics on the dataset they studied. A similar study by Murray et al. [56] investigated the impact of gender and nationality in the peer review panel for the biosciences journal *eLife* for the period (2012-2017). Their results show that women and authors from emerging countries were underrepresented to be recommended as editors and peer reviewers. Likewise, a study by Holman et al. [57] highlights the issue of gender gap in the scientific publications of different knowledge domains by analyzing the PubMed and arXiv scholarly databases that have more than 36 million authors from over 100 countries in computer science, math, physics, and medical science. Yin et al. [58] showed that ignoring the diversity in reviewers can lead to a source of bias in the reviewing process. They assumed that the greater the social connection among the group the greater is the bias in the reviewing process. Hence, their measurement of diversity was based on the notion of social influence among the reviewers. In their model, they considered three parameters that are: the reputation of the reviewer in the scientific community, the co-authorship network, and the coverage. They compared their model of one of the baseline methods in review panel formation algorithms and their model was better in terms of topics coverage and reviewers' impact in the scientific community.

This problem has received the attention of the NSF since one of its priorities is to promote diversity and inclusion. This was reflected in the project selection funded by this organization. The decisions of proposal selection depended on the feedback from peer reviewers. Hence, forming a diverse group of reviewers was the key parameter to achieve that goal. Accordingly, NSF has developed an automated reviewer selection system as explained by Hettich and Pazzani [8] that called "*Revaide*". This system considered different parameters to select reviewers and some of them are based on demographic data such as gender and EPSCOR (i.e., states that do receive much federal research funding). If the review panel does not contain a female or a reviewer from EPSCORE states, then the system tries to include these groups in the panel. However, the authors do not mention the mechanism to do that and if how they can mitigate the utility loss due to this selection.

The problem of bias in a peer review process was not limited to the gender and race but it can be seen from other angles such as the geolocation of the reviewer. For example, a study in [59] shows that the US dominated the peer review process by 32.9% while their publication represents 25.4% from the whole publications between 2013 and 2017. It also shows that 11 countries reviewed more than two-thirds of the articles published by *Web of Science* for the same period. The study also shows that editors usually selected reviewers from developed regions or from their own region which is one reason that can be concluded from the lower rate of reviewers from the emerging countries. Another study shows by Freeman and Hung [60] investigated the race of authors in more than 2.5 million scientific papers published by US authors for the period of (1985 to 2008). They highlighted that authors from the same race tend to co-author more frequently and publish lower quality work in terms of citation and journal impact. However, a work will produce a higher citation and get published in a high impact journal when researchers collaborate with different researchers who are geographically isolated and from different ethnicity. Their conclusion promotes the idea that fairness and greater diversity lead to increase team productivity and generate a high impact contribution. AlShebli et al. [61] studied the relationship between diversity and research quality. They studied different types of diversity parameters such as ethnicity, gender, academic and affiliation. They analyzed more than 6 million scientists who cover different domains. Their result shows that there is a strong relation between the ethnic diversity of a team and the quality of the produced work. According to their study, a team that is ethnically diverse will produce a higher impact work as compared to a homogeneous team.

3 Research Plan

The overall goal of this research is to design and build expert recommendation algorithms that promote diversity and achieve fairness while minimizing the utility loss, i.e., the drop in the expertise that could occur by prioritizing diversity. Our work focuses on recommending an expert to join a team or building a team as a whole through a sequence of recommendations. This team could be as simple as a small-scale project workgroup or as complex as hiring people to join an organization. In this research, we consider assembling a program committee for a specific conference or adding to an existing program committee as examples of an expert recommendation process.

Although some recent research focuses on assembling a diverse group in different cases, none of them has handled forming a diverse group of researchers or a program committee. Also, previous work focused on promoting diversity based on a specific demographic feature (e.g., age, gender, …etc.). In contrast, we develop algorithms that consider multiple demographic features when recommending an expert to join a group in academic fields. To our knowledge, this is the first research that develops such algorithms that promote diversity and fairness in academia by generating a ranked list of experts considering a variety of demographic features. Our first goal is to construct a profile to represent researchers that includes two main components: The first is an *expertise profile* that quantifies the competence of a researcher in a specific academic field. The second is a *demographic profile* that represents the demographics of a researcher based on certain demographic features.

The second goal of this research is to promote diversity by maximizing the participation of members of underrepresented and/or protected demographic groups while minimizing the expertise drop that we might get by focusing on demographics rather than just expertise when forming a team. We develop and evaluate three novel recommendation algorithms that recommend scholars to a program committee for a specific conference. The first recommends an expert based on expertise only, the second maximizes the diversity by ranking researchers based their overall diversity calculated by combining all their diversity features. The third approach is a hybrid approach that tries to balance between promoting diversity and the need to have experienced individuals using linear optimization. In all of these approaches, we measure the diversity gain and the utility loss. Our research evaluates these algorithms using these two proposed measurements and concludes by determining the best one that can achieve the goal that maximizing the diversity with the lowest expertise utility loss.

The third goal of this research is to enhance diversity by considering another diversity definition proposed in [62], that is, diversity by proportionality. Diversity by proportionality has been referred to as *statistical parity* in Artificial Intelligence, and *demographic parity* in social science. Our goal here is to design algorithms that promote the diversity based on this definition and achieve the best possible balance between diversity and the expertise. We propose two algorithms: 1) our *Voting Diversity* algorithm inspired by [62] and the seat-allocation problem in electoral systems; and 2) our *Hill Climbing Diversity* algorithm that iteratively recommends experts based their ability to close the demographic gap between the proposed recommendation set and the distribution of the population from which we pick our candidates. In both algorithms, we incorporate linear optimization to achieve the best possible balance between diversity and expertise as we did with the *hybrid* algorithm in the previous goal. We evaluate these two algorithms and comparing them to the *hybrid* algorithm from the previous experiment by measuring diversity gain and the expertise savings in the recommendation list generated by each algorithm. The research concludes by finding the best algorithm that achieves the highest possible diversity gain with the best expertise saving, which is the main goal of this research.

3.1 Goal 1: Expert Profiling

Given researcher's name and her or his affiliation, our goal is to build a complete profile for that researcher. The profile consists of two main parts: expertise and demographic information. The expertise profiling contains the expertise score that is quantified by finding the *h*-index for a researcher. Different *h*-index measurements have been proposed; however, we use the *h*-index score as provided by Google Scholar. We build a tool that extracts the *h*-index and current affiliation name from Google Scholar. The second part is extracting the demographic information for that researcher. We consider gender, ethnicity, career stage, institution rank, and geolocation to determine the demographic profile for a researcher. We use some tools to determine the gender and ethnicity from the name and extract other information from the scholar's publication. Finally, we assign a diversity score based on these parameters to determine how diverse is that researcher.

3.1.1 Expertise Profiling

Expertise profiling can be defined as a record that shows the proficiency of specific knowledge domains that an expert possesses [10]. It can be viewed as a vector of scores that shows the competency of each skill for that expert. In academia, there has been considerable interest in developing expert profiles due to the demand of having experts to review papers, participate in conference program committees or grant review panels, or finding talented individuals to join research teams. There have been different attempts to measure the amount of expertise that an expert has in academia as in [6][8][9][15][63][64]. One method that we use in this research is the *h*-index as a metric to assess the scientific performance of a researcher. *h*-index was proposed by Hirsch in 2005 to measure the researcher's quality and productivity [9]. It is a robust single-number metric that uses the number of publications to indicate the quality of the researcher's output and the citation count to represent the quality of the expert's work. *h*-index scores are also employed by funding bodies and employers to determine funding, career decisions, promote and award committees [65][66]. Using a single score number to assess researcher expertise helps to rank those candidates and finally makes these decisions much easier [67]. It has been incorporated in many scholarly databases such as Google Scholar, Web of Science, Scopus, and Publish or Perish. In this research, we use *h*-index as provided by Google Scholar as our metric to represent the expertise score for each researcher, as it tends to offer more excellent coverage and accuracy for computer scientists compared to other bibliometric databases [18]. Accordingly, we build a tool that takes the researcher's full name and the affiliation and returns the researcher's full name, affiliation, *h*-index, and the total citations for the publications of that researcher based on the researcher's Google Scholar profile. We develop the tool to support batch processing where we submit a file that contains a list of names and affiliations of researchers, and the tool returns two files. The first is for every researcher that the tool was able to get their Google Scholar profile, an expertise profile is created and stored in that file. However, in case any researcher had no Google Scholar profile, the tool would exclude this researcher from the dataset and write the name and the affiliation of that researcher to another file to be used later for further verification.

3.1.2 Demographic Profiling

3.1.2.1 Protected Class

A protected class can be defined as that group of individuals that are protected by the law or the authorities against discrimination [45]. Decision making systems have the ethical and legal responsibility to comply with such a requirement by establishing the required policies that restrain any discrimination against these protected classes. In most cases, these classes are demographic features such as gender, race, ethnicity or any other parameters. For example, U.S. federal law protects individuals from discrimination or harassment based on the following nine protected classes: gender, race, age, disability, color, faith or religion, national origin, parental status, or genetic information [45].

3.1.2.2 Protected Classes in Academia

Many academic institutions and scholars realized the significance of including the demographic features in an expert profile for the reason that such data can be used in discrimination countermeasures, achieving fairness goals, complying with state and local regulations with respect to fair and diverse employment opportunities. In this research, we represent the demographic profile using five important features that have been considered major sources of bias in the academic environments that are: gender [51][53], race [52][53], career stage [68], institution geolocation [59], institution ranking [59].

We build the demographic profile for a researcher by incorporating different techniques of feature predictions and web crawling to collect the attributes of the demographic profile as some features are explicitly included in researcher personal home pages while others might not be included due to privacy concerns, and hence prediction tools are employed to predict such a feature. We assign a binary weight for each demographic feature; that is if a feature is part of a protected group then it gets the value of 1; otherwise, the feature weight is set to zero. However, we notice that each protected class can have many segments within; hence, we propose a continuous weight for each class using different mechanisms based on the segments in that demographic feature as shown in Table 3.1. The next section explains these mechanisms in detail.

Table 3.1 – Protected and not protected classes

Demographic Feature	Non-Protected Class	Protected Class
Gender	Male	Female
Ethnicity	White or Asian	Non-White or Non-Asian
Career Stage	Senior Researcher	Student
	Associate Professor	Ph.D.
	Professor	Junior
	Reader	Graduate
	Senior Lecturer	Postdoc
		postdoctoral
		Research fellow
		Researcher
		Adjunct Professor
		Assistant Professor
		Lecturer
Geolocation	Developed Countries	Developing Countries
University Rank	Less than the mean	Equal or greater than the mean

3.1.2.3 Methods

To predict gender and race, we use *NameSor* software described in [28]. The software provides the facility to use the full name (i.e., first and last name) and classify it based on a database of name information of more than 4 billion names [29] with the help of a novel machine-learning algorithm to provide a matching probability for the gender and race. One challenge is identified by [69] with respect to predicting gender with Chinese names. We manually validated any name that had a gender confidence probability of 0.6 or less by checking scholar information on the web. We found that a gender matching accuracy of 80% with respect to Chinese names and 92% percent with respect to others, and we manually rectified any discrepancies. Once scholar gender is predicted, we need to map it from the corresponding binary and continuous values by using the concept of protected parameters. The result of *NamSor* provides two categorical values and hence we define the same weight for binary and continuous demographic profiles where the protected parameter for gender feature is female as presented in Table 3.2 that is assigned a value of 1 and male value of 0.

However, the accuracy was not as high when predicting race, specifically predicting African American as the system provides an accuracy of 15% by labeling many White scholars as African American. Hence, we manually verified every scholar labeled as African American by *NamSor* using the available information on the web (e.g., homepage of a researcher) to correct any classification error. Nevertheless, the software predicts other races with an acceptable accuracy of 75-80%. Moreover, we can see in Table 3.2 that the protected parameter for the feature gender is any race that is not White or Asian. We propose this classification based on our work in [70] and what has been presented in chapter 2. Moreover, the continuous weight for this feature was the complement of the representation of those races in the computer science communities (i.e., everyone who has a B.Sc. in computer science) [71]. For example, the race weight for an African American researcher is 0.944 which is the complement of their distribution in the computer science community that is (5.6%).

Career stage is extracted from the Google Scholar (GS) profile. To determine the categories inside this feature, we used the academic ranks in the US and the UK, which we found that they cover the majority of the academic ranks in the world. The binary weight for this feature has been assigned by having two classes (junior = 1, senior = 0). We define the senior researcher as any researcher who has the academic rank of an associate professor, a senior lecturer, or above. Any academic rank lower than associate professor is considered a junior researcher. However, for the continuous weight, we found that we have five categories insides this feature and hence we define the weight using a step of 0.2, where the lowest rank (e.g., Ph.D. student) has a value of 1 and high rank (e.g., distinguished professor) has a value of 0.

Institution has been collected from the Google Scholar profile of the researcher or home page of that researcher. Then, we use the TIMES computer science university ranking system [72] to determine the rank of a university and mapped it to 0 if the university rank is less than the mean of TIMES computer science university ranking and 1 otherwise. However, the continuous weight has been calculated by dividing this rank by the lowest rank in TIMES computer science university ranking system that is 1260. Hence, the university of rank 1 has the lowest possible value and the one with a rank of 1260 gets the value of 1.

The last attribute is geolocation. We collected this attribute by finding the geolocation of the affiliated Institution of a researcher. We then determine the country of this Institution as provided by the TIMES ranking system. We assign a binary weight for this feature by finding whether the country is a developed country (Binary value: 0) and a developing country (Binary value: 1) as per the last United Nations countries classification [73]. However, the continuous weight has been calculated by finding a country's scientific impact, as measured by *h*-index, in the computer science community as provided in [74]. We divide the *h*-index value by the maximum *h*-index and consider this the continuous weight for the country feature of the demographic profile of the researcher. Table 3.2 provides a summary of the protected features along with their binary and continuous weight.

Table 3.2 – Protected classes within a demographic feature$^{(+\text{ binary},*\text{ continuous})}$

Demographic Feature	Protected Class	Binary Weight	Continuous Weight
Gender	Female	1	1
Ethnicity	Black	1	0.944
	Hispanic	1	0.925
	Others	1	0.995
Career Stage	student	1	1
	Ph.D.	1	1
	Junior	1	1
	Graduate	1	1
	Postdoc	1	0.8
	Postdoctoral	1	0.8
	Research fellow	1	0.8
	Researcher	1	0.8
	Adjunct	1	0.8
	Assistant	1	0.6
	Lecturer	1	0.6
Geolocation	Developing+/Greater or equal to 10th percentile of weight distribution*	1	1-(*h*-index/950)
University Rank	Equal or greater the mean	1	rank/1260

3.2 Goal 2: Diversity Maximizing Recommendation

From goal 1, we were able to get a complete expert profile that reflects his or her expertise and demographic information. As we have seen, the expertise is measured by an accumulative score and the same is applied to the demographic information. Our goal at this stage is to add an expert to a group that has been tasked to achieve a predefined goal. We consider adding a member to a conference program committee (PC) as an example of this class. We can address this as a problem and examine solutions to it from three main perspectives. First, we discuss the problem from an expertise retrieval only and develop an algorithm to achieve this task from that perspective. Next, we look to this problem from computational social science and discuss the concept of protected groups and how this can be applied to achieve fairness when adding a member to a program committee. Finally, we design a hybrid approach that considers these two approaches and some optimization techniques. We aim that such an approach provides a trade-off between the utility loss that results from considering adding a protected group member to a group and the level of diversity that such group has. Optimization techniques are utilized to achieve that goal.

3.2.1 An Expertise Retrieval Approach (EXP)

As discussed in section 3.1, we quantified the skills of scholars using *h*-index, where the higher the score is, the higher the expertise that such scholar has. To add a researcher to a team (e.g., conference program committee), we get the expertise score (i.e., *h*-index in our case) from a conference author's profiles and recommend the scholar that has the highest expertise score. We refer to this approach as **EXP**, and we consider this our baseline algorithm for this goal. One advantage of this approach is that it maximizes the expertise in the process of the recommendation. However, it might lead to a systematic bias by favoring one demographic group over the other by failing to address the issue of demographic diversity. For example, the expertise-only approach does not consider the issue of the gender gap and the race gap. Hence we might end up with a team of the same race or gender. Another concern is by favoring highly-cited researchers; it minimizes the opportunities for junior researchers to attend such research teams, which negatively impacts their chances to advance their careers. Additionally, most highly-cited researchers are employed by the top rank universities, and this approach would less favor those researchers from lower-tier universities.

Algorithm 1 Expertise-Based Recommendation (EXP)

```
Input : E, list of researcher's profiles with $n$ cardinality, each $e$ in E is a triplet of ($e_{name}$, $e_{dem}$, $e_{exp}$) that represents the name, the demographic profile, and expertise score of researcher $e$. N is the required number of candidates.
Output: A list of ranked researchers of size N
```

```
for each $e$ in $E$ do
        $e_{sum\_div}$ ← Sum of all feature scores in $e_{dem}$
end
$E_{EXP}$ = SORT(E, $e_{exp}$ , $e_{sum\_div}$, DESC) //Rank E based on expertise score in a descending order,
    if two or more scholars have the same $e_{exp}$, sort based on the diversity score $e_{sum\_div}$
candidates ← Select N profiles from top $E_{EXP}$
```

3.2.2 A Diversity Approach (DIV)

In this approach, we look at the problem from the lens of social science and the concept of social inequality and bias. For goal 1, we developed a demographic profile besides the expertise profile. In this research, we focus on five demographic parameters that are: gender, ethnicity, geolocation, career stage, and university rank. Also, we defined a protected parameter for each category and as appeared in Table 3.2. To add a person to a specific group using this approach, we calculate the diversity score for a researcher. The diversity score is simply the sum of the weights of demographic features in the scholar demographic profile, and as given by equation 3.1. For example, if the demographic profile for a researcher is (gender = woman, race= African American, career Stage = professor, university rank = 2, country = United States) then the corresponding diversity score with binary profile is (1+1+0+0+0 = 2). We refer to this algorithm as (**DIV**) wherein the scholars are ranked according to their diversity score in descending order. If two or more researchers have the same score, the algorithm picks the one that has the highest expertise score.

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

Algorithm 2 Diversity-Based Recommendation (DIV)

Input : **E**, list of researcher's profiles with *n* cardinality, each *e* in **E** is a triplet of (e_{name}, e_{dem}, e_{exp}) that represents the name, the demographic profile, and expertise score of researcher *e*. N is the required number of candidates.

Output: A list of ranked researchers of size N

```
for each $e$ in $E$ do
        $e_{sum\_div}$ ← Sum of all features in $e_{dem}$
```

```
3. end
4. $E_{DIV}$ = SORT(E, $e_{sum\_div}$ , $e_{exp}$, DESC) //Rank E based on diversity score in a descending order, if two or more scholars have the same $e_{sum\_div}$, sort based on the expertise score $e_{exp}$
5. candidates ← Select N profiles from top $E_{DIV}$
```

3.2.3 A Hybrid Approach (H)

The previous two approaches present some pros and cons. The first approach enhances the expertise of the team but fails to address the problem of forming a diverse team. The diversity approach solves that problem, but again it might cause a drop in the expertise level of a team. Hence, a *hybrid* approach is introduced in this research to solve that. We introduce a tuning parameter (α) to tackle this issue and as presented in equation 3.2:

$$\text{Score}(\mathbf{H}) = \alpha * \text{Score}(\mathbf{DIV}) + (1-\alpha)\ \text{Score}\ (\mathbf{EXP}) \qquad (3.2)$$

Our goal, in this case, is minimizing the utility loss that might result from favoring diversity over expertise during recommendation. In this research, we test different values for α and the result is compared to the previous two approaches to find the effectiveness of this approach. To make both scores comparable, we measure the performance using score scaling such that Score (**EXP**) and Score (**DIV**) are normalized so that both scores get a value between (0 and 1). The following equation is used in our normalization process:

$$\text{Score}_{\text{norm}} = \frac{Score_i - \min\ (\text{Score})}{\max(Score) - \min\ (Score)} \qquad (3.3)$$

This equation maps each score to the interval [0,1] and the minimum score is 0 while the maximum score is 1. We experimentally evaluate the impact of the tuning parameter α by comparing the results to both expertise and diversity approaches. Our goal is to find the best balance between diversity gains and expertise loss.

Algorithm 3 A Hybrid-Based Recommendation (H)

```
Input : E, list of researcher's profiles with $n$ cardinality, each $e$ in E is a triplet of ($e_{name}$, $e_{dem}$, $e_{exp}$) that represents the name, the demographic profile, and expertise score of researcher $e$, N is the required number of candidates, $\alpha$ is a tuning parameter
Output: A list of ranked researchers of size N
for each $e$ in $E$ do
        $e_{sum\_div} \leftarrow$ Sum of all features in $e_{dem}$
        $e_{hybrid} \leftarrow (\alpha * e_{sum\_div}) + ((1-\alpha) * e_{\text{exp}})$
end
$E_{\text{Hybrid}}$ = SORT(E, $e_{hybrid}$, DESC) //Rank E based on $e_{hybrid}$ in a descending order
candidates $\leftarrow$ Select N profiles from top $E_{\text{Hybrid}}$
```

3.3 Goal 3: Demographic Parity Recommendation

In this goal, we present another perspective to the definition of diversity, that is diversity by proportionality, where the outcome of the recommended items at any given rank should approach the distribution of the features that define the population of items. The term has been investigated in the information retrieval community in [62] to diversify search results for a given query in a search engine. It has also been referenced as the statical parity or the demographic parity [50] in the machine learning and artificial intelligence community.

Demographic parity has also been associated with group fairness, where its main goal is to achieve a fair representation of the population as a group. However, it draws some criticism that it fails to address the issue of individual fairness [50]. Previous research on demographic parity has focused on fair recommendations based on a single feature, e.g., age, or race, or gender [44]. We extend this work by developing two new algorithms that aim for demographic parity across multiple features simultaneously.

This work presents two new algorithms that address multifaceted demographic parity. One is inspired by the electoral system and the problem of seat allocation for the competing parties [62]. The other is motivated by the problem of local search in mathematical optimization, wherein we iteratively identify the gap in distribution by feature between the recommended items at a given rank and the proportion of those features in the item pool. In Chapter 4, we evaluate these two algorithms and compare them to the hybrid algorithm from the previous goal to determine which algorithm(s) are best at achieving group and individual fairness while minimizing the loss of expertise.

3.3.1 A Hill Climbing Diversity Approach

In this approach, we develop an expert recommendation algorithm that considers the demographic gap with respect to the demographic parity between a group and the population of experts (i.e., researchers' pool). At each step, the algorithm tries to build a solution towards its ultimate goal. i.e., demographic parity with the pool of available researchers. This approach is inspired by the mathematical optimization technique of Hill Climbing [75], wherein an incremental change is applied to a solution as it iteratively improves an objective function.

The algorithm starts with defining a demographic profile for the researchers' pool that has the same number of demographic features that define the researcher's demographic profile. Hence, we have $\overrightarrow{P_{parity}}$= <$w_1$, w_2 , …, w_n>, where *n* is the number of demographic features that define the researcher's demographic profile. We have also $\vec{D}$ = <d_1, d_2, d_3, …, d_n> that defines the demographic profile of a researcher. To calculate the weight *w* for each protected feature in, we consider the proportion of the protected class as defined in Table 3.1 in the researcher's pool.

Now, we have $\overrightarrow{P_{Group}}$ = <x_1, x_2, x_3, …, x_n> that represents the demographic parity of the newly formed group after adding the recommended expert. Initially is a vector of zeros. We also have $\overrightarrow{P_{Gap}}$ that represents the difference between $\overrightarrow{P_{parity}}$ and $\overrightarrow{P_{Group}}$. We iteratively calculate $\overrightarrow{P_{Gap}}$ as in equation 3.4, where *1 < i < k*, and *k* is the required size of recommendation.

$$\overrightarrow{P_{Gap(i)}} = \overrightarrow{P_{parity}} - \overrightarrow{P_{Group(i-1)}} \qquad (3.4)$$

Iteratively, we calculate two scores for each researcher. The first one is the diversity that is given by equation 3.5 and it represents the similarity between the gap profile and a researcher as given by the cosine similarity and the second is the expertise score as given by the h-index. The score for both diversity and expertise be normalized as in equation 3.3 to have a value between 0 and 1. Then we use the linear optimization with the tuning parameter α as in equation 3.2 to combine the score from the diversity and the expertise. Finally, the expert that has the highest combined score is recommended and removed from researchers' pool. As soon as the recommendation is done, we calculate the new $\overrightarrow{P_{Gap}}$ and the $\overrightarrow{P_{Group}}$ based on this recommendation. Iteratively, the process continues until we recommended the required number of expert *K*. We may notice that it is possible that we might reach our goal that is $\overrightarrow{P_{Group}} = \overrightarrow{P_{par\imath ty}}$ at any step of *K*, hence $\overrightarrow{P_{Gap}} = 0$, which causes the diversity score = 0. However, as soon as this case occurs, we restart the recommendation process from the beginning by having $\overrightarrow{P_{Group(\imath-1)}} = 0$ and $\overrightarrow{P_{Gap}} = \overrightarrow{P_{par\imath ty}}$ and continue with the recommendation. The pseudo-code for this algorithm is given in Algorithm 4.

$$\text{Score(Diversity)} = \cos(\overrightarrow{P_{Gap(\imath)}} , \vec{D}) = \frac{\sum_{j=1}^{m} p_j \, d_j}{\sqrt{\sum_{j=1}^{m} p_j^2} * \sqrt{\sum_{j=1}^{m} d_j^2}} \qquad (3.5)$$

Algorithm 4 Hill Climbing Diversity Recommendation

Input : E, list of researcher's profiles with *N* cardinality, each *e* in **E** is a triplet of (e_{name}, e_{dem}, e_{exp}) that represents the name, the demographic profile, and expertise score of researcher *e*, K required number of candidates, α is a tuning parameter

Output: candidates, a list of ranked candidates of size K

```
Calculate $\overrightarrow{P_{par\imath ty}}$ Based on $e_{dem}$ in R // calculate the demographic parity for the researchers' pool
$\overrightarrow{P_{Group(0)}}$ = vector (n,0) // Initialize demographic parity vector for the new group to all zeros
    with length of n demographic features
for i = 1 to K do
    $\overrightarrow{P_{Gap(\imath)}} = \overrightarrow{P_{par\imath ty}} - \overrightarrow{P_{Group(\imath-1)}}$
    If $\overrightarrow{P_{Gap(i)}}$ is a vector of zeros do: // check if we reach the parity before having the required size of recommendation
        $\overrightarrow{P_{Group(i-1)}}$ = vector (n,0)
        $\overrightarrow{P_{Gap(i)}}$ = $\overrightarrow{P_{parity}}$
    end
    for j = 1 to N do
        Diversity_score[j] ← COSSIM($P_{gap(i)}$, $e_{(j)dem}$)
        Expertise_score[j] ← $e_{(j)exp}$
        DIV_NORM[j]= NORM(Diversity_score)
        EXP_NORM[j] = NORM(Expertise_score)
    end
    for j = 1 to N do
        Gap_score[i] ← (α * Diversity_score[j]) + ((1-α) * EXP_NORM[j])
    end
    Gap_score_ordered = SORT(Gap_score, DESC)
    New_member = Gap_score_ordered.pop() // get the highest score candidate
    E.remove(new_member) // remove the candidate from researchers' pool
    Candidates[i] = new_member    // add the candidate to the final recommendation list
    $\overrightarrow{P_{Group(i)}}$ = DIST(Candidates) // update the new group vector based on the distribution after adding a new candidate
end
Output Candidates
```

3.3.2 Voting Diversity Approach

The problem of finding a representative group that reflects the multiple features that define a subgroup of a population has been studied in different venues. In political science, for example, the problem of seat allocation for the competing parties in an election process can be seen as a good example of this category. We see that the problem of forming a diverse group of experts that reflect the demographics of the population of experts is similar to this problem based on the following analogy: The rank of each expert in the proposed recommendation list can be seen as the seat number that is allocated to a competing party (i.e., demographic feature in our case) based on the number of votes it gets (the percentage of this feature in the researcher's pool).

Based on the above analogy, we present our *Voting Diversity* expert recommendation algorithm that is an adaption of D'Hondt method, named after Belgian mathematician Victor D'Hondt who described this technique in 1878 [76] . In the United States it is named, Jefferson Method, after Thomas Jefferson, who introduced the method for proportional allocation of seats in the United States House of Representatives in 1792 [77]. Both methods give the same outcome; however, the methods of presenting the calculation are different. This method is the standard technique that is used to allocate seats in the election of 50 countries in different parts of the world.

Our recommendation algorithm starts by assigning seats based on the winning demographic features using D'Hondt method, a sub-group for every member of this demographic feature (aka. party) is created. Then we use linear optimization with the tuning parameter (α) considering their diversity score and expertise score to get the final score for every member in this subgroup. Finally, the one with the highest score is recommended. In the coming sections, we describe D'Hondt method, and we follow by our *Voting Diversity* expert recommendation algorithm.

3.3.2.1 D'Hondt method

This method assigns the available seats in sequential order for the competing parties. For each seat, it computes a quotient for all of the parties based on the votes they receive and the number of seats they have been assigned so far as in equation 3.6. This seat is then assigned to the party with the largest quotient, which helps maintain the overall proportionality. Finally, it increases the number of seats assigned to the chosen party by one. The process repeats until all seats are assigned. Pseudo-code for this procedure is provided as Algorithm 5. In this procedure, D = {D_1, D_2, ..., D_n} is the set of parties and M_i = {m_{i1} , m_{i2} , ..., m_{in}} is the set of members in party M. v_i indicates the percentage of votes D_i possesses in the pool and s_i the number of seats that have been assigned to D_i so far.

$$\text{Quotient} = \frac{v_i}{1+ s_i} \qquad (3.6)$$

Algorithm 5 D'Hondt method

Input : D, a set if of parties of cardinality m , V is a vector of votes each party in D gets, N number of seats, M_i is the set of members for a party D_i

Output: candidates, a list of ranked candidates

```
$\vec{S}$ = vector (0, $m$) # vector to track the number of seats each party gets
quotient = vector (0, $m$) # vector to store D'Hondt quotation value for each party
for s = 1 to N do
    for i = 1 to m do
        quotient[i] = $\frac{v_i}{1+ s_i}$        # calculate D'Hondt quotation for each party
    end
    $i^*$ ← $\arg_{\max}$ (quotient)  # $i^*$ is the index of the party in the set D that wins the seat
    $d^*$ ← pop $M_i^*$ # get the member from the winning party $D_i$ member list $M_i$
    $s_i^*$ ← $s_i^*$ + 1  # update the number of seats in $\vec{S}$ for the winning party $D_i$
    candidates[s] ← $d^*$ # add the member to the candidates list
end
output candidates
```

3.3.2.2 Method

In this section, we present our adaption of the seat allocation problem. D'Hondt method assigns every seat to only one competing party based on the value of the quotation as given in Algorithm 5 above. However, in our case, we assume that the seat can be assigned to a certain demographic feature and at the same time can be shared with multiple demographic features. The other adaption that we use here is that our goal is not only to achieve group fairness, but we want to maintain individual fairness within the same demographic segment. Hence, we consider the linear optimization with α to combine the expertise and diversity scores for the same demographic segment. To put this into perspective, we illustrate the step for our *Voting Diversity* approach in

Algorithm 6. The algorithm first starts with determining the votes (i.e., proportion) of the protected and the non-protected groups in the population (i.e., the researchers' pool). Then it begins assigning seats based on the group that has the highest quotient considering we have only two parties (i.e., protected and non-protected group). The process continues iteratively until the assignment of all available seats. The assignment with a non-protected group is based on the expertise score. However, if the seat had to be assigned for the protected group, then we would consider having five parties (i.e., demographic features) competing for that seat. As with the protected and non-protected group quotation calculation, we have voting vector *V* that stores the proportion of each feature in the pool and *S* vector that keeps track of the seats allocated so far for each feature, where *s* is initially a vector of zeros. *V* and *S* are used to calculate the quotation for each feature and the one that has the highest quotation is chosen to recommend a member based on.

The final stage is to assign that seat to a member of that group. This is done in three steps. First, we create a sub-group for every one that has this winning feature in his profile (i.e., non-zero feature score). Second, we calculate the diversity score by summing the score of demographic features in scholar's demographic profile. The last step is to apply our linear optimization algorithm on this group with the tuning parameter (α) to get the final score for everyone in this group. The one that has the highest score is recommended and removed from researchers' pool. The seating vector is updated according to the feature scores in the recommended member. For example, if the recommended researcher based on a winning feature that is gender has the following demographic profile that is (gender = woman, race= African American, career Stage = assistant professor, university rank = 22, country = Singapore), then this means that his demographic profile is <1,1,1,0,1> with a diversity score of 1+1+1+0+1= 4, the seat is divided among every protected feature that is non-zero, hence and assuming that this the first seat to be assigned, the *s* vector will be <0.25,0.25,0.25,0,0.25>. The process is repeated whenever a seat of the protected parameter has to be assigned. The pseudo for this algorithm can be seen as in Algorithm 6.

Algorithm 6 Voting Diversity Recommendation

Input: E, list of researchers, each *e* in **E** is triplet of (e_{name}, e_{dem}, e_{exp}) that represents the name, the demographic profile, and expertise score of researcher *e*. Those researchers have at least one protected parameter in their e_{dem}.
N, list of researchers with no protected parameter in their demographic profile, each *e* in **N** is triplet of (e_{name}, e_{dem}, e_{exp}) and they are sorted in a descending order according to their e_{exp}
α, a tuning parameter that has a value ∈ [0, 1]
Output: candidates, a list of ranked candidates

```
$\vec{S}$ = vector (0, 2) vector to track the number of seats for protected and non protected
V = <$v_p$ , $v_n$ > // calculate proportion for protected and non-protected features in the population
Q = vector (0, 2) // vector to store D'Hondt quotation of protected ($q_p$) and non-protected features ($q_n$)
VP = <$v_1$ , $v_{2, \ldots,}$ $v_n$ > // calculate proportion for each protected features
QP = vector (0, n) > // vector to store D'Hondt quotation for n protected features
$\overrightarrow{SP}$ = vector (0, $n$) # vector to track the number of seats for each protected feature
for all seats in the ranked list S do
    for $v_i$ in V do
        Q[i] = = $\frac{v_i}{1+ s_i}$ # calculate D'Hondt quotation for protected and non protected
    end
    If $q_{p >=} q_n$
        for all features $D_j$ ∈ D do // D is the set of demographic features in $e_{dem}$
            QP[j] = = $\frac{vp_i}{1+ sp_i}$ //calculate D'Hondt quotation for every protected feature
        end
        winning_feature_index = index(argmax(QP))
        feature_members = subgroup(E, $e_{dem}$(winning_feature_index)!= 0)
        For each $e$ in feature_members do
            $e_{sum\_div}$ ← Sum of all features in $e_{dem}$
            $e_{hybrid}$ ← (α * $e_{sum\_div}$) + ((1-α) * $e_{exp}$)
        end
        $E_{candid}$ = SORT(feature_members, $e_{hybrid}$, DESC) //Rank subgroup based on $e_{hybrid}$ in a descending order
        new_member = ( $E_{candid}$.pop(0)) # select the first candidate
        E.pop(candidate) # remove the recommended candidate from the recommendation pool
        For i = 1 to n do
```

```
            $SP[i] = SP[i] + (candidate_{dem}[i]/ candidate_{sum\_div})$   //update the seats for each feature, since one or more features can be in researcher's demographic profile, we divide by the sum of feature scores so that each feature gets its share in the allocated seat
        end
      $S_p = S_p + 1$ // keep tracking of the seats assigned to a protected group (i.e. parent seats)
    else:
      new_member = (N.pop(0))  //remove the selected scholar from the non-protected
      $S_n = S_n + 1$ // keep tracking of the seats assigned to non-protected group
    Candidates[i] = new_member // add the new candidate to the recommended candidates
end
output candidates
```

In this chapter, we have discussed our three goals which answer the questions of (1) how to model a researcher in terms of expertise and demographics, (2) how to address the diversity in expertise recommendation (3) how to maintain the demographic parity as a diversity measure in the recommendation process. We have also provided a description of the algorithms that we developed to address each of the aforementioned goals. In the next chapter, we will describe the datasets being used as well as the methods employed to evaluate each of our goals.

4 Evaluation

4.1 Dataset:

We will test our recommendation algorithms by recommending scholars to join an existing conference program committee (CPC). Hence, we created a dataset that includes the CPC and authors of the top three Association for Computing Machinery(ACM) conferences that have high impact factors [78]. The Association for Computing Machinery (ACM) is an international non-profit scientific association with more than 100,000 members as of 2019. It is considered the oldest and largest computing society in the world, and it has 37 Special Interest Groups (SIGs) that reflect different computing disciplines [79]. We collected information about all the authors and PC members for SIG-CHI (The ACM Conference on Human Factors in Computing Systems), SIG-COMM (The ACM Conference on Data Communication), and SIG-MOD (ACM Conference on Management of Data) for the year of 2017. As previously noted, these Conferences' Program Committees (CPCs) were found to be less diverse than the set of authors whose papers were accepted to present at the conference [70]. Using information available from their Google Scholar page and home page, we collected five demographic features that are: gender, race, career stage, geolocation, and university rank using our demographic profile module that we discussed in Chapter 3, discarding researchers in industry and that missing demographic information. The total profiles in our dataset are 1,217 distributed as shown in Table 4.1.

Table 4.1 Evaluation data set of academic authors with demographic profiles

Conference	PC members	Authors
SIGCHI17	213	436
SIGMOD17	130	290
SIGCOMM17	23	125

4.2 Metrics

4.2.1 Expertise Evaluation

We evaluate the quality of the expertise in the ranked list generated by our algorithms using non-Discounted Cumulative Gain (nDCG). The nDCG is a standard information retrieval metric proposed by Järvelin and kekäläinen in [80] to measure the quality of documents retrieved in a ranked list. Since its introduction in 2002, it has been extensively used to determine the quality of ranking provided by search engine algorithms and recommendation algorithms [81].

To determine the quality of an item, the nDCG requires a relevancy score for that item that determines how a document is relevant to a submitted query in a search engine, or how a recommended item matches the preference of a user in a recommendation system. The core concept of nDCG is a position of a document in a ranked list determines the gain that this document contributes to the quality of ranking and that an item with high relevancy should be ranked higher than a document with a marginally relevant one [80].

To put this in perspective, let's assume that we have a set of non-binary relevance judgments for *n* items. Then, the Cumulative Gain (CG) is given by:

$$CG = \sum_{i=1}^{n} score\ (i) \qquad (4.1)$$

The cumulative gain presented in the previous equation does not consider the ranking of an item, and hence in such a scheme of ranking whether less relevant or high relevant items are ranked higher makes no difference. For example, a search that retrieves three items with a relevancy score of 1, 2, 1 would have the same CG with a search that has this relevancy order 1,1,2. This is problematic, and hence the need to consider the item rank in such metric becomes a necessity. One solution proposed in [80] is that the relevancy of an item is discounted by the log of that rank. The logarithmic function would be the best fit in this setting as it gradually, but not abruptly, reduces the item score as its rank gets higher. For example, log 2 = 1 and log 1024 = 10 with base 2, thus a document at the position 1024 would still get one-tenth of its relevancy score. Since the log of rank 1 is zero, some literature suggested the score of the item at ranked 1 should not get discounted [80]. However, others suggested that we increment the rank by 1 [82].

$$DCG = (\sum_{i=1}^{n} \frac{score\ (i)}{\lg(1+i)}) \qquad (4.2)$$

To put a stronger emphasis on the relevancy of an item, a modified Discounted Cumulative Gain (DCG) is proposed in [82]. We adopt the latter approach in our research. Accordingly, the Discounted Cumulative Gain (DCG) is given by:

$$DCG = (\sum_{i=1}^{n} \frac{2^{score(i)}-1}{\lg(1+i)}) \qquad (4.3)$$

Different ranking can generate different DCGs based on a set of items. However, to make the DCG comparable among different ranks, the DCG is normalized by the Ideal DCGs (IDCG) that results from sorting all of these items in terms of their relevancy score in a descending order and as given by equation 4.4:

$$nDCG = \frac{DCG}{IDCG} \qquad (4.4)$$

In our evaluation, we consider the expertise score, as provided by the *h*-index, the relevancy score and we evaluate the quality of the expertise in the ranked list of researchers provided by our algorithms using the nDCG, in a similar way to the evaluation of a ranked list of items in traditional recommendation algorithms. The DCG is calculated using equation 4.3 for each ranked list and the IDCG is the ranking that is produced by the expertise recommendation approach.

4.2.2 Diversity Evaluation:

In this section, we presented two metrics to evaluate the diversity, the *multi-feature normalized Discounted Cumulative Gain* (**mnDCG**) mnDCG and the *Cumulative Proportionality for Ranking* **(CPR**). The use of each metric depends on our definition of diversity. In this research, we explored two definitions, the first one is to maximize the representation of the minorities, where mnDCG is used, and the corresponding evaluation can be seen in experiment one. However, the second defines the diversity is that the representation of a certain demographic segment in a group should match the distribution of that segment in the population. To evaluate that, we use the CPR metric as defined in section 4.2.3 to measure the diversity gain.

4.2.2.1 Multi-Feature Normalized Discounted Cumulative Gain (mnDCG)

In this research, we have considered many demographic features in the demographic profile of a researcher. In the literature that we explored, the metrics to evaluate the diversity in a ranked list consider only one attribute. Hence, we have to come up with a new evaluation metric that supports evaluating ranking based on different features.

In the previous section, we discussed the nDCG and its popularity as a metric to evaluate ranking. However, nDCG supports only one feature in calculating the gain at each rank. Nevertheless, we argue that a modified nDCG version can help provide a promising metric for evaluating ranking based on multi features.

In this section, we provide our proposed metric that is the *multi-feature normalized Discounted Cumulative Gain* (**mnDCG**) to evaluate the ranking based on multi-features. We will explain the modification that we did along with examples.

First, let's assume that we have *m* features, we calculate the DCG per feature as in equation 4.3. Once DCG is calculated, then Ideal Discounted Cumulative Gain (IDCG), is calculated *for each feature* by ranking candidates in a descending order based on that feature. Now, nDCG for that feature can be calculated using equation 4.4. The process repeats itself for all features and the mnDCG is the average nDCG gain over all features as shown in 4.5.

$$mnDCG = \frac{1}{k}\sum_{j=1}^{k} nDCG_f \qquad (4.5)$$

Metric 1 Multi-feature normalized Discounted Cumulative Gain (mnDCG)

```
Input: R, a ranked list of a candidate with of n cardinality
D, a set of candidate demographic profiles of n cardinality
d ∈ D, where d is a vector of length m  // This represents the demographic profile
f ∈ F that has a cardinality of m        // Demographic feature set
Output: multi-feature normalized Discounted Cumulative Gain (mnDCG)
Initialize nDCG, IDCG, nDCG, mnDCG
// IDCG Calculation
For i ← 1 to m // loop for all features
    feature_all_score = subset(R,d[i])      // get all feature scores from R to a subset
    feature_all_score_ranked = ORDER(feature_all_score)
    For j ← 1 to n
        $IDCG[i] = IDCG[i] +(( 2^{feature\_all\_score\_ranked[j]} -1) / (lg(j+1))$
    End
End
// DCG Calculation
For i ← 1 to m
    feature_all_score = subset(R,d[i])
    For j ← 1 to n
        $DCG[i] = DCG[i] +(( 2^{feature\_all\_score[j]} -1) / (lg(j+1))$
    End
End
// nDCG Calculation
For i ← 1 to m
    nDCG[i]= DCG[i]/IDCG[i]
    mnDCG= mnDCG + nDCG[i]
End
mnDCG = mnDCG / m
Output mnDCG
```

For example, let assume that we have the following demographic profiles in a dataset of three candidates as shown in Table 4.2. Let's also assume that we have two algorithms that generate the following ranks R1= <C1, C2, C3> and R2 = <C1, C3, C2>. Our hypothesis is that a metric should favor R2 instead of R1 since R2 brings more diversity than R1. Now, we apply our proposed metric (mnDCG) to verify if it satisfies this condition.

Table 4.2 Example of demographic profiles

Candidate	Gender	Race	Career Stage	Geolocation	University Rank
C1	1	1	1	0	0
C2	1	1	0	0	0
C3	1	0	0	1	0

First, we calculate the DCG per feature as in equation 4.3 for each algorithm. We would end up having the following:

DCG_{R1} = <2.13, 1.63, 1, 0.5, 0 >

and DCG_{R2} = <2.13, 1.5, 1, 0.631, 0 >.

Then, using equation 4.4, we calculate the IDCG per each feature and as follows:

IDCG = <2.13, 1.63, 1, 1, 0 >

Using equation 5, the final mnDCG for R1 and R2 is:

$mnDCG_{R1}$ = 0.7, $mnDCG_{R2}$ = **0.71013**

As we can see, R2 has a better gain relative to R1 and this confirms our hypothesis. Later, in the result section, we present our evaluation for the proposed algorithms in goal 1 using nDCG, mnDCG, and the F-measure.

4.2.2.2 Cumulative Proportionality for Ranking (CPR):

CPR is a metric first proposed by Dang and Croft in [62] as a diversity by proportionality measure to quantify the proportion of query topics in a set of ranked documents retrieved at rank *K*. It has been inspired by Gallagher index [76] which is a statistical analysis methodology used in electoral systems to measure the difference between the percentage of votes each party receives and the percentage of seats it gets.

In this research, we utilize this metric to measure the diversity as defined in goal 2, where the goal of the diversity is to achieve the demographic parity that is the demographics in the generated ranking need to be the same as the demographics of the group. The metric starts with calculating the disproportionality at rank *K* and as in equation 4.7:

$$\text{DP@K} = \sum_{feature} c_i \, (v_i - s_i)^2 \quad (4.7)$$

where v_i is the percentage of this demographic feature in the pool and s_i is the percentage of seats that have been assigned by an algorithm to that feature at the rank *K*. We argue that in the resulted expert list, having a demographic feature that already has enough experts in the ranked list is not as bad as a demographic feature that does not have enough experts; hence we have c_i such that:

$$c_i = \begin{Bmatrix} 1 & v_i \geq s_i \\ 0 & otherwise \end{Bmatrix} \quad (4.8)$$

In equation 4.8, c_i acts as a penalizing parameter for underrepresenting any demographic feature (i.e., $s_i < v_i$) in the expert list generated by our proposed algorithms, but not for overrepresenting them ($s_i > v_i$). However, it is important to note that we are measuring the *disproportionality* at this stage, hence if the percentage of the feature in the resulted list is greater than the parity (over-represented), this means that the list is not disproportional (i.e., it is proportional) and we penalized the result set by having $c_i = 0$. This means that we consider the list is proportional even if a feature is overrepresented but not underrepresented.

Now, a perfect disproportional expert list for a feature *i* is a list that does not have this feature in the resulted expert list, hence Ideal DP@K is given by equation 4.9:

$$\text{Ideal DP@K} = \sum_{feature} (v_i)^2 \quad (4.9)$$

Next, we calculate the proportionality measure in a ranked list by normalizing the DP score with Ideal-DP so that the results are comparable among all expert ranked lists:

$$\text{PR@K}= 1- \frac{DP@K}{Ideal\ DP@K} \quad (4.10)$$

Finally, we calculate the Cumulative Proportionality (CPR) measure for a ranking as in equation 4.11:

$$\text{CPR@K} = \frac{1}{K}\ \sum_{i=1}^{k} PR@i \quad (4.11)$$

4.2.3 F-Measure:

The F-measure is a popular information retrieval metric that has been used to evaluate the performance of search engine results. It represents the harmonic mean between the precision and the recall of a retrieved list of documents [81][83]. In our evaluation, the F-measure is used to calculate a harmonic mean between expertise and diversity. We have the expertise gain measured by the nDCG and the diversity gain measured by mnDCG, or in one case the CPR. The F-measure is given by equation 4.12:

$$\text{F-measure} = \frac{2*Expertise\ gain*Diversity\ gain}{Expertise\ gain+Diversity\ gain} \quad (4.12)$$

The highest possible value of an F-measure is 1, indicating perfect expertise and diversity gains, and the lowest possible value is 0, if either the expertise or diversity gain is zero. In this research, we seek to maximize the F-measure, where the best algorithm is the one that attains the highest F-measure.

4.2.4 Expertise Savings

It refers to how much of the expertise is kept after incorporating the diversity. To measure this, we use the nDCG as described in section 4.2.1. The nDCG requires two inputs: ranking provided by an algorithm, and an ideal ranking. Since we evaluate the expertise in this metric, then the ideal ranking is the ranking produced by the expertise recommendation approach) as this approach ranks the candidates according to their qualifications (i.e., *h*-index in our case).

4.3 Experiment 1: Diversity Maximizing Evaluation

In this experiment, we evaluate the algorithms that we presented in goal two, i.e., the diversity and expertise-based recommendation algorithms. We use the RAND recommendation algorithm as our baseline in this experiment. The RAND algorithm randomly picks candidates from a conferences' authors pool to be selected to a PC. Our evaluation is based on measuring the demographic diversity gain in each algorithm; hence we tested our proposed algorithms on each conference separately and measure the diversity gain by recommending candidates in the top K ranking produced by these algorithms using different values of K. We tested different values for K to evaluate the impact of K on the diversity of the recommendation generated by every algorithm. We tested three different values for K that are: 50, 100, and K = PC size. We reported the diversity gain using mnDCG since our goal in this experiment is to maximize the representation of the protected groups in a ranked set of candidates produced by our ranking algorithms and as in Table 4.3.

Table 4.3 mnDCG Diversity gain rankings produced by RAND, expertise, and diversity algorithms with binary demographic profile

Conference	**Rank@K**	**RAND**	**DIV**	**EXP**
SIGCHI17	50	0.179	0.632	0.103
	100	0.253	0.633	0.112
	PC Size (213)	0.280	0.700	0.164
SIGCOMM17	50	0.374	0.682	0.275
	100	0.487	0.804	0.503
	PC Size (23)	0.251	0.583	0.179
SIGMOD17	50	0.207	0.573	0.136
	100	0.245	0.630	0.183
	PC Size (130)	0.298	0.651	0.220

As Table 4.3 presents, the DIV algorithm always outperforms the other algorithms with respect to diversity gain. We also notice that the expertise algorithm produces the poorest diversity performance as compared to other algorithms, including random, indicating that it generates program committees that do not reflect the diversity of the community. With respect to the recommendation size, we can notice that with K=50, we can achieve approximately 0.6 of diversity gain with the diversity-based approach; however, the gain slightly changes as K gets higher. Nevertheless, the diversity gain in the expertise recommendation approach steadily increases with the value of K indicating a high correlation between the diversity gain and the size of the recommendation. This shows that the size of the recommendation does not have the same impact on diversity as with the expertise-based approach.

Table 4.4 mnDCG Diversity gain ranking produced by RAND, expertise, and diversity algorithms with continuous demographic profile (* statistically significant p < 0.05)

Conference	Rank@K	RAND	DIV	EXP*
SIGCHI17	50	0.281	0.655	0.142
	100	0.282	0.656	0.153
	PC Size (213)	0.329	0.698	0.212
SIGCOMM17	50	0.340	0.686	0.266
	100	0.458	0.788	0.498
	PC Size (23)	0.195	0.601	0.183
SIGMOD17	50	0.229	0.549	0.151
	100	0.289	0.612	0.194
	PC Size (130)	0.305	0.656	0.225

We repeat the same experiment having continuous wights to represent a researcher's demographic profile. Table 4.4 depicts the result of this experiment where the diversity gain, as measured by mnDCG, has been calculated for each algorithm. We compare this result to the ones provided by the binary profile in Table 4.3. The results show similar trends for the DIV algorithm with a slight advantage for the continuous profile with no statistically significant difference.

Moreover, we can argue that the list of candidates produced by the DIV algorithm for the binary and the continuous might not be the same due to having different demographic weight representations. For example, a researcher of a black race would be more favorable to be selected than a one of a Hispanic race with continuous weight representation assuming that they have the same other demographic features, while this is not the case with the binary as they both have the same race feature weight. The conclusion that can be drawn here is that DIV algorithm with continuous weights can provide a slightly better diversity gain during the selection process. However, the impact is not clear on the individual level because of different sets of candidates produced when using binary and continuous profiles.

Unlike DIV, the EXP algorithm provides the lowest diversity gain among the three algorithms. Interestingly, EXP results show that the continuous-weight demographic profile provides a considerable improvement in the diversity gain versus the binary profile. This result is statistically significant.

Finally, it can be argued that the previous experiment has presented some bias in the evaluation as measuring the diversity in an algorithm that promotes the demographic diversity would always favor such an algorithm. However, the goal of this experiment was not to show which algorithm performs better in terms of diversity, but to quantify the amount of diversity each algorithm produces, the impact of the size of the recommendation on each algorithm and evaluate two demographic feature representation models. This experiment has provided some evidence that we need to diversify the scientific community as we get the least diversity gain with the expertise recommendation approach where the dominant demographics have a higher opportunity of exposure and hence end up having higher publications and as a result higher expertise score. However, to get this rectified, and to promote demographic diversity in the scientific community, we proposed the *hybrid* approach in section 3.2 that balances the demographic diversity with expertise. We hypothesize that this approach promotes the underrepresented groups by selecting those candidates who are demographically diverse and have the best skill score among their peers in these groups and bounding the loss of expertise that comes at the cost of promoting diversity.

4.4 Experiment 2: Hybrid Algorithm Evaluation

The goal of this experiment is to evaluate the performance of the *hybrid* algorithm, presented in section 3.2 that incorporates the results from the diversity algorithm (DIV) and the expertise algorithm using a linear tuning parameter (α). Our experiment consists of two stages: first, we find the best value setting for α where we can find the best balance between diversity and expertise (i.e., the best F-measure). Second, we evaluate the demographic composition provided by this algorithm by trying to form a PC for each conference and study whether the binary or continuous profile can have a major impact on the demographics of the formed group. In both experiments, we consider the expertise recommendation approach as our baseline algorithm.

4.4.1 Hybrid Tuning

In this experiment, we empirically determine the value of α that maximizes the diversity gain and minimizes the expertise loss that might result from incorporating the diversity achieving the best possible balance. We report the expertise savings to represent the amount of expertise retained after incorporating diversity, and the diversity gain relative to the baseline expertise algorithm. We use F-measure to combine the two diversity and expertise gains into a single metric. We report the result using α steps of 0.1 to find the best value of α where $\alpha = 0$ indicates the expertise only algorithm and α 1.0 indicates the diversity only algorithm. We tested our *hybrid* algorithm on the three conferences in our dataset by recommending authors to join a PC from the Author's pool for each conference through the process of ranking authors by our *hybrid* algorithm. We report the average result for each metric as it is presented in Table 4.5 and Table 4.6. We started by having the binary demographic profile as our input. We observe that the highest F-measure is achieved when α is 0.4 indicating a 60% contribution from the expertise ranking and 40% from the diversity algorithm. The lowest F-measure is achieved when α = 1 (i.e., diversity only approach) with an average F-measure value of 0.662 indicating that selection based on expertise only outperforms selection based on diversity only.

Table 4.5 Hybrid algorithm evaluation with binary demographic profile

α	Diversity Gain	Expertise Gain	F-Measure	Diversity Gain	Expertise savings
0	0.597	1.000	0.748	0.00%	100.00%
0.1	0.612	0.999	0.759	2.53%	99.89%
0.2	0.637	0.993	0.776	6.60%	99.26%
0.3	0.677	0.971	0.798	13.44%	97.06%
0.4	0.716	0.935	**0.811**	**19.96%**	**93.48%**
0.5	0.773	0.854	0.810	29.37%	85.37%
0.6	0.807	0.788	0.797	35.07%	78.78%
0.7	0.826	0.731	0.776	38.38%	73.11%
0.8	0.828	0.671	0.741	38.61%	67.06%
0.9	0.833	0.691	0.755	39.44%	69.10%
1	0.824	0.646	0.724	37.94%	64.63%

We repeat the same experiment, but with author's continuous demographic profile this time and we reported the same metrics that we had when testing our *hybrid* algorithm with the binary. As indicated in Table 4.6, we can notice the same trends that we got when using binary profile where the F-measure gradually increases reaching the best possible score when α = 0.4 and slowly decreasing to have its minimum at α = 1. Moreover, we can see that using continuous profile has resulted in a slightly better F-measure as compared to using binary profile for this dataset. However, this cannot be generalized as the result is not statistically significant (p = 0.25).

Table 4.6 Hybrid algorithm evaluation with continuous demographic profile

α	Diversity Gain	Expertise Gain	F-Measure	Diversity Gain%	Expertise Savings %
0	0.614	1.000	0.761	0.00	100.00
0.1	0.629	0.999	0.772	2.40	99.86
0.2	0.656	0.990	0.789	6.90	98.99
0.3	0.693	0.966	0.807	13.03	96.58
0.4	**0.731**	**0.924**	**0.816**	**19.21**	**92.44**
0.5	0.774	0.851	0.810	26.16	85.05
0.6	0.812	0.760	0.786	32.34	76.05
0.7	0.822	0.718	0.767	33.93	71.81
0.8	0.826	0.688	0.751	34.55	68.83
0.9	0.827	0.675	0.743	34.68	67.46
1	0.825	0.663	0.735	34.40	66.33

4.4.2 Hybrid Demographic Composition

The diversity gain has quantified the diversity in the recommendation list provided by the *hybrid* algorithm. However, it does not completely show the impact on each demographic parameter in the proposed list. Hence, we did another experiment where we form a PC using the *hybrid* algorithm with both binary and continuous demographic profiles from a pool of the current PC members and authors and compare it to the real PC. We use the following setting for the *hybrid* algorithm (α = 0.4, K= PC size) since α = 0.4 attains the best F-measure as indicated in the previous experiment. We report the percentage of each protected parameter for each feature across the real PC, and the PC formed using expertise recommendation (baseline), *hybrid* with binary profile, and *hybrid* with the continuous profile for each conference. The results show that our *hybrid* algorithm has increased the representation of the majority of demographic groups on average across the three conferences for both binary and continuous profiles with a slight gain when using continuous profiles. For instance, Figure 4.1 presents the demographic composition of the SIGCHI17 real and formed PCs. As we can see that female and ethnic minorities percentages have been largely boosted in the PC produced by our *hybrid* algorithm with both binary and continuous profiles. Nevertheless, the only demographic parameter that the real PC and *hybrid* have a smaller margin difference is the career stage. Moreover, our *hybrid* recommendation approach outperforms the baseline algorithm in all demographic parameters with a small expertise loss of 8%, which is a small penalty to increase the diversity. Furthermore, the use of continuous profile has slightly performed better than the binary profile in the case of female representation in SIGCHI17. However, it largely outperforms the binary for the case of the low-tier universities, where the binary has not even outperformed the real PC. In contrast to SIGCHI17 *hybrid*-proposed PC, the PC generated by SIGCOMM17 using binary profiles have slightly better demographic percentages as compared to *hybrid* with continuous as shown in Figure 4.2. Apart from this difference, the *hybrid* algorithm enhances the representation of the protected parameters across all features except the career stage as compared to the baseline and the real PC.

Testing our *hybrid* algorithm with SIGMOD17 does not bring new trends, where the *hybrid* with binary profile and continuous produce similar demographics in their proposed PC and they outperform the base and the real PC as shown in Figure 4.3.

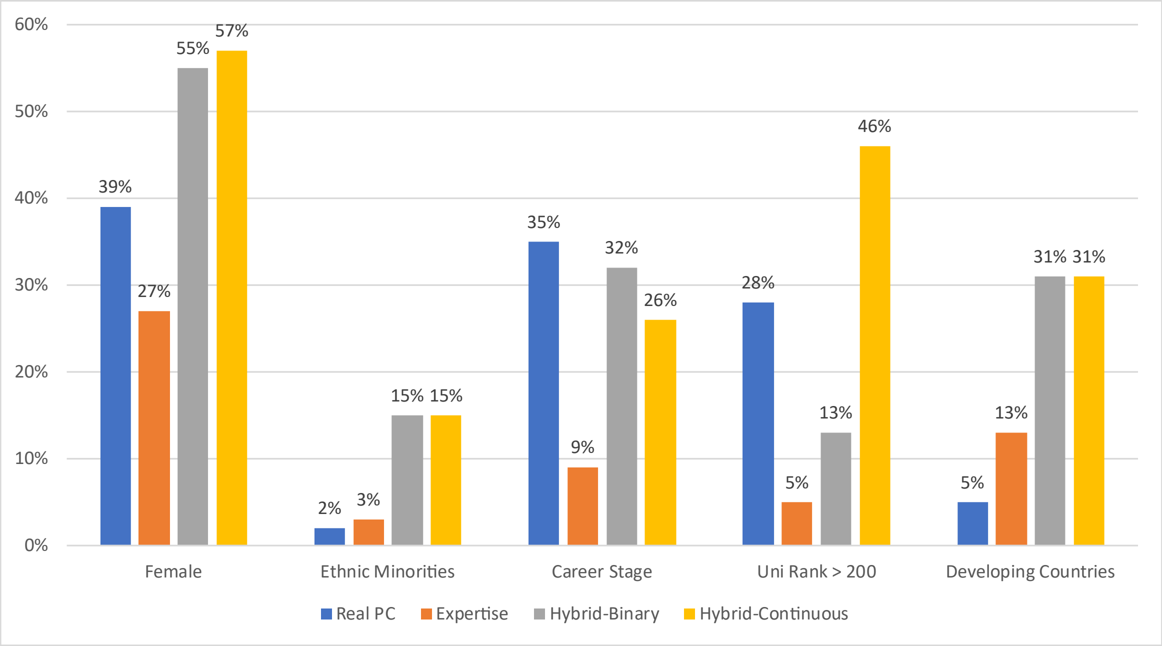

Figure 4.1 Demographic structure of real SIGCHI17 PC and PCs produced by expertise and hybrid with binary and continuous profiles [α = 0.4, K = PC size]

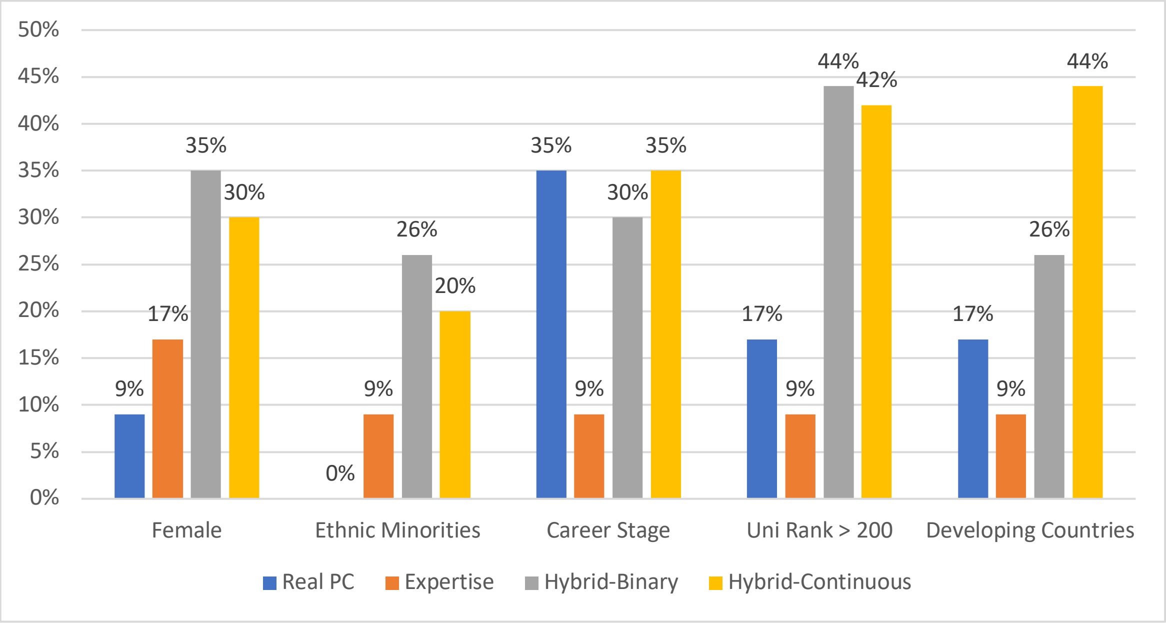

Figure 4.2 Demographic structure of real SIGCOMM17 and PCs produced by expertise and hybrid with binary and continuous profiles [α = 0.4, K = PC size]

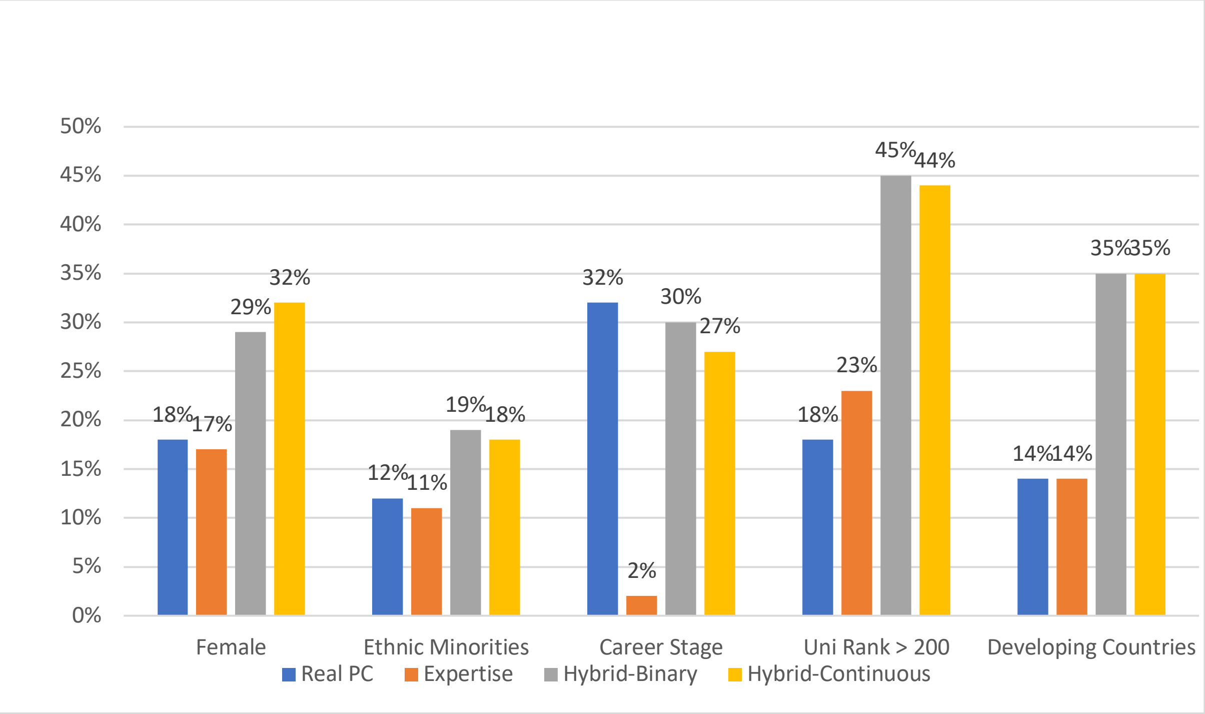

Figure 4.3 Demographic structure of real SIGMOD17 PC and PCs produced by expertise and hybrid al algorithms with binary and continuous profiles [α= 0.4, K = PC size]

Overall, many conclusions can be drawn from this experiment. First, we note that there is no significant difference in term of demographic diversity when using binary and continuous demographic profiles with the *hybrid* algorithm. Second, the *hybrid* algorithm has outperformed the baseline algorithm (i.e., expertise recommendation approach) when using binary or continuous profile and for all demographics. Third, the average expertise loss of the three conferences in our dataset is 13.81% for binary and 20.8% for continuous as shown in Table 4.7, this along with no significant difference in demographic diversity between indicates that binary profile performs better than the continuous profile with the *hybrid* algorithm. Lastly, the *hybrid* algorithm enhances the demographic diversity for all the protected parameters versus the real PC except the career stage. This is due to the fact that a *hybrid* algorithm score with $\alpha = 0.4$ has 60% preference toward diverse researcher with higher expertise (i.e., higher *h*-index).

Table 4.7 Hybrid utility loss [α = 0.4, K = PC size]

Conference	Utility Loss - Binary	Utility Loss - Continuous
SIGCHI17	13.91%	18.82%
SIGCOMM17	14.91%	22.98%
SIGMOD17	12.60%	20.61%
Average	13.81%	20.80%

4.5 Experiment 3: The Demographic Parity

The previous section has discussed the evaluation of how to maximize diversity by increasing the representation of the underrepresented and minority classes in a group. In this section, we evaluate the diversity through the statistical parity approaches presented in section 3.3. We evaluate the three proposed algorithms (*Hybrid*, *Voting Diversity*, and *Hill Climbing Diversity*) by first determining the best value of α that achieves the best F-measure. Second, we study and compare the demographic composition of the PCs generated by these algorithms.

4.5.1 Alpha Setting

In this experiment, we try to study the impact of α in our linear optimization formula implemented in these three algorithms aiming to determine the best α value that achieves the best possible balance between the diversity and the expertise scores. To make the score comparable among the conferences in our dataset, we select K = 50, where K is our ranking cutoff (i.e., the size of the recommendation). These algorithms recommend researchers from a pool of researchers that includes the current PC and the conference authors. First, we evaluate our algorithm using the binary demographic profile, and then we repeat the same experiment but with the continuous demographic profile of researchers. In both cases, we report expertise savings, diversity gain, and the F-measure to combine the two diversity and expertise gains into a single metric using α steps of 0.1 to find the best value of α since all three algorithms evaluated in this section have linear optimization as part of their design. An α value of 0.0 indicates that we select based on the expertise only algorithm and an α value of 1.0 indicates the ranking is based on the diversity score for that group. However, this concept has not the same generalization over the three algorithms since every algorithm implemented the linear optimization equation on a different subset of candidates. For example, the linear optimization in the general *hybrid* algorithm is being implemented on all the candidates at once. However, the case is not the same when it comes to the *Voting Diversity* approach as the algorithm first selects a subset for the feature that we want to choose a candidate from and then the linear optimization is used to calculate the score based on that. Finally, the *Hill Climbing Diversity* algorithm utilizes linear optimization every time a candidate got selected by getting the score from the gap distance as a diversity score and the expertise score based on the *normalized h-index*.

Table 4.8 and 4.9 present the performance of our algorithms when testing on our dataset. In these tables, we report the average of each metric over the three conferences that we have in our dataset. Both tables provide the expertise gain using nDCG, the diversity gain as measured by the Cumulative Proportionality Ratio (CPR), and the F-measure that represents that harmonic mean between the diversity and the expertise. The tables also show the diversity gain percentage as compared to the baseline (expertise recommendation approach) and the expertise savings.

As we can see in Table 4.8 and 4.9, the *hybrid* and *Hill Climbing Diversity* algorithms show the best performance when α approaches 0.4 when testing this with both binary and continuous demographic scholar profiles. This indicates that a nearly equal value of expertise and demographic diversity leads to the best result. However, the case is not the same with the *Voting Diversity* approach as it tends that the linear optimization does not help in improving the performance of this algorithm with both continuous and binary researcher’s profiles, where the best results is attained when the value of α is around 0. Although the continuous demographic profile provides a broad representation of the demographic segments of a researcher, the outcomes that we get from incorporating it does not seem to provide better diversity and expertise gains or show any new trends that is different from the binary profiles and as presented in Table 4.9, similar to the result of the previous experiment. Indeed, including the binary demographic profile has provided slightly better results, as provided by the F-measure. Overall, the *Hill Climbing Diversity* algorithm outperforms other algorithms when α approaches 0.4 making this the best possible balance between the diversity and the expertise gains that we get in our experiment. To put this in perspective, we present Figure 4.4 and Figure 4.5 to illustrate the expertise savings produced by our algorithm with respect to the change of the value of α. The change appears to be approximately linear with the *hybrid* algorithm slightly outperforming the *Hill Climbing Diversity* algorithm; however, when α = 0.4 the *Hill Climbing Diversity* algorithm attains better results with almost steady expertise savings while the *hybrid* is sharply declined. The *Voting Diversity* approach produces expertise savings less than other algorithms. Nevertheless, it produces better results than the *hybrid* when α is greater than 0.6. Additionally, The *Hill Climbing Diversity* algorithm has a minimal drop in the expertise with the change of α as compared to the *Voting Diversity* approach and the *hybrid* algorithm. We argue that a conclusion can be drawn here that is recommending a scholar by considering the demographic representation of the pool of candidates would minimize the expertise loss as compared to recommending a diverse researcher with respect to the population. This can be seen clearly when α = 1.0 (i.e., recommending based only the demographics profile of a researcher) the maximum expertise loss would not exceed 19% in the *Hill Climbing Diversity* algorithm compared to 55.16 % for the *hybrid*.

Table 4.8 Algorithm evaluation with binary demographic profile

α	Algorithm	EXP(nDCG)	CPR	F-Measure	Diversity Gain (%)	Expertise Savings (%)
0	Hybrid	1.000	0.368	0.532	0.000	100.000
	Voting	0.795	0.911	0.848	101.226	85.939
	Hill Climbing	1.000	0.368	0.532	0.000	100.000
0.1	Hybrid	0.998	0.391	0.558	6.882	99.845
	Voting	0.794	0.909	0.847	100.900	86.099
	Hill Climbing	0.994	0.446	0.608	20.255	99.769
0.2	Hybrid	0.985	0.420	0.586	14.267	99.121
	Voting	0.783	0.912	0.842	101.557	84.984
	Hill Climbing	0.944	0.625	0.743	60.179	96.942
0.3	Hybrid	0.934	0.488	0.639	29.316	95.534
	Voting	0.774	0.912	0.837	101.351	84.182
	Hill Climbing	0.874	0.822	0.846	90.366	91.275
0.4	Hybrid	0.849	0.581	0.689	46.101	89.071
	Voting	0.729	0.910	0.809	100.704	80.297
	Hill Climbing	0.825	0.902	**0.862**	**100.772**	**87.471**
0.5	Hybrid	0.664	0.701	0.679	65.428	75.279
	Voting	0.672	0.891	0.765	96.865	75.412
	Hill Climbing	0.808	0.920	0.860	102.962	86.172
0.6	Hybrid	0.465	0.797	0.586	83.044	61.943
	Voting	0.495	0.856	0.627	90.763	64.685
	Hill Climbing	0.781	0.928	0.848	104.397	83.949
0.7	Hybrid	0.357	0.827	0.496	88.602	54.352
	Voting	0.462	0.839	0.595	87.937	63.614
	Hill Climbing	0.769	0.931	0.842	105.370	83.492
0.8	Hybrid	0.309	0.828	0.444	88.737	51.184
	Voting	0.455	0.837	0.589	87.466	62.942
	Hill Climbing	0.765	0.899	0.826	99.650	83.763
0.9	Hybrid	0.309	0.829	0.444	88.875	51.136
	Voting	0.455	0.837	0.589	87.466	62.942
	Hill Climbing	0.751	0.886	0.813	99.754	82.852
1	Hybrid	0.190	0.900	0.313	99.847	44.854
	Voting	0.405	0.848	0.548	89.272	60.120
	Hill Climbing	0.735	0.891	0.805	100.968	81.988

.

Table 4.9 Algorithm Evaluation with continuous demographic profile

α	Algorithm	EXP(nDCG)	CPR	F-Measure	Diversity Gain(%)	Expertise Savings(%)
0	Hybrid	1.000	0.395	0.560	0.000	100.000
	Voting	0.781	0.901	0.835	87.453	78.077
	Hill Climbing	1.000	0.395	0.560	0.000	100.000
0.1	Hybrid	0.997	0.416	0.580	5.394	99.731
	Voting	0.776	0.899	0.831	86.981	77.599
	Hill Climbing	0.994	0.457	0.624	15.367	99.436
0.2	Hybrid	0.982	0.444	0.605	12.241	98.200
	Voting	0.772	0.896	0.828	86.619	77.244
	Hill Climbing	0.955	0.592	0.723	44.188	95.508
0.3	Hybrid	0.926	0.507	0.650	26.201	92.573
	Voting	0.757	0.895	0.818	86.092	75.739
	Hill Climbing	0.877	0.782	0.826	72.293	87.695
0.4	Hybrid	0.816	0.594	0.684	42.718	81.623
	Voting	0.729	0.889	0.799	85.116	72.851
	Hill Climbing	0.827	0.856	**0.841**	**81.971**	**82.722**
0.5	Hybrid	0.625	0.713	0.666	61.090	62.513
	Voting	0.673	0.880	0.760	83.396	67.334
	Hill Climbing	0.795	0.881	0.835	84.027	79.483
0.6	Hybrid	0.424	0.791	0.548	71.370	42.448
	Voting	0.565	0.844	0.675	77.292	56.462
	Hill Climbing	0.775	0.898	0.831	86.355	77.487
0.7	Hybrid	0.305	0.827	0.441	75.305	30.508
	Voting	0.466	0.815	0.593	71.815	46.644
	Hill Climbing	0.753	0.899	0.818	85.051	75.318
0.8	Hybrid	0.254	0.840	0.385	76.480	25.425
	Voting	0.436	0.813	0.567	71.819	43.588
	Hill Climbing	0.701	0.897	0.785	85.549	70.068
0.9	Hybrid	0.229	0.858	0.356	79.648	22.922
	Voting	0.427	0.813	0.559	72.027	42.667
	Hill Climbing	0.696	0.887	0.779	84.822	69.640
1	Hybrid	0.205	0.869	0.327	80.960	20.479
	Voting	0.424	0.812	0.556	71.671	42.380
	Hill Climbing	0.642	0.866	0.735	83.062	64.199

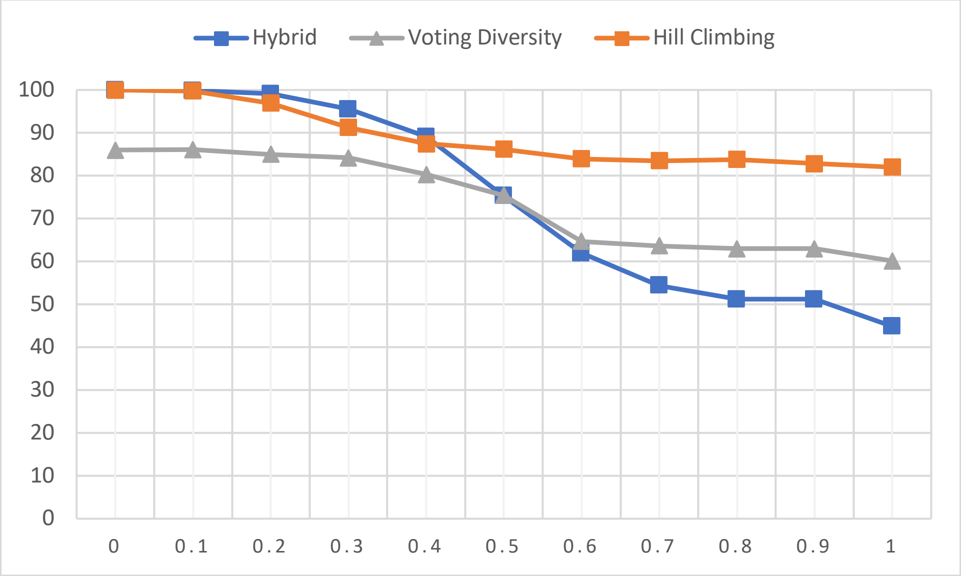

Figure 4.4 Impact of α on expertise savings – binary demographic profile

Testing with continuous demographic profile has not revealed major trends as presented in Fig 4.5. Nevertheless, we can notice that the expertise savings is less for all algorithms when testing with continuous demographic profile, giving the binary a bonus for the case. Moreover, we can see that the expertise savings sharply declines when α approaches 0.5 reaching its lowest point at α =1 with expertise savings of 20%.

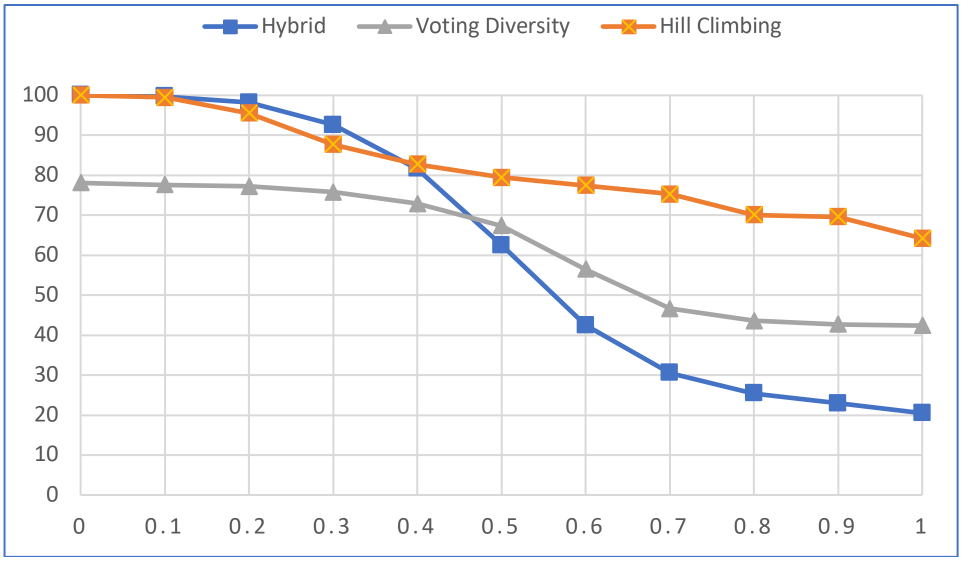

Figure 4.5 Impact of α on expertise savings – continuous demographic profile

Figure 4.6 presents the impact of the change of α on the diversity gain for the proposed algorithms. The trends can be seen as the opposite of the ones we discuss in the expertise gain since α is a rewarding parameter for the diversity score in our linear optimization formula. However, this can be valid for both the *hybrid* and to a certain extent *Hill Climbing Diversity* algorithm, but with the *Voting Diversity* approach the change occurs at a slower rate with no major impact of α. This behavior can be justified as that the core concept of the *Voting Diversity* approach produces a demographic subgroup based on the winning demographic feature before applying the linear optimization to balance the expertise and the diversity for that subgroup, and hence the demographic proportion is relatively maintained. Nevertheless, the *Hill Climbing Diversity* algorithm shows best diversity gain among all algorithms, and this occurs when α is 0.4, a nearly equal amount contribution of both diversity and expertise. This balance, along with having the best diversity gain at α around 0.4, results in having the best F-measure among all algorithms. The trends presented in Fig 4.7 when testing the same with continuous demographic profile do not present a different conclusion. However, the only difference that we can notice is that the diversity gain is slightly less than with the binary, making the binary demographic representation of researchers the preferable profile to be used with these algorithms.

Nevertheless, we note that the difference between the *Hill Climbing Diversity* and *Voting Diversity* is not major when it comes to the score of F-measure and that explains why that result is not statistically significant ($p = 0.33$ for binary, and $p = 0.38$ for continuous). However, both algorithms are statistically significantly ($p < 0.05$) better than the *Hybrid* algorithm. Nevertheless, when it comes to expertise savings, the *Hill Climbing Diversity* algorithm statistically significantly ($p < 0.05$) outperforms the other algorithms.

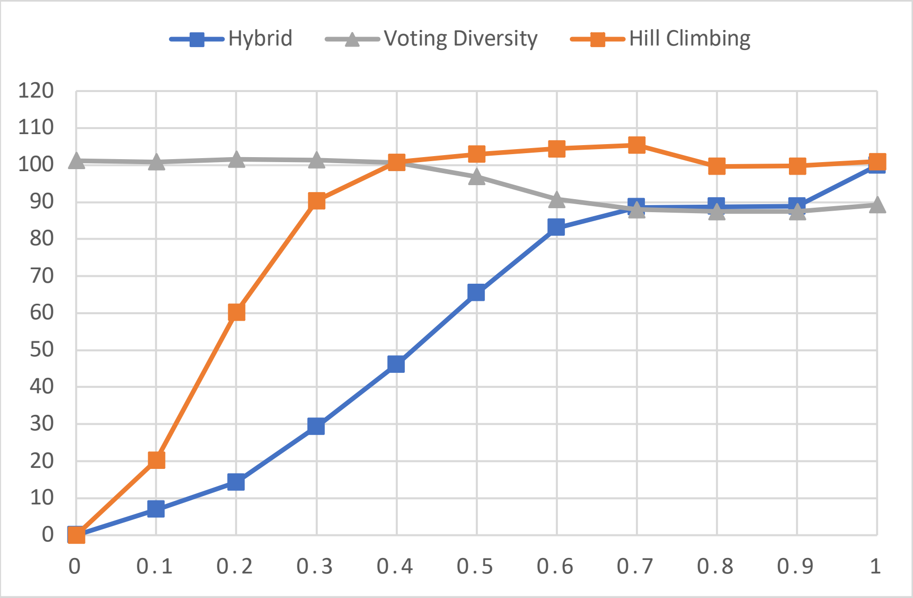

Figure 4.6 Impact of α on diversity gain (%) – binary demographic profile

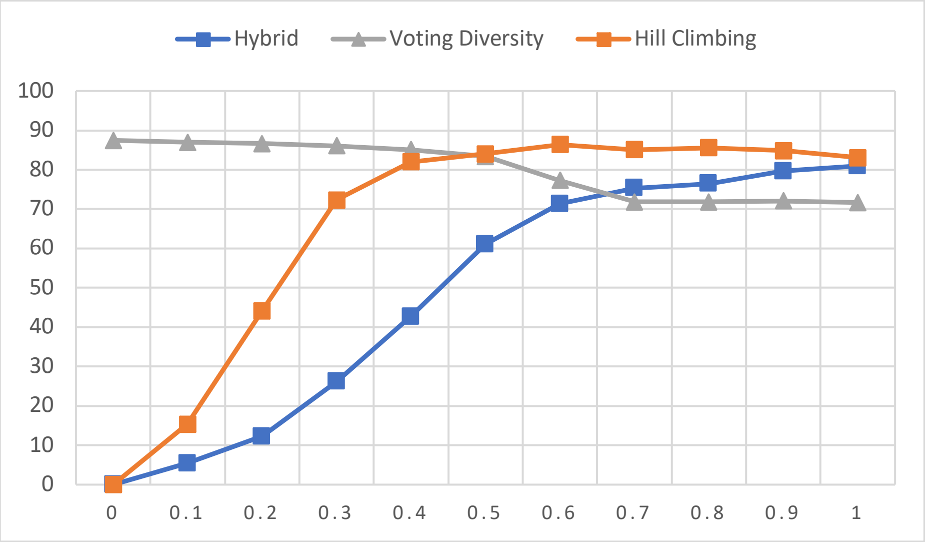

Figure 4.7 Impact of α on diversity gain (%) – continuous demographic profile

4.5.2: Demographic Composition and Parity Matching

4.5.2.1 Experiment Settings

The previous experiment determined the best α setting for our algorithms. Although the CPR metric quantified how well these algorithms produce recommendation lists that match the distribution of the pool, we still need to understand and analyze the demographic composition of the PC and whether it leads to a diverse and better demographic representation as compared to the existing PCs in our dataset. Hence, we did another experiment by recommending a number of researchers that have the same size of the PC members for each conference. We illustrate and discuss the change in the representation per each demographic feature and for each conference in our dataset. We focus on the protected group in each feature to investigate how well its representation is and as provided by each algorithm. We compare the distribution of each feature provided by our proposed algorithm to the feature distribution in the researchers' pool. We choose the value of α to be 0.4 since it is the setting that provides the best possible balance between the diversity and expertise and as illustrated in the previous section.

4.5.2.2 Results

Figure 4.8 shows the demographic distribution for the SIGCHI 17 as provided by each algorithm. As we can see the *hybrid* (balanced) approach overrepresents the protected groups as compared to the distribution of the pool which leads to a negative bias. Moreover, the result for this algorithm shows that it does not guarantee that all demographic features would be well represented, and some could be underrepresented with respect to the pool distribution (e.g., career stages as in Fig 4.8). In contrast, we can see that both the *Voting Diversity* and *Hill Climbing Diversity* approaches guarantee a fair representation of each feature that is almost equivalent to its representation in the researcher's pool with a slight advantage for the *Hill Climbing Diversity* algorithm. For example, the *Hill Climbing Diversity* algorithm produces a SIGCHI PC composed of 40% junior researchers, a distribution similar to that in the pool. Similarly, the *Voting Diversity* approach provides a PC of 38% junior researchers, only slightly less than the pool distribution. Nevertheless, the *Voting Diversity* approach overrepresents members of ethnic minorities and researchers from developed countries as compared to the distribution of the pool. Additionally, this comes at the cost of the expertise (as shown in Table 4.8 and 4.9). Nevertheless, the *Hill Climbing Diversity* algorithm shows a promising result as it guarantees that the PC distribution it generates is similar to the statistical parity of the pool as shown in Fig 4.8. Moreover, the *Hill Climbing Diversity* algorithm attains the best balance between the diversity and the expertise among all algorithms as discussed in the previous section. Additionally, the result shows a better representation for the protected parameters as compared to the current PC. However, a difference can be seen with respect to the geolocation, where the current PC has a higher representation as compared to the distribution of the researchers from developed countries in the pool.

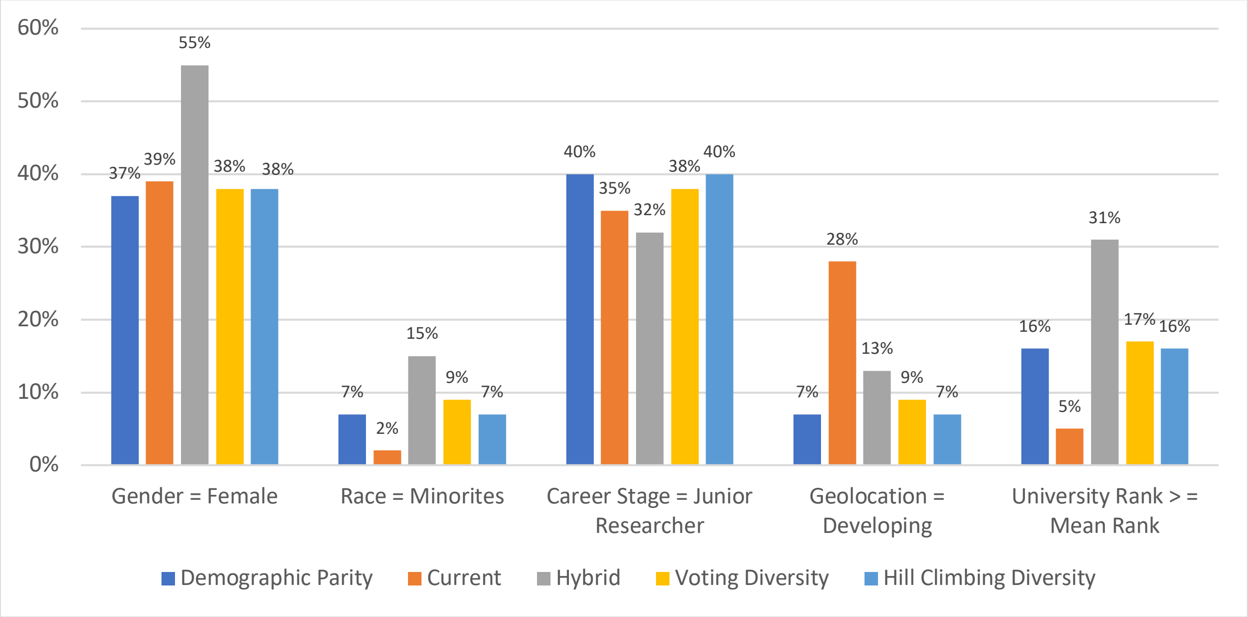

Figure 4.8 Demographic difference between the real PC, pool distribution, and the proposed algorithms with α = 0.4 for SIGCHI17 binary profiles

Likewise, we repeated the same experiment, but with the continuous demographic profile as a researcher profile. As shown in Figure 4.9, the results with the continuous profile reflect the same trends, i.e., the protected parameter is again overrepresented in the balanced approach, and the *Hill Climbing Diversity* algorithm outperforms other algorithms in terms of the balance between diversity and expertise. In contrast to the result of the binary profile, the continuous profile increased the participation of researchers from the developing countries, overrepresenting them in the proposed PCs. The only feature for which the continuous profile falls short is career stage, where the recommended junior researchers are less represented than in the pool or the existing PC.

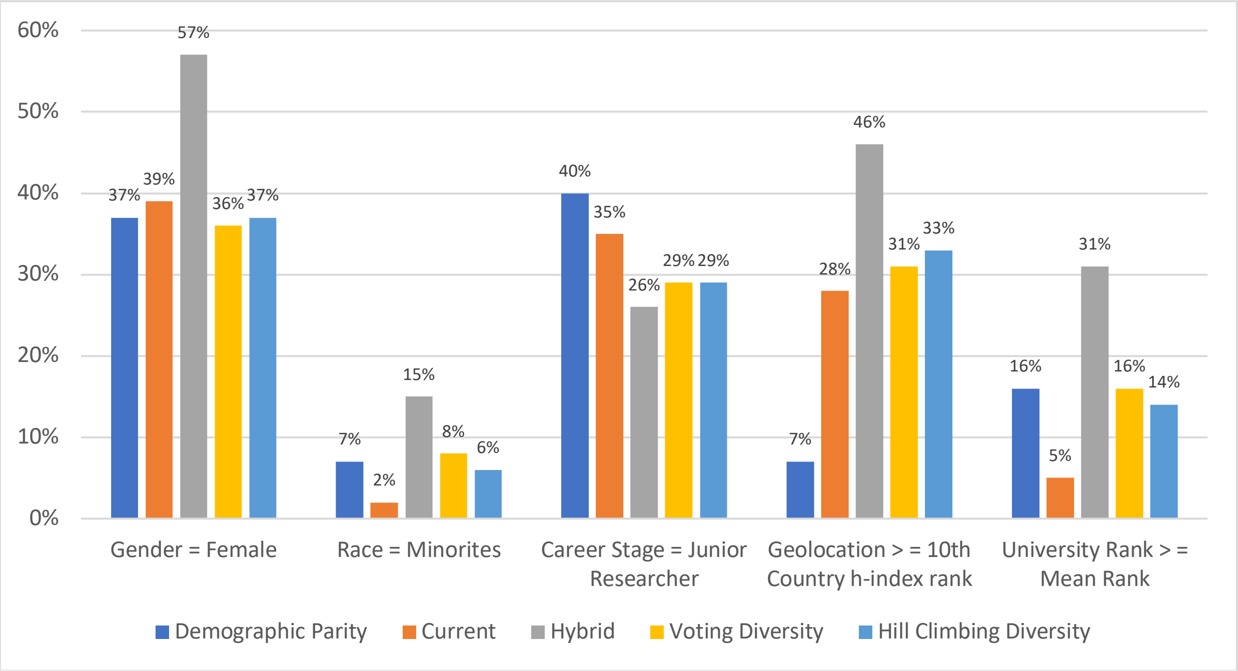

Fig 4.9 Demographic difference between the real PC, pool distribution, and the proposed algorithms with α = 0.4 for SIGCHI17 continuous profiles

To draw a solid conclusion about the performance of our algorithms, we evaluated them on the other conferences in our dataset (SIGCOMM17, and SIGMOD17) with both the binary and continuous researcher's profile. The results for the binary profile, presented in Figures 4.10 and 4.12 respectively, show similar trends to the ones for SIGCHI17 that we discussed in the previous section. The only difference of note is the representation from the developing countries, where the *Hill Climbing Diversity* approach shows trends similar to the distribution of the pool and does not maximize this feature as with SIGCHI17. The results of evaluating our algorithms with the continuous profiles of SIGCOMM17 and SIGMOD17 can be seen in Fig 4.11 and 4.13, respectively. The trends in these two graphs confirm the conclusion seen when evaluating with

SIGCHI17.

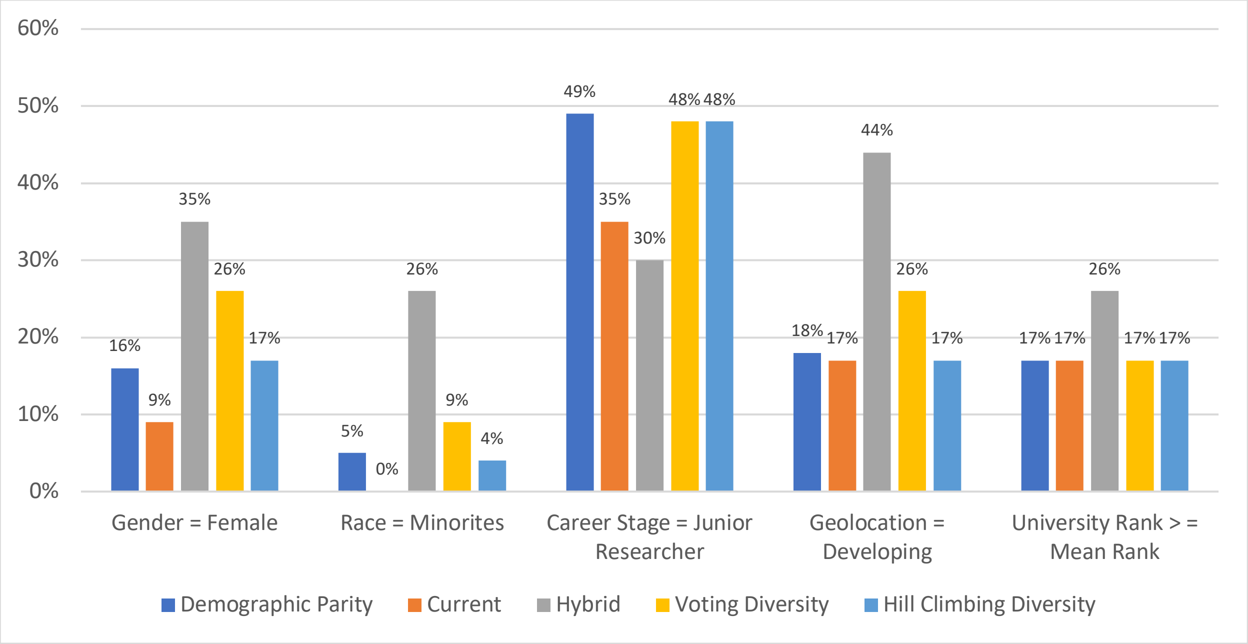

Figure 4.10 Demographic difference between the real PC, pool distribution, and the proposed algorithms with α = 0.4 for SIGCOMM17 binary profiles

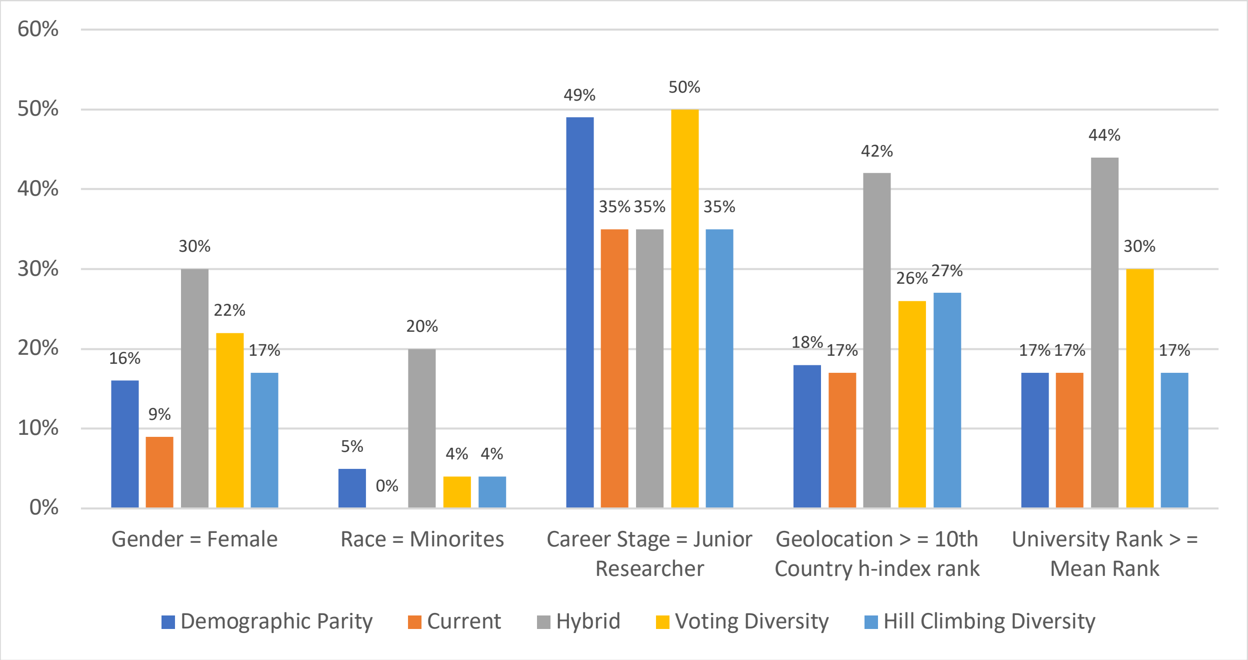

Figure 4.11 Demographic difference between the real PC, pool distribution, and the proposed algorithms with α = 0.4 for SIGCOMM17 continuous profiles

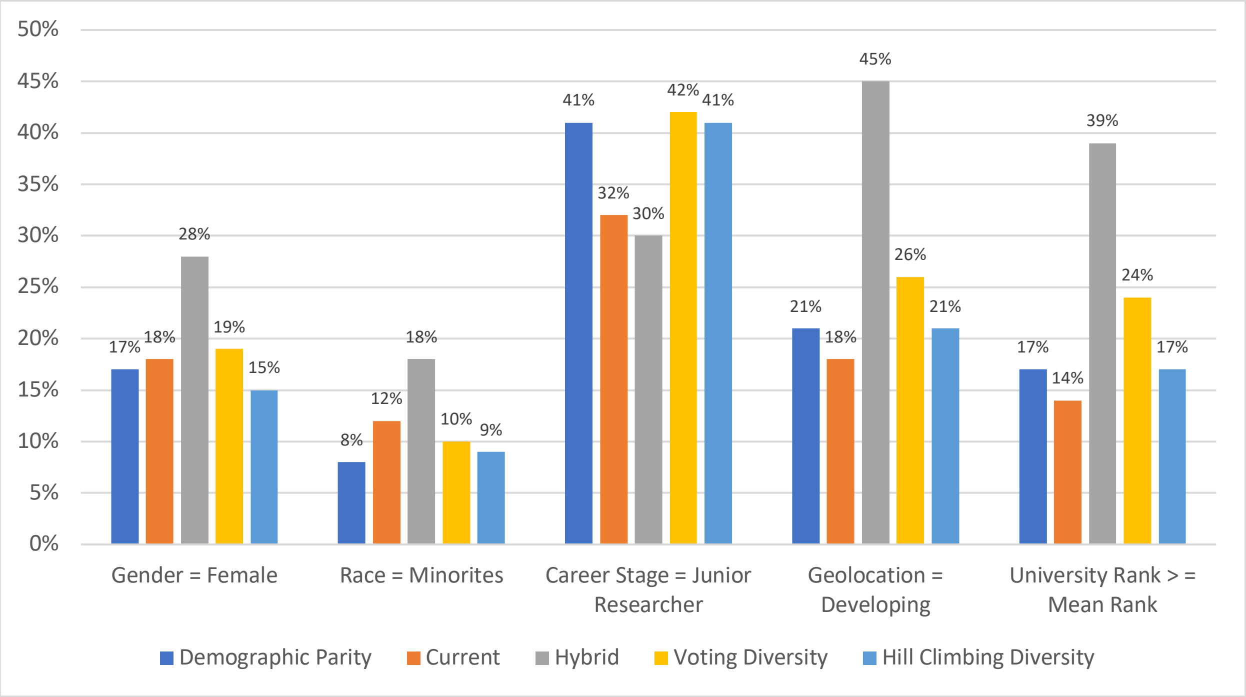

Figure 4.12 Demographic difference between the real PC, pool distribution, and the proposed algorithms with α = 0.4 for SIGMOD17 binary profiles

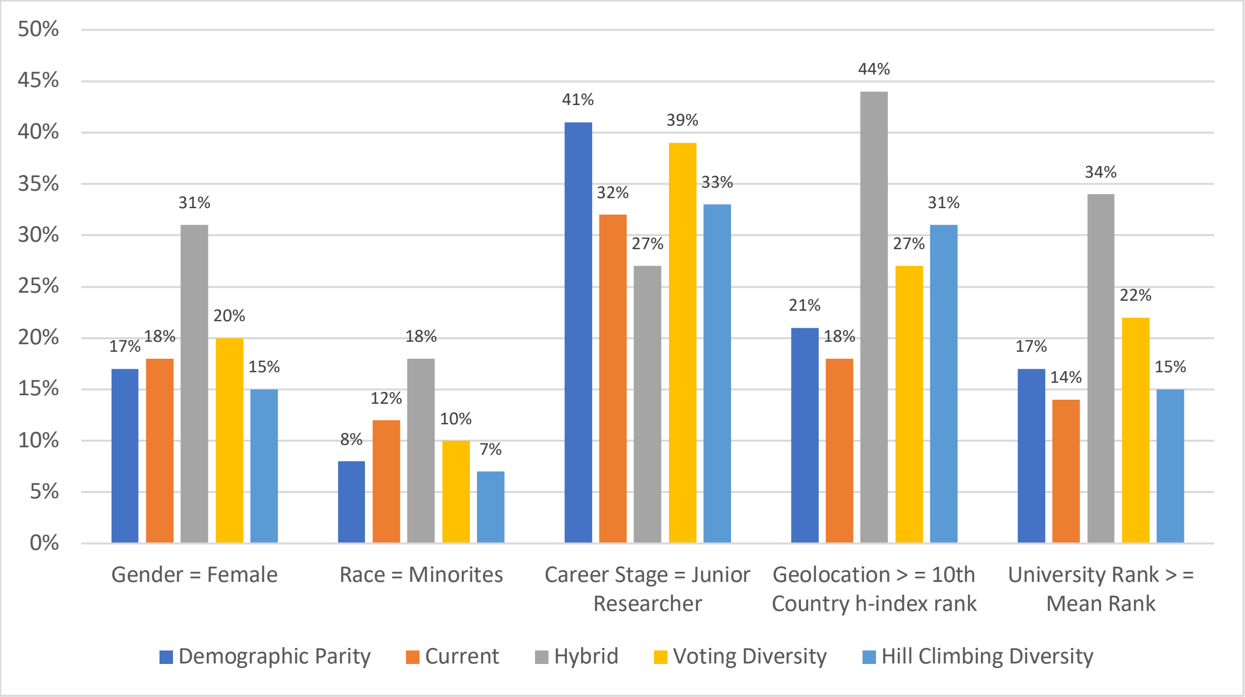

Figure 4.13 Demographic difference between the real PC, pool distribution, and the proposed algorithms with α = 0.4 for SIGMOD17 continuous profiles

In summary, all algorithms provide more diverse demographic representations as compared to the existing PC. Second, the *Hill Climbing Diversity* algorithm shows the best results with respect to balancing between diversity and expertise and provided the closest demographic parity. This result is statically significant when comparing this to *hybrid* algorithm, but is it is not when comparing this to *Voting Diversity* approach. Nevertheless, *Hill Climbing Diversity* outperforms other algorithms with a statistically significant result with respect to expertise savings. Third, the *hybrid* approach increases the diversity of the demographic segments in the recommendation list; however, there is no guarantee that it would achieve the statistical parity. Finally, incorporating the continuous demographic profile can lead to overrepresentation of some demographic segments (e.g., geolocation) and lead to a possible decrease in the expertise; that makes the binary profile a better choice to achieve the best balance between the expertise and the diversity, the main goal in this research.

5 Conclusions and Future Work

In this chapter, we provide our research contributions to the fields of expert recommendation, fair ranking, group formation, and algorithmic fairness. We also discuss possible future research directions suggested by our results.

5.1 Conclusions

Expert recommendation has become a prominent sub-discipline of the information retrieval field. Its importance comes from the exponential growth of the available knowledge and how it is challenging to identify experts in such a rapidly evolving world where innovation and creativity determine the success of a society. In this research, we study some of the challenges of expert recommendation systems and introduce algorithms to tackle these issues. We address these problems by answering three research questions that serve as guidelines to our work: 1) how to model the expertise and the demographics of a researcher? 2) How can we maximize the representation of protected groups at the top of a ranked list of recommended experts? 3) How to produce a recommendation that reflects the demographic parity of the recommended expert population. The answers to these questions are the core of our work. By answering these questions, we have made contributions in four areas, namely, expert recommendation, fair ranking, group formation, and algorithmic fairness.

To summarize, we made several contributions, described in more detail below:

- We created a dataset of 1217 demographic profiles for researchers in academia.
- Focusing on researcher modeling, we have quantified researcher's expertise by using one of the most well-respected bibliometric that is *h*-index.
- We have also identified five major demographic features that have been shown to be sources of explicit or implicit bias in academia, i.e., gender, race, career stage, academic

institution ranking, and affiliation geolocation. These features have been used to develop a more comprehensive way to represent demographics in researcher profiles We evaluated two ways to represent the demographic profile of a researcher by having binary and continuous weights for the demographic features in an expert profile. We use these demographic features within an expert recommender system in academia to achieve fairness and increase demographic diversity.

- We develop three scholar recommendation algorithms: 1) the expertise model; 2) a new diversity model; and 3) a hybrid approach that balances diversity gains against loss of expertise. We consider a specific example of expert recommendation in academia that is recommending researchers to join a conference program committee and we consider the expertise model as our baseline algorithm. We developed a new evaluation metric to measure the diversity in the proposed ranking by each algorithm that is the Multi-dimensional non-Discounted Cumulative Gain (mnDCG), that measures gain across multiple dimensions. We have also used the nDCG (Discounted Cumulative Gain) to report the expertise gain in the proposed recommendation list. Our evaluation shows our diversity approach provides a better diversity gain; however, this comes with the cost of expertise. Hence, we developed a hybrid recommender system that incorporates linear optimization through a tuning parameter (α). Our evaluation shows that the best value for (α) is approximately 0.4, i.e., 40% weight to the diversity recommendation and 60% weight to the expertise recommendation, with 92.44% of expertise saving and nearly 20% of diversity gain. The results are statistically significant; however, comparing binary and continuous results does not produce a highly significant difference.

- We introduced new expert recommendation algorithms by exploring another concept of diversity that is *diversity by proportionality*, where the goal is to find a representative recommendation at any given rank that reflects the demographic parity of the recommendation pool. To achieve that, we develop two algorithms, the Hill Climbing Diversity, and Voting Diversity. The Hill Climbing Diversity algorithm utilizes mathematical optimization techniques to find the demographic gap among the features with respect to the proposed expert group and the demographic parity of the recommendation pool. The other algorithm, Voting Diversity, has been inspired by the problem of seat allocation in the electoral voting system where voting is similar to the feature ratio in the recommendation pool and the number of seats is the size of the recommendation. We evaluated these algorithms on the same dataset of the three ACM conferences by using the Cumulative Proportionality measure (CPR) as a diversity metric and nDCG as expertise metric and we reported the expertise and diversity gains with F-measure as the balance between the diversity and the expertise and we compare the performance of these two algorithms to the hybrid algorithm from the previous goal. The results show that all algorithms provide a better demographic diversity as compared to the existing PC. We have also observed that Hill Climbing Diversity and Voting Diversity algorithms guarantee the demographic parity with a slight advantage for the voting diversity. We have also seen that the use of continuous profile has increased the representation of some features as compared to the binary, which leads the binary to be a better choice when it comes to maintaining the demographic parity. Finally, we find that Hill-Climbing shows the best balance between the diversity and the expertise making it the preferred algorithm among the proposed approaches with expertise saving of approximately 83% and F-measure of 0.841 compared to 0.799 for voting approach and 0.684 for the hybrid approach.

Finally, we believe that our proposed algorithms can be applied in any expert recommendation setting and it is not only limited to academia considering the required change in the expert demographics and expertise profiles. We consider this research is the first step towards addressing the issue of equity and diversity in the design of recommendation algorithms. The research supports civil rights laws designed to prevent discrimination so that all individuals have equal opportunity when applying for a job, financing, credit card, or other opportunities regardless of gender, race, or disability. Through this research, we provided the first step towards tools that will allow identifying the demographic parameters that are considered major sources of bias and ensure that minorities are well represented and have the same advantages that the majorities have. This work also provides new algorithms to form diverse groups, supported by studies that show that a diverse team can produce better outcomes. The research promotes the idea of allowing young researchers and scholars from developing countries or states that have less access to research funding, to share their work and build their professional networks. This has the potential to positively impact their academic development and enrich their research communities. We believe that this research contributes to the development of diverse communities that provide equal opportunities to all members. We trust that creating such an environment will enrich cultural exchange and positively impact the individuals and their communities.

5.2 Future Work

We believe that there is a potential for this work to be extended in different domains. First, the demographic profile design can be further developed to have a wide range of demographic features. We also see that the demographic profile can be customized based on the environment where it applies. For example, the protected feature for gender in education is female, while in nursing for example is male Also, we plan to study the demographic composition of different academic conferences in other domains. We also believe that our algorithms can be further evaluated in many expert recommendation areas and it is not only limited to the educational environment. We also plan to test different proportional representation methods and evaluate the impact on the overall performance of our proposed algorithms.

6 References

[1] Winans, M., Faupel, D., Armstrong, A., Henderson, J., Valentine, E., McDonald, L., Walters, D., Waite, J., Trapani, M., & Magill, E. (2016). 10 key marketing trends for 2017 customer expectations and ideas for exceeding customer expectations. ftp://ftp.www.ibm.com/software/in/pdf/10_Key_Marketing_Trends_for_2017.pdf

[2] Coles, P., Cox, T., Mackey, C., & Richardson, S. (2006). The toxic terabyte-how data dumping threatens business efficiency. *IBM Global Technical Services*

[3] Reinhardt, W., Schmidt, B., Sloep, P., & Drachsler, H. (2011). Knowledge worker roles and actions—results of two empirical studies. *Knowledge and Process Management*, *18*(3), 150-174.

[4] Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., & Si, L. (2012). Expertise retrieval. *Foundations and Trends® in Information Retrieval*, *6*(2–3), 127-256.

[5] Yimam-Seid, D., & Kobsa, A. (2003a). Expert-finding systems for organizations: Problem and domain analysis and the DEMOIR approach. *Journal of Organizational Computing and Electronic Commerce*, *13*(1), 1-24.

[6] Chandrasekaran, K., Gauch, S., Lakkaraju, P., & Luong, H. P. (2008). Concept-based document recommendations for citeseer authors. Paper presented at the *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, 83-92.

[7] Price, S., & Flach, P. A. (2017). Computational support for academic peer review: A perspective from artificial intelligence. *Communications of the ACM*, *60*(3), 70-79.

[8] Hettich, S., & Pazzani, M. J. (2006). Mining for proposal reviewers: Lessons learned at the national science foundation. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 862-871).

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

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

[11] Becerra-Fernandez, I. (2006). Searching for experts on the web: A review of contemporary expertise locator systems. *ACM Transactions on Internet Technology (TOIT)*, *6*(4), 333-355.

[12] Craswell, N., Hawking, D., Vercoustre, A., & Wilkins, P. (2001). P@ noptic expert: Searching for experts not just for documents. Paper presented at *the Ausweb Poster Proceedings, Queensland, Australia* (Vol. 15, p. 17).

[13] Reichling, T., & Wulf, V. (2009). Expert recommender systems in practice: Evaluating semi-automatic profile generation. Paper presented at the *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 59-68.

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

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

[16] Pagel, P. S., & Hudetz, J. A. (2011). H-index is a sensitive indicator of academic activity in highly productive anaesthesiologists: results of a bibliometric analysis. *Acta Anaesthesiologica Scandinavica, 55*(9), 1085-1089.

[17] Tang, L., & Hu, G. (2018). Evaluation woes: Metrics can help beat bias. *Nature*, *559*(7714), 331-332.

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

[19] Al-Shamri, M. Y. H. (2016). User profiling approaches for demographic recommender systems. *Knowledge-Based Systems, 100*, 175-187.

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

[21] Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial intelligence review*, *13*(5), 393-408.

[22] Michael, J. (2007). 40000 namen, anredebestimmung anhand des vornamens. *C'T,* 182-183.

[23] Perez, I. S. (2021). *Gender-guesser*. https://pypi.python.org/pypi/gender-guesser

[24] Vanetta, M. (2021). *Gender Detector*. https://github.com/malev/gender-detector

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

[26] Ye, J., Han, S., Hu, Y., Coskun, B., Liu, M., Qin, H., & Skiena, S. (2017). Nationality classification using name embeddings. Paper presented at the *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management,* 1897-1906.

[27] Strømgren, C. (2021). *Genderize.io*. https://genderize.io/

[28] Carsenat, E. (2013). Onomastics for Business: can discrimination help development? *ParisTech Review*.

[29] Menéndez, D. A., González-Barahona, J. M., & Robles, G. (2020, April). Damegender: Writing and Comparing Gender Detection Tools. In *SATToSE*.

[30] Brocco, M., Hauptmann, C., & Andergassen-Soelva, E. (2011, August). Recommender system augmentation of HR databases for team recommendation. In *2011 22nd International Workshop on Database and Expert Systems Applications* (pp. 554-558). IEEE.

[31] Mathieu, J. E., Tannenbaum, S. I., Donsbach, J. S., & Alliger, G. M. (2014). A review and integration of team composition models: Moving toward a dynamic and temporal framework. *Journal of Management, 40*(1), 130-160.

[32] Rabanca, G. (2017). *Approximation Algorithms for Effective Team Formation* (Doctoral dissertation). https://academicworks.cuny.edu/gc_etds/2353/

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

[34] Chhabra, M., Das, S., & Szymanski, B. (2013). Team formation in social networks. *Computer and Information Sciences III,* 291-299.

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

[36] Neshati, M., Beigy, H., & Hiemstra, D. (2014). Expert group formation using facility location analysis. *Information Processing & Management, 50*(2), 361-383.

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

[38] Palubeckis, G., Ostreika, A., & Rubliauskas, D. (2015). Maximally diverse grouping: An iterated tabu search approach. *Journal of the Operational Research Society, 66*(4), 579-592.

[39] Gallego, M., Laguna, M., Martí, R., & Duarte, A. (2013). Tabu search with strategic oscillation for the maximally diverse grouping problem. *Journal of the Operational Research Society, 64*(5), 724-734.

[40] Lai, X., & Hao, J. K. (2016). Iterated variable neighborhood search for the capacitated clustering problem. *Engineering Applications of Artificial Intelligence, 56*, 102-120.

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

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

[43] Maass, K. L., Lo, V. M. H., Weiss, A., & Daskin, M. S. (2015). Maximizing diversity in the engineering global leadership cultural families. *Interfaces, 45*(4)*,* 293-304.

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

[45] US Equal Employment Opportunity Commission (2019). *Facts about Discrimination in Federal Government Employment Based on Marital Status, Political Affiliation, Status as a Parent, Sexual Orientation, and Gender Identity.* https://www.eeoc.gov/federal/otherprotections.cfm

[46] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259-268). ACM.

[47] Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. Paper presented *at the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 35-50.

[48] Kamiran, F., Žliobaitė, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems, 35*(3), 613-644.

[49] Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. Paper presented at *the Proceedings of the 26th International Conference on World Wide Web*, 1171-1180.

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

[51] Bornmann, L., & Daniel, H. D. (2004). Reliability, fairness and predictive validity of committee peer review. *BIF Futura*, *19*, 7-19.

[52] Perna, L. W., Gerald, D., Baum, E., & Milem, J. (2007). The status of equity for black faculty and administrators in public higher education in the south. *Research in Higher Education, 48*(2), 193-228.

[53] McConner, M. (2014). Black women in leadership: An assessment of the gender inequality and racism that exists among black women leaders in higher education. *The Journal Higher Education Management, 29*(1), 78-87.

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

[55] Lerback, J., & Hanson, B. (2017). Journals invite too few women to referee. *Nature, 541*(7638), 455-457. doi:10.1038/541455a

[56] Murray, D., Siler, K., Lariviére, V., Chan, W. M., Collings, A. M., Raymond, J., & Sugimoto, C. R. (2018). Gender and international diversity improves equity in peer review. *BioRxiv*, 400515.

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

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

[59] Publons. (2018). *Global State of Peer Review.* https://publons.com/static/Publons-Global-State-Of-Peer-Review-2018.pdf

[60] Freeman, R. B., & Huang, W. (2015). Collaborating with people like me: Ethnic co-authorship within the united states. *Journal of Labor Economics, 33*(S1), S289-S318.

[61] AlShebli, B. K., Rahwan, T., & Woon, W. L. (2018). Ethnic diversity increases scientific impact. *arXiv Preprint arXiv:1803.02282.*

[62] Dang, V., & Croft, W. B. (2012, August). Diversity by proportionality: an election-based approach to search result diversification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (pp. 65-74).

[63] Pedreshi, D., Ruggieri, S., & Turini, F. (2008, August). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 560-568).

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

[65] Conductscience. (November 7, 2019). *Measuring Research Success: H-index Scores in Science*. https://conductscience.com/measuring-research-success-h-index-scores-in-science/

[66] Tang, L., & Hu, G. (2018). Evaluation woes: Metrics can help beat bias. *Nature*, *559*(7714), 331-332.

[67] Schreiber, W. E., & Giustini, D. M. (2019*). Measuring scientific impact with the h-index: a primer for pathologists. American journal of clinical pathology*, *151*(3), 286-291.

[68] Gallo, S. A., Thompson, L. A., Schmaling, K. B., & Glisson, S. R. (2020). The participation and motivations of grant peer reviewers: a comprehensive survey. *Science and engineering ethics*, *26*(2), 761-782.

[69] Mattauch, S., Lohmann, K., Hannig, F., Lohmann, D., & Teich, J. (2020). A bibliometric approach for detecting the gender gap in computer science. *Communications of the ACM*, *63*(5), 74-80.

[70] Salman,O., Gauch, S., Alqahatni, M., & Ibrahim, M. (2020). The Demographic Gap in Conference Program Committees. In *the Proceeding of the 17th International Conference on Applied Computing (AC 2020)*, (pp. 46-52).

[71] National Science Board. (2020). *The State of U.S. Science and Engineering 2020.* https://ncses.nsf.gov/pubs/nsb20201/u-s-and-global-education

[72] Times Higher Education. (2020). *World University Rankings 2019 by subject: computer science*. https://www.timeshighereducation.com/world-university-rankings/2019/subject-ranking/computer-science

[73] United Nations (2019). *World Economic Situation and Prospects 2019*. https://www.un.org/development/desa/dpad/wp-content/uploads/sites/45/WESP2019_BOOK-web.pdf

[74] SJR. (2020). *SJRScimago Journal & Country Rank*. https://www.scimagojr.com/countryrank.php?area=1700

[75] Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

[76] Gallagher, M. (1991). Proportionality, disproportionality and electoral systems. *Electoral studies*, *10*(1), 33-51.

[77] Balinski, M. L., & Young, H. P. (1978). The Jefferson method of apportionment. *Siam Review*, *20*(2), 278-284.

[78] Aminer.(2020). *Conference Rank: Computer Science*. https://www.aminer.cn/ranks/conf

[79] ACM. (2021). *About the ACM Organization*. https://www.acm.org/about-acm/about-the-acm-organization

[80] Järvelin, K., & Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. Paper presented at the *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 41-48.

[81] Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39, pp. 234-265). Cambridge: Cambridge University Press.

[82] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005, August). Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning* (pp. 89-96).

[83] Van Rijsbergen, C. J. (1979). Information Retrieval . Butterworth-Heinemann.