Proportionality and Fairness in Voting and Ranking Systems

by

Kanav Mehra

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2023

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Part of this thesis includes first-authored content from the following published article.

Kanav Mehra, Nanda Kishore Sreenivas, and Kate Larson. Deliberation and Voting in Approval-Based Multi-Winner Elections. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence, IJCAI 2023*.

The work in this paper was a collaborative effort. As the first author, I was responsible for the overall conceptualization, formalization, methodology, research, implementation, and writing. Nanda specifically contributed to the conceptualization, formalization, and implementation of the deliberation model, along with assistance in overall writing and reviewing. Kate contributed with overall supervision and conceptualization. Chapter 3 is an extended version of this paper.

Abstract

Fairness through proportionality has received significant attention in recent social choice research, leading to the development of advanced tools, methods, and algorithms aimed at ensuring fairness in democratic institutions.

Citizen-focused democratic processes where participants deliberate on alternatives and then vote to make the final decision are increasingly popular today. While the computational social choice literature has extensively investigated voting rules, there is limited work that explicitly looks at the interplay of the deliberative process and voting. In this thesis, we build a deliberation model using established models from the opinion-dynamics literature and study the effect of different deliberation mechanisms on voting outcomes achieved when using well-studied voting rules. Our results show that deliberation generally improves welfare and representation guarantees, but the results are sensitive to how the deliberation process is organized. We also show, experimentally, that simple voting rules, such as approval voting, perform as well as more sophisticated rules such as proportional approval voting or method of equal shares if deliberation is properly supported. This has ramifications on the practical use of such voting rules in citizen-focused democratic processes.

Intricately designed proportional voting rules offer robust theoretical and axiomatic fairness guarantees that can prove valuable in similar scenarios beyond the realm of elections. In the second part, we capitalize on these properties and introduce innovative fair-ranking algorithms based on proportional voting methods. Specifically, we define the general task of fair ranking, which involves generating a list of items that is fairly ordered with respect to a given query, as a voting problem. Our findings reveal that proportional voting rules deliver exceptional performance, frequently matching or surpassing the performance of existing benchmarks in terms of aggregate fairness and relevance metrics. These discoveries present exciting avenues for further research and applications, endorsing the widespread adoption of proportional voting rules in domains where fairness is a priority.

Acknowledgements

First, I would like to thank my advisor Professor Kate Larson. Throughout this journey, she has been incredibly supportive, positive, and inspiring. Thanks for giving me the freedom to explore my interests while guiding me through the process. I started this journey as a novice researcher, and Kate's guidance has definitely made me a better researcher. She does not just support you through the research journey but also shapes you to be a better researcher as a result of the process. I would like to sincerely thank her for her presence and motivation – I could not have asked for a better advisor.

Second, I would like to thank my committee members – Professor Edith Law and Professor Robin Cohen, for agreeing to read the thesis and providing insightful feedback.

I would like to sincerely thank my lab members – Nanda, Dave, Ben, Marvin, Valerie, Wei, and Sriram for their constant support and feedback. Our weekly group meeting discussions that often revolved around both academic and random thoughts were an integral part of this journey.

Finally, and most importantly, I would like to thank my family – my parents and my brothers for being a constant pillar of unconditional support and belief. My friends and brothers, Ayush, Akshansh, and Sidhant, thank you for tolerating me and being the voice of sanity when I needed it the most. Finally, thank you to all the friends in Waterloo for being a part of this journey.

Dedication

This is dedicated to my family and my friends. Thank you for believing in me, motivating me, and being a part of this journey.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Participatory democracy is a governance model that revolves around active citizen involvement in decision-making processes [18, 28]. It has gained significant global popularity in recent times and is manifested through various approaches like participatory budgeting, citizens' assemblies, and community forums [29, 58]. By emphasizing inclusivity and accessibility, participatory democracy aims to ensure that citizens feel heard and directly influence routine governance choices. The goal is to foster a healthy debate between the participants and develop a transparent form of governance, reflecting the direct will of the people.

Deliberation and voting are two crucial parts of citizen-focused participatory democratic processes. Deliberation serves as the initial phase, promoting constructive discussions among citizens. Its purpose is to facilitate a healthy exchange of ideas, enabling voters to refine their preferences and enhancing the availability of collective information [23]. The ultimate goal is to foster consensus-building and improve the decision-making process. Although the deliberative process aims to foster consensus formation, it is important to acknowledge that unanimity is not always achievable in practical scenarios. Therefore, voting becomes necessary to make decisions. Voting holds significance as it provides everyone with an equal opportunity to express their opinions through the ballot objectively. It also accommodates the diversity of preferences that may arise after the deliberation phase, which is crucial for inclusive decision-making. Hence, it is established that both deliberation and voting are crucial aspects of participatory democracy procedures, and it is essential to understand the relationship between them.

Having listed the key features of a participatory democratic process, it is important to consider the desired conditions necessary to maximize the benefits of these features.

Inclusivity, representation, and simplicity should be at the core of a healthy citizen-focused democratic process. A simple, inclusive, and fair election mechanism goes a long way in garnering trust and promoting citizen participation [67]. The broader scope of this thesis is to understand the key factors contributing to favorable election outcomes in participatory democracy. A part of this research endeavors to leverage computer science methods to take a step toward establishing best practices for designing effective participatory democracy protocols. We hope our findings can serve to guide the development of several practical democratic institutions, helping them maximize and unlock the true potential of a healthy democratic system. Well-informed democratic platforms have the power to play a pivotal role in fostering alignment through collective decision-making and tackling pressing global challenges like climate change, healthcare, and the economy.

Social choice theory provides a rich technical and mathematical framework to develop a holistic understanding of participatory democracy processes, and elections in general. Dating back to the 18th century, literature in this space has extensively studied, both theoretically and experimentally, a wide variety of collective decision-making procedures. Furthermore, social choice theory offers an extensive classification of various outcomes arising from collective decision-making, offering well-grounded justifications for determining what constitutes a favorable outcome. These justifications encompass multiple dimensions and objectives, including welfare, representation, fairness, and efficiency. To explain this further, consider that each voter obtains some utility from the collective decision, depending on whether the decision matches the voter's preference. Informally, the utilitarian argument for welfare is concerned with selecting an outcome that maximizes the total utility across voters, thereby selecting the most supported candidate(s). On the other hand, the objective of diversity or representation advocates for selecting candidate(s) in a way that maximizes the number of voters who derive some non-zero utility. Finally, proportionality serves as a desirable criterion that balances welfare and diversity, embodying the notion that sufficiently large "cohesive" voter groups deserve to be "fairly" represented [2]. Proportionality is a foundational concept in this thesis, serving as the key metric we aim to fulfil. Within this study, we regard proportionality, defined formally through various axioms, as the fundamental basis for achieving a fair voting outcome. As a result, our research questions are motivated by the objective of enhancing proportionality guarantees within participatory democracy processes, as well as leveraging the fairness attributes offered by proportional voting rules beyond traditional election contexts.

In this thesis, we aim to address research questions broadly at the intersection of participatory democracy, proportionality, and fairness, utilizing techniques in computer science and social choice theory. The first research question revolves around the design of effective participatory democracy mechanisms. It delves into the influence of deliberation on voting processes and aims to explore simple, transparent, and effective designs for deliberation mechanisms that optimize voting outcomes. Specifically, one of the goals is to study whether effective deliberation can circumvent the need for complicated voting rules and vastly improve voting outcomes even for simple, explainable voting rules. However, we recognize the axiomatic guarantees of proportionality and fairness offered by certain intricate proportional voting rules. There is extensive scope to harness the advantageous properties offered by these rules, extending their utility beyond electoral contexts. By preserving their core principles, these voting algorithms can be suitably adapted to address related problems. Thus, our second research question arises from the motivation to explore the application of such sophisticated voting rules, known for their established fairness guarantees, in domains beyond traditional elections. This exploration would enable us to harness the advantageous properties of these rules while accepting a trade-off in explainability when it is deemed acceptable. Particularly, we view the general fair-ranking problem, where the task is to fairly rank a list of items given a query, as a voting problem. Voting is essentially a ranking task, where the objective is to aggregate voter preferences and generate a ranking of candidates. Thus, our focus lies in assessing the suitability and performance of established proportional voting rules as algorithms for achieving fair rankings.

1.1 Contributions

In the context of the research motivation and questions discussed above, this thesis makes the following main contributions to the social choice theory and proportionality literature:

- First, we establish the crucial interaction between deliberation and voting in participatory democracy processes and identify a clear gap in the literature on this topic. To address this, we construct a dynamic agent-based model of deliberation, drawing inspiration from prevalent opinion dynamics models. Our focus is to investigate the impact of different deliberation strategies on voting outcomes, specifically examining approval-based multi-winner elections.

- Through our empirical analysis, we provide evidence that well-designed deliberation strategies, which prioritize exposure to diverse groups and opinions, significantly enhance the quality of deliberation, protect minority preferences, and achieve better voting outcomes. Notably, our research highlights the noteworthy finding that in the presence of effective deliberation, even simple voting rules can be as potent as more complex voting rules that do not incorporate deliberation.

- Finally, we provide a novel, out-of-election-domain application for proportional voting rules: fair ranking. We investigate the connection between proportionality objectives in voting and the statistical parity concepts commonly explored in broader fairness research. Lastly, we develop four original fair-ranking approaches utilizing a well-known proportional voting rule, and we empirically evaluate their performance in various fair-ranking scenarios.

1.2 Outline

The rest of the thesis is organized as follows. Chapter 2 presents a detailed background on social choice from the multi-winner voting and proportionality perspectives. These ideas provide the necessary groundwork required to understand the remainder of the thesis. In Chapter 3, we present our models, experiments, and findings to the first research question on the interaction between deliberation and voting in participatory democracy. Chapter 4 describes the next set of methods and experiments on fair-ranking algorithms using voting rules, addressing the question of fair ranking through proportional voting. Finally, Chapter 5 concludes the thesis, discussing the limitations, future work, and broader impact.

Chapter 2

Background

In this chapter, we provide relevant background on social choice from the perspective of multi-winner voting and proportionality. First, we define the committee election problem and cover several desired properties that motivate the selection of an optimal committee. Next, the concept of proportionality is formally defined within the context of approval-based multi-winner elections, using an axiomatic approach. Finally, we provide a detailed definition of various approval-based committee voting rules. These concepts provide the necessary foundation required to understand the remainder of the thesis.

2.1 Approval-Based Multi-Winner Elections

Social choice theory constitutes the experimental and theoretical framework employed to examine scenarios of collective decision-making. Within these scenarios, the aim is to aggregate individual opinions, preferences, or interests in order to attain a collective decision that optimizes a predetermined social objective. Election scenarios are pervasive within society, spanning diverse domains that include politics, scheduling, expert opinion aggregation, contest outcome determination, and technological applications such as recommendation systems. These scenarios involve the process of selecting preferred options, candidates, or outcomes through various decision-making mechanisms and play a fundamental role in numerous social, organizational, and technological contexts.

Among these scenarios, we are concerned with *multi-winner voting*, where a committee must be selected to represent the interests of some larger group. There are many settings where a committee must be selected. For example, say faculty members in the Computer

Voter	**Approval Ballot**
v_1	$\{a, b, c\}$
v_2	$\{a, e\}$
v_3	$\{d\}$
v_4	$\{b, c, d\}$
v_5	$\{b, c\}$
v_6	$\{b\}$

Table 2.1: Sample approval-based election for the candidate set $C = \{a, b, c, d, e\}$ and $n = 6$ voters. The approval ballots A_i are shown.

Science department are currently engaged in a voting process to choose a committee that will consist of the next set of board members. Furthermore, these elections see many different applications such as facility location [34], participatory budgeting (PB) [18], search result diversification [71], and more. *Multi-winner* voting has been well studied within the social choice literature, with a focus on understanding how the "best" committee can be selected. However, even defining what is meant by "best" is no trivial undertaking. In some contexts, such as aggregation of expert judgments, the desired committee should ideally be of *excellent* quality and consist of the highest-rated k alternatives. However, in other tasks such as choosing k locations for constructing a public facility (*e.g.* hospitals, fire stations), it is preferable to ensure *diversity*, such that as many voters as possible have access to the facility. This tension between the two desired voting outcomes, *excellence* and *diversity* [50], makes the problem challenging and interesting from the perspective of the election designer. Thus, the desired voting objective would dictate the choice of the ideal voting rule. Finally, the format or design used to record voter preferences is also integral to the overall quality of the voting process. We focus on approval-based elections, where voters express preferences by sharing a subset of approved candidates. Approval ballots are used in practice due to their simplicity and flexibility [12, 11, 5]. They also offer scope for deliberation as often voters are left to decide between many different alternatives.

We now provide the formal notation for approval-based committee elections. Let $E = (C, N)$ be an election, where $C = \{c_1, c_2, ..., c_m\}$ and $N = \{1, ..., n\}$ are sets of m candidates and n voters, respectively. Each voter $i \in N$, has an *approval ballot* $A_i \subseteq C$, containing the set of its approved candidates. The *approval profile* $A = \{A_1, A_2, ..., A_n\}$ represents the approval ballots for all voters. For a candidate $c_j \in C$, $N(c_j)$ is the set of voters that approve c_j and its *approval score*, $V(c_j) = |N(c_j)|$. Let $S_k(C)$ denote all k-sized subsets of the candidate set C, representing the set of all possible committees of size k. Given approval profile A and desired committee size $k \in \mathbb{N}$, the objective of a multi-winner election is to select a subset of candidates that form the winning committee $W \in S_k(C)$. An *approval-based committee rule*, $R(A, k)$, is a social choice function that takes as input an approval profile A and committee size k and returns a set of *winning committees.*[1] For any voting rule $R(A, k)$, we will use W_R to denote its selected committee (after tie-breaking). A sample election is shown in Table 2.1.

2.2 Properties

As discussed in the previous section, we ideally want our voting rules to exhibit certain desired properties, representing the principles that should govern the selection of winners given individual ballots. In this thesis, we analyze and compare voting rules across three dimensions: *social welfare*, *representation*, and *proportionality.* Intuitively, the *welfare* objective focuses on selecting candidates that garner maximum support from the voters. *Representation* cares about *diversity*; carefully selecting a committee that maximizes the number of voters represented in the winning committee. A voter is represented if the final committee contains at least one of its approved candidates. The formal definitions for these objectives are given below.

Definition 1 (Utilitarian Social Welfare) *For a given approval profile A and committee size k, the utilitarian social welfare of a committee W is:*

$$SW(A, W) = \sum_{i \in N} \sum_{c \in W} u_i(c), \tag{2.1}$$

$u_i(c) \in \mathbb{R}$ *is the utility voter i derives from candidate c.*

Definition 2 (Representation Score) *For a given approval profile A and committee size k, the representation score of a committee W is defined as:*

$$RP(A, W) = \sum_{i \in N} \min(1, |A_i \cap W|) \tag{2.2}$$

Consider the election from Table 2.1 and a winning committee $W = \{b, c\}$ of size $k = 2$. For this example, assume a simple setting where a voter gets unit utility from an approved candidate being in the selected committee (0 otherwise). Now, from the above definitions, the $SW(A, W)$ is equal to 7 (since b is approved four times and c is approved thrice, the total utility is 7). Similarly, the $RP(A, W)$ in this case is equal to 4 (since v_2 and v_3 don't approve any candidate in W).

[1] A tie-breaking method is used to pick one winning committee in cases where multiple winning committees are returned.

2.3 Proportionality

In most real-world election scenarios, excellence and diversity depict opposite ends of a spectrum and cannot be optimized simultaneously. *Proportionality* serves as an important third objective to capture a compromise between welfare (excellence) and representation (diversity). It requires that if a large enough voter group collectively approves a shared candidate set, then the group must be "fairly represented". Definitions of proportionality differ based on how they interpret "fairly represented".

Definition 3 (T-Cohesive Groups) *Consider an election $E = (C, N)$ with n voters and committee size k. For any integer $T \geq 1$, a group of voters N' is T-cohesive if it contains at least Tn/k voters and collectively approves at least T common candidates, i.e. if $|\cap_{i \in N'} A_i| \geq T$ and $|N'| \geq Tn/k$.*

Definition 4 (Justified Representation (JR)) *A committee W of size k satisfies JR if, for each 1-cohesive group $N' \subseteq N$, there exists at least one voter in N' that approves at least one candidate in W.*

Definition 5 (Proportional Justified Representation (PJR)) *A committee W of size k satisfies PJR if for each integer $T \in \{1, ..., k\}$ and every T-cohesive group $N' \subseteq N$, it holds that $|(\cup_{i \in N'} A_i) \cap W| \geq T$.*

Definition 6 (Extended Justified Representation (EJR)) *A committee W of size k satisfies EJR if for each integer $T \in \{1, ..., k\}$, every T-cohesive group $N' \subseteq N$ contains at least one voter that approves at least T candidates in W, i.e. for some $i \in N'$, $|A_i \cap W| \geq T$.*

A voting rule is considered to satisfy Justified Representation (JR), Proportional Justified Representation (PJR), or Extended Justified Representation (EJR) if it consistently generates a committee that fulfills the respective criterion. EJR is recognized as one of the most robust axioms of proportionality, implying the fulfillment of PJR, which, in turn, implies JR [2, 65]. Unlike EJR [2], where the focus is on a single group member, PJR provides a more natural requirement for group representation. However, EJR provides stronger guarantees for average voter satisfaction [65]. It is worth noting that verifying whether a given committee satisfies EJR or PJR is computationally hard, whereas JR can be verified in polynomial time [3].

2.4 Approval-based Committee Voting Rules

In this section, we define the set of approval-based multi-winner voting rules that form the basis of our analysis and describe some of their key properties. In the next chapter, we will discuss the reasons for selecting to study these rules and their relevance to our analysis.

- **Approval Voting (AV):** For an approval profile A, the AV-score of committee W is $sc_{av}(A, W) = \sum_{c \in W} V(c)$. This rule selects k candidates with the highest individual approval scores. The formal definition is $R_{AV}(A, k) = \arg\max_{W \in S_k(C)} sc_{av}(A, W)$.

- **Approval Chamberlin-Courant (CC):** The CC rule [20], $R_{CC}(A, k)$, picks committees that maximize representation score $RP(A, W)$. Given profile A, $R_{CC}(A, k) = \arg\max_{W \in S_k(C)} RP(A, W)$. It maximizes voter coverage by maximizing the number of voters with at least one approved candidate in the winning committee.

- **Proportional Approval Voting (PAV):** [76] For profile A and committee W, the PAV-score is defined as $sc_{pav}(A, W) = \sum_{i \in N} h(|W \cap A_i|)$, where $h(t) = \sum_{i=1}^{t} 1/i$. The PAV rule is defined as $R_{PAV}(A, k) = \arg\max_{W \in S_k(C)} sc_{pav}(A, W)$. Based on the idea of diminishing returns, a voter's utility from having an approved candidate in the elected committee W decreases according to the harmonic function $h(t)$. It is a variation of the AV rule that ensures proportional representation, as it guarantees EJR [2]. PAV reduces to AV when committee size $k = 1$, but computing the winning committee for PAV is NP-hard [4].

- **Method-of-Equal-Shares (MES):** $R_{MES}(A, k)$, also known in the literature as Rule-X [60, 61], is an iterative process that uses the idea of budgets to guarantee proportionality. Each voter starts with a budget of k/n and each candidate is of unit cost. In round t, a candidate c is added to W if it is q-affordable, *i.e.* for some $q \geq 0, \sum_{i \in N(c)} \min(q, b_i(t)) \geq 1$, where $b_i(t)$ is the budget of voter i in round t. If a candidate is successfully added then the budget of each supporting voter is reduced accordingly. This process continues until either k candidates are added to the committee or it fails. In case of failure, another voting rule is used to select the remaining candidates.

Example 1 *Considering the sample approval election in Table 2.1, we will look at the winning committees of size $k = 2$ chosen by each rule defined above. AV results in the committee $\{b, c\}$, selecting the most approved candidates. CC selects the winning committee $\{a, b\}$, maximizing the voter representation. Finally, PAV and MES also select $\{b, c\}$ (among others).*

Chapter 3

Deliberation and Voting in Approval-Based Multi-Winner Elections

3.1 Introduction

Citizen-focused democratic processes such as citizens' assemblies [29] and participatory budgeting [18] offer extensive scope for discussion over the multitude of possible alternatives. For example, deliberation is an important phase in most implementations of participatory budgeting as it allows voters to refine their preferences and facilitates the exchange of information, with the objective of reaching consensus [5]. While deliberation is a vital component of democratic processes [36, 40], it cannot completely replace voting because, in reality, deliberation does not guarantee unanimity. Even if deliberation encourages agreement, voting is still necessary to aggregate individual preferences post-deliberation. Accordingly, we argue that it is essential to understand the relationship between voting and deliberation. To this end, we use an agent-based deliberation model and study the effect of different deliberation mechanisms on the outcomes derived from well-studied voting rules.

In practice, participatory democratic processes must be simple and explainable to ensure citizen trust and engagement. Lack of transparency discourages participation, especially from under-represented communities. We argue that the "complexity" of a voting rule can be measured along three axes — computational complexity (for some voting rules it is computationally hard to determine the winning committee [4] while for others it is polynomial), the cognitive burden on the voter [11], and the ease of explaining the voting rule. The first dimension, computational complexity, is well-defined and extensively explored in social choice theory research. In contrast, the cognitive burden on the voter is subjective and difficult to gauge, as it pertains to the mental effort required by the voter to collect and articulate their preferences. The aim is to reduce this burden by facilitating a straightforward ballot design, which can be achieved through the implementation of approval ballots. Lastly, the third dimension relates to the simplicity of explaining the voting algorithm, specifically the method used to determine the winners. Complicated rules may provide strong performance guarantees, but they are often hard to explain to the layperson. In this work, we argue that effective deliberation can circumvent the need for complicated voting rules and vastly improve voting outcomes even for simple rules such as classical approval voting (AV).

We present an agent-based model of deliberation and explore various alternatives for structuring deliberation groups. We evaluate standard multi-winner voting rules, both before and after voters have the opportunity to deliberate, with respect to standard objectives from the literature, including social welfare, representation, and proportionality. We show that deliberation, in almost all scenarios, significantly improves welfare, representation, and proportionality. However, the results are sensitive to the deliberation mechanism; increased exposure to diverse opinions (or agents from different backgrounds) enhances the quality of deliberation, achieves higher consensus, protects minority preferences, and in turn achieves better voting outcomes. Finally, our results indicate that in the presence of effective deliberation, *simple*, explainable voting rules such as approval voting perform as well as more sophisticated, *complex* rules. This can serve to guide the design and deployment of voting rules in citizen-focused democratic processes.

3.2 Background

In this section, we provide a detailed background. We cover the existing research in this area and then define the objectives utilized for comparing our methods. We then list the approval-based committee voting rules used for our analysis and present our motivation for choosing them. Finally, we provide a brief background of opinion dynamics models to describe the choice and design of our deliberation model.

3.2.1 Related Work

The social choice literature has extensively studied the quality of approval-based multi-winner voting rules. From the quantitative perspective, a recent paper by Lackner and

Skowron provides an in-depth theoretical and empirical analysis of different approval-based multi-winner voting rules with respect to (utilitarian) social welfare and representation guarantees [50]. Fairstein et al. extended this work to study the welfare-representation trade-off in the more general PB setting [32]. The traditional axiomatic approach, on the other hand, provides a qualitative evaluation, *i.e.* whether a voting rule satisfies a property or not. For approval-based rules, recent work has focused heavily on proportionality axioms [2, 65, 3, 14, 49, 69]. We refer the reader to an extensive survey on the properties of multi-winner rules by Faliszewski et al. [33].

Deliberation, specifically within social choice, has been studied through various approaches. From the theoretical perspective, a wide variety of mathematical deliberation models have been proposed [21, 87]. For example, recent work has looked at iterative small-group deliberation methods for reaching consensus in collective decision-making problems [38, 30]. Elkind et al. propose a consensus-reaching deliberation protocol based on coalition formation [27]. A recent experimental study shows that deliberation leads to meta-agreements and single-peaked preferences under specific conditions [63]. Another paper looks at deliberation and voting simultaneously, but their work is limited to the *ground-truth* setup with ordinal preferences over three alternatives [59]. They do not study the impact of deliberation on the quantitative and qualitative properties of voting rules.

In this chapter, we bridge the gap between deliberation and voting literature. To our knowledge, we are the first to experimentally study the effect of deliberation on voting outcomes across different deliberation strategies.

3.2.2 Objectives

Our analysis is based on different standard objectives [50]. In particular, we consider objectives across three dimensions: welfare, representation, and proportionality.

Utilitarian Ratio: This ratio compares the (utilitarian) social welfare achieved by $W_R = R(A, k)$ to the maximum social welfare achievable:

$$UR(R) = \frac{SW(A, W_R)}{max_{W \in S_k(C)} SW(A, W)} \tag{3.1}$$

Representation Ratio: This ratio measures the diversity of the committee $W_R = R(A, k)$, by comparing the representation score achieved by W_R to the optimal representation score amongst all k-sized committees:

$$RR(R) = \frac{RP(A, W_R)}{max_{W \in S_k(C)} RP(A, W)}. \tag{3.2}$$

Note that the CC rule maximizes representation, $RR(R_{CC}) = 1$.

Utility-Representation Aggregate Score: This score captures how well a voting rule, $R(A, k)$ balances both social welfare and representation:

$$URagg(R) = UR(R) * RR(R) \tag{3.3}$$

Voter Satisfaction: Given $W_R = R(A, k)$, the voter satisfaction is measured as the average number of candidates approved by a voter in W:

$$VS(R) = \frac{\sum_{i \in N} |A_i \cap W_R|}{|N|} \tag{3.4}$$

Finally, we are interested in experimentally verifying whether or not the generated profile instances **satisfy EJR**, **PJR**, or **JR**. To this end, we count the number of profile instances that satisfy these three properties.

3.2.3 Voting Rules

We study the approval-based **multi-winner voting rules** formally defined in 2.4: Classical Approval Voting (AV), Approval Chamberlin-Courant (CC) [20], Proportional Approval Voting (PAV) [76], and Method-of-Equal-Shares (MES) [61]. We elect to study these rules since they exhibit a wide range of properties, allowing for comparisons to be drawn across several axes. The following example illustrates the difference between our chosen voting rules and the properties used to compare them.

Example 2 *Consider an election profile with $n = 55$ voters and $m = 9$ candidates, such that 30 voters approve candidates $\{c_1, c_2, c_3\}$, 20 voters approve $\{c_4, c_5, c_6\}$, and 5 voters approve $\{c_7, c_8, c_9\}$. Let $k = 3$. Given this profile, the AV-winning committee is $W_{AV} = \{c_1, c_2, c_3\}$, the CC-winning committee is $W_{CC} = \{c_2, c_5, c_8\}$ (among others), the PAV-winning committee is $W_{PAV} = \{c_1, c_2, c_4\}$ (among others), and finally MES also results in*

$W_{MES} = \{c_1, c_2, c_4\}$. *These committees vary significantly and cover different properties. As per the definitions in section 2.2,* W_{AV} *achieves the highest utilitarian social welfare* $SW(A, W_{AV}) = 90$ *(assuming unit utilities) but a poor representation score* $RP(A, W_{AV}) = 30$ *and fails EJR. Contrarily,* W_{CC} *achieves a poor welfare score of* 55 *but the optimal representation score of* 55 *as it covers all voters. Both PAV and MES select a proportional committee that satisfies EJR and reflects a compromise between the two extremes.* W_{PAV} *or* W_{MES} *achieves a welfare score of* 80 *and a representation score of* 50.

In the above example, we highlight the wide range of properties covered by these rules. Firstly, AV is known to maximize social welfare under certain conditions on voters' utility functions [50, 49], however, there are no guarantees that AV satisfies proportionality as defined by the EJR criterion [2]. Conversely, CC maximizes diverse representation, but its welfare properties are less well understood. Both PAV and MES guarantee EJR and maintain a balance between diverse representation and social welfare. Finally, we argue that AV can be viewed as being *simple* in terms of computational complexity and explainability, whereas, PAV and MES are *complex* along at least one of these axes. Thus, this collection of multi-winner voting rules covers the set of properties we are interested in better understanding.

3.2.4 Opinion Dynamics Models

Opinion dynamics is the study of opinion or belief diffusion and spread in a population. These models aim to formally and mathematically capture the complex interactions and dynamics that shape the formation, diffusion, and evolution of opinions among individuals or groups. Furthermore, these models typically consider individuals as agents with certain attributes and simulate the interaction between several agents over time, allowing for opinion evolution and observing the factors leading to that change.

In our work, we model deliberation between voters using agent-based opinion dynamics models and discuss two well-established models below.

DeGroot's Classical Model

According to DeGroot's model [22], an agent's updated opinion is simply the weighted sum of opinions from various sources (itself included). The weights were static, and could be different for different agents. So, for two agents x and y, x updates its opinion as:

$$x(t+1) = w_{xx}x(t) + w_{xy}y(t) \tag{3.5}$$

where $x(t)$ denotes the opinion of agent x at time t, w_{xx} and w_{xy} denote x's weights on its own opinion and y's opinion, respectively. Note that the weights should sum up to 1, and therefore, $w_{xy} = 1 - w_{xx}$.

Bounded Confidence Model

Hegselman and Krause later presented the Bounded Confidence (BC) model [44], which introduced a global confidence level Δ. In the original paper, agents were on a network, and agents updated their opinions based on opinions of their neighbors. In the BC model, an agent x considered a neighbor's (y) opinion only if the neighbor's opinion was within x's confidence interval $[x(t)-\Delta, x(t)+\Delta]$. In the initial version, there were no distinct weights and all opinions within the confidence interval were weighted equally. When simplified for just two agents x and y, the opinion update for x is given by:

$$x(t+1) = \begin{cases} 1/2(x(t) + y(t), & \text{if } y(t) \in [x(t) - \Delta, x(t) + \Delta] \\ x(t), & \text{otherwise} \end{cases} \tag{3.6}$$

The BC model captures the idea of confirmation bias, and BC and its several modified versions have largely remained popular to date in the field of opinion dynamics.

3.3 Model

In this section, we describe our deliberation and voting model in detail, within the framework defined in the previous sections. We first define our underlying agent population and how we model their initial preferences. We then discuss the deliberation process, through which agents exchange information and update their preferences. Finally, we observe that deliberation is often done, not at the full population level, but instead in smaller subgroups. We discuss different ways these deliberation subgroups can be created.

3.3.1 Voting Population: Preferences and Utilities

Our agent population N is divided into two sets — a *majority* and *minority*, where the number of agents in the majority is greater than that in the minority. Agents' initial preferences depend on their population group. Consistent with previous work [50], our preference model is based on the ordinal Mallows model. The rankings are then converted to an approval ballot using the top-ranked candidates. In particular, we assume an agent i's initial preference ranking, P_i^0, is sampled from a Mallows model [55], with reference rankings, Π_{maj} and Π_{min}, for the majority and minority populations respectively.[1]

We further assume that agents have underlying cardinal utilities for candidates, consistent with their ordinal preferences. For agent i, these utilities are represented by a vector $U_i = \langle u_i(c_1), u_i(c_2), \ldots, u_i(c_m)\rangle$, where $u_i(c_x) \geq u_i(c_y)$ if and only if $c_x \succeq_i c_y$ in P_i^0, and $u_i(c_x) \in [0, 1]$. We work in this cardinal space as it allows us to leverage standard deliberation models and measure welfare across voting rules in settings where voters derive some utility from elected candidates who were not on their ballot. Our particular instantiation of utility functions subsumes earlier work (e.g. [50]) and is consistent with utility models used in the social choice literature (e.g. [62, 31]).

3.3.2 The Deliberation Process

Deliberation is defined as a "discussion in which individuals are amenable to scrutinizing and changing their preferences in the light of persuasion (but not manipulation, deception or coercion) from other participants" [24]. Deliberation thus requires a group of peers with whom to deliberate and a methodology for changing preferences. In this section, we describe the process in which agents update their preferences, deferring details about peer groups until later.[2]

Deliberation is an iterative process, involving, at each step, a speaker and listeners. The speaker makes a report, based on their preferences, and the listeners update their own preferences based on this information. In this work, we use a variation of the Bounded Confidence (BC) model to capture the (abstract) deliberation process [44]. The BC model is a particularly good match for modelling deliberation in groups because it was intended to "describe formal meetings, where there is an effective interaction involving many people at the same time" [19]. In the BC model, listeners consider the speaker's report (e.g. utilities for different candidates) and update their opinions/preferences of the candidates independently, only if the speaker's report is not "too far" from their own. The notion of distance is captured by a confidence parameter for each listener, Δ_i, where agents may have different confidence levels [53, 78]. The BC model was designed for one-dimensional

[1] The Mallows model is a standard noise model for preferences. It defines a probability distribution over rankings over alternatives (i.e. preferences), defined as $\mathbb{P}(r) = \frac{1}{Z}\phi^{d(r,\Pi)}$ where Π is a reference ranking, $d(r, \Pi)$ is the Kendall-tau distance between r and Π, and Z is a normalizing factor.

[2] As is common in much of the deliberation literature (e.g [24, 59]), we assume agents are non-strategic and truthfully reveal their utilities.

opinion spaces. However, agents in our model discuss and update utilities derived from all m candidates in C, making it a multi-dimensional space. We make a simplifying assumption that agents' utilities for all m candidates are independent of each other and apply the BC model to each dimension (candidate) independently.[3]

Given time step t, some agent, x, selected as the speaker, makes its report (which reveals x's thoughts and utilities for the candidates). Each listener updates its own preferences across candidates $c_j \in C$ according to the following rule

$$u_i^{t+1}(c_j) = \begin{cases} (1 - w_{ix})u_i^t(c_j) + w_{ix}u_x^t(c_j), & \text{if } |u_i^t(c_j) - u_x^t(c_j)| \leq \Delta_i \\ u_i^t(c_j), & \text{otherwise} \end{cases} \tag{3.7}$$

where $w_{ix} \in [0, 1]$ is the *influence weight* that i places on x's perspective. It is known that opinions from sources similar to oneself have a higher influence than opinions from dissimilar sources [80, 54]. To capture this phenomenon, we let w_{ix} take on one of two values, contingent on the relationship between i and x. In particular,

$$w_{ix} = \begin{cases} \alpha_i, & \text{if } \{i, x\} \subset N_{maj} \vee \{i, x\} \subset N_{min} \\ \beta_i, & \text{otherwise.} \end{cases} \tag{3.8}$$

In words, if i and x are both members of the majority group (N_{maj}) or the minority group (N_{min}) then $w_{ix} = \alpha_i$, otherwise $w_{ix} = \beta_i$ where $\alpha_i \geq \beta_i$.

3.3.3 Deliberation Groups

In the real world, deliberation typically happens in small discussion or peer groups [29, 39]. To this end, we divide the agent population into g sub-groups of approximately equal size. The deliberation process is conducted within these sub-groups where one *round* of deliberation is complete when all agents in each group have had the opportunity to speak.

We want to explore how group-formation strategies influence the deliberation process and the final decision made through voting. Our strategies are informed by common heuristics or rationale used in practice and none rely on private/unknown information such as the agents' underlying utilities or preferences. We do, however, assume that whether an agent is a member of the majority or minority group is public information and allow

[3] This assumption might be restrictive in the participatory budgeting setup where voter utilities are dependent on project costs and a total budget. However, in our case, i.e. general committee elections, this assumption is not too restrictive as the voters could view every candidate independently.

group-formation strategies to use such information. Finally, we consider both single-round and iterative group-formation strategies where agents are divided into different groups in each round [29].

3.3.4 Single-Round Group-Formation Strategies

Homogeneous group: Each group contains only agents who are members of N_{maj} or N_{min}. That is, there is no mixing of minority and majority agents. If groups formed organically without a central planner, such structures are most likely to form. The concept of homophily, the tendency for people to connect and socialize with those sharing similar characteristics, beliefs, and values, dates as far back as Plato, who wrote in Phaedrus that "similarity begets friendship", and there is evidence that adults, in particular, preferentially associate with those of similar political persuasions [56].

Heterogeneous group: Each group is selected such that the ratio of the number of majority agents to the number of minority agents within the group is approximately equal to the majority:minority ratio in the overall population. Each group created through this strategy is *diverse* and representative of the overall agent population. This strategy is already popular among practitioners in the real world. Citizens' Assembly of Scotland diversifies deliberation groups based on age, gender, and political affiliation [39].

Random group: Each group is created by randomly sampling agents from the population (without replacement) with equal probability.

Large Group: This is a special case where the deliberation process runs over the entire population of agents. Considering time constraints, limited attention spans, and other physical limitations, such a strategy is not typically used in practice. However, we include this strategy as a Utopian baseline because it ensures maximum exposure to the preferences of every other agent in the system.

3.3.5 Iterative Group-Formation Strategies

Iterative random: This strategy assigns agents to groups at random, but these assignments are done in each iteration or round. It is an iterative version of the random group division discussed above.

Iterative golfer: This strategy is a variant of the social golfer problem [43, 52] from combinatorial optimization. The number of rounds, R, is fixed *a priori*, and the number of times any pair of agents meet more than once is minimized. We refer the reader to Appendix A for details. A similar approach is used in Sortition Foundation's GroupSelect algorithm [77], which is used by several nonprofits for group-formation in participatory budgeting sessions [39].

3.4 Experimental Setup

We now describe our experimental setup. We first describe our population of agents. We then describe the process in which agents deliberate, before discussing details about the voting processes.

Our election setup consists of 50 candidates ($|C| = 50$)[4] and 100 voters, with 80 agents in the majority group (N_{maj}) and 20 in the minority group (N_{min}). Agents' initial preferences are sampled using a Mallows model, with $\phi = 0.2$. The reference ranking used while sampling a preference ordering depends on whether the agent belongs to N_{maj} or N_{min}. Reference rankings, Π_{maj} and Π_{min}, are sampled uniformly from all linear orders over C. Due to this sampling process, agents in either the majority or minority group have fairly similar preferences (as ϕ is relatively small) but the two groups themselves are distinct. To instantiate agents' utility functions, we generate m samples independently from the uniform distribution $\mathbf{U}(0, 1)$, sort it, and then map the utilities to the candidates according to the agent's preference ranking. For the BC model, all three parameters $(\Delta_i, \alpha_i, \beta_i)$ are sampled from uniform distributions over the full range for each parameter.[5]

When deliberating, agents are divided into 10 groups (except for the *large group* strategy). This is similar to the Citizens' Assembly of Scotland, which ran over 16 sessions; in each session, the 104 participants were divided across 12 tables [39]. For iterative deliberation, the deliberation continues for $R = 5$ rounds. We consider different approval-based multi-winner voting rules to elect $k = 5$ winners. We do not use a fixed ballot size in this work to allow agents more flexibility. Accordingly, we use a flexible ballot size, such that each agent's ballot is of size b_i, where b_i is sampled from $\mathcal{N}(2k, 1.0)$. Agent i's approval vote is then the set consisting of its b_i top-ranked candidates from its preference ranking.

[4] Since project proposals are typically invited from the participants in PB [18, 5] there are a large number of candidates to choose from (e.g., PB instances in Warsaw, Poland had between 20-100 projects (36 on average).[75, 32]).

[5] We ran experiments where all parameters were drawn from a normal distribution. There were no significant differences from the results reported here.

As a baseline, we apply every voting rule to the agent preferences *before* deliberation. We then run the different deliberation strategies, freezing agents' utilities once deliberation has concluded. We then apply every voting rule to the updated preferences. We use the Python library *abcvoting* [48], and use random tie-breaking when a voting rule returns multiple winning committees.[6]

To avoid trivial profiles, *i.e.*, profiles where an almost perfect compromise between welfare and representation is easily achievable, we impose some eligibility conditions. An initial approval profile A^0 is eligible only if $\mathrm{RR}(AV, A^0) < 0.9 \land \mathrm{UR}(CC, A^0) < 0.9$. This is a common technique used in simulations comparing voting rules based on synthetic datasets [50]. This entire simulation is repeated $10,000$ times and the average values are reported. To determine statistical significance while comparing any two sets of results, we used both the t-test and Wilcoxon signed-rank test, and we found the p-values to be roughly similar. All pairs of comparisons between deliberation group strategies for a given voting rule are statistically significant ($p < 0.05$) unless otherwise noted.

3.5 Results

We have several goals for our experiments. First, we use the metrics introduced in Section 3.2.2 to compare voting rules where there is no deliberation. This allows us to establish a *baseline* to compare against. We then explore the impact that deliberation has on the outcome achieved by the different voting rules, including comparing different deliberation mechanisms so as to best understand how the structure of the deliberating groups affects the final outcome, including comparing against an idealized situation where all agents share information and deliberate together in one large group. In the rest of this section, we describe our findings, which we organize according to the different metrics.

3.5.1 Impact of Deliberation on Preferences

Variance

To understand how deliberation processes shape and change agents' preferences, we compare the average variance in the agents' utilities before (*initial*) and after deliberation (Figure 3.1). As expected, deliberation reduces disagreement amongst agents, moving all towards a consensus. Processes where agents are exposed to more, diverse, agents (e.g.

[6] The code is available at: `https://github.com/kanav-mehra/deliberation-voting`.

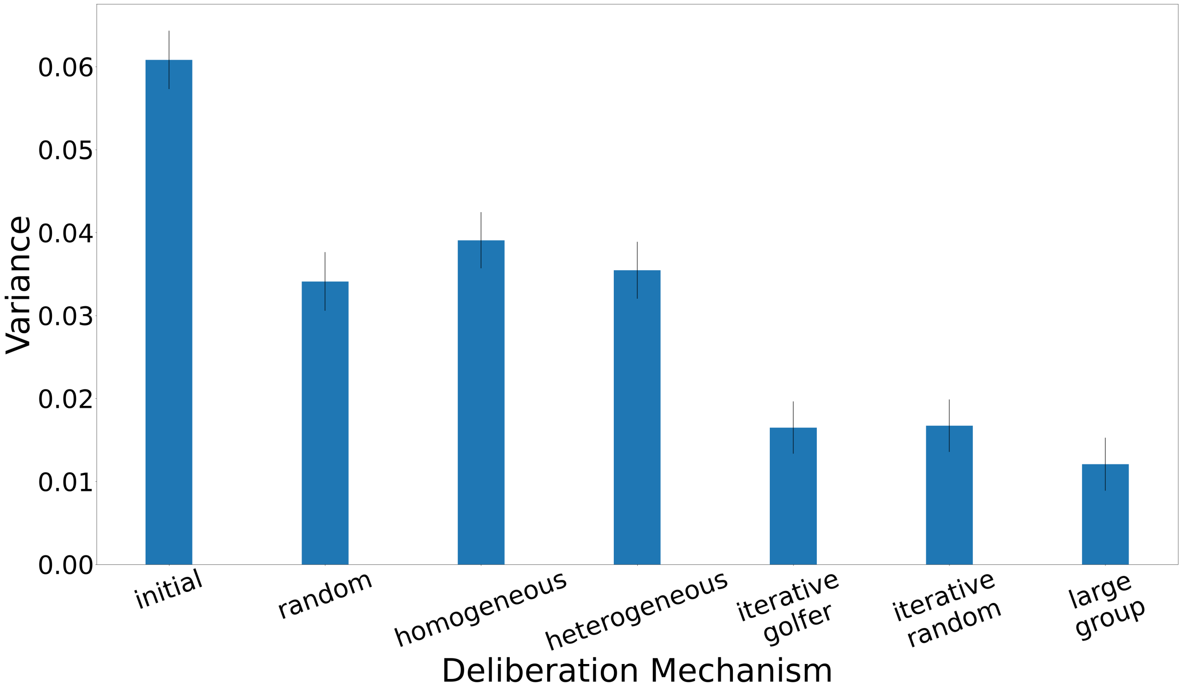

Figure 3.1: Average variance of agents' utilities for candidates. Lower variance implies a higher degree of consensus in the population.

the iterative variants and the *large group*) see the largest reduction in variance across the population. While achieving greater agreement is desirable, it should not be achieved by disregarding initial minority opinions. We delve into this topic in Section 3.6.

Inter-group Ballot Disagreement

In Figure 3.1 we introduce a measure of consensus in the population as the average variance in agents' utilities and show that deliberation reduces disagreement amongst agents. To complement this analysis and further understand the impact of deliberation on agents' preferences, we introduce another metric that computes the disagreement between the majority and minority voters based on their ballots. More generally, given two approval ballots A_a and A_b the disagreement score between the ballots is computed as:

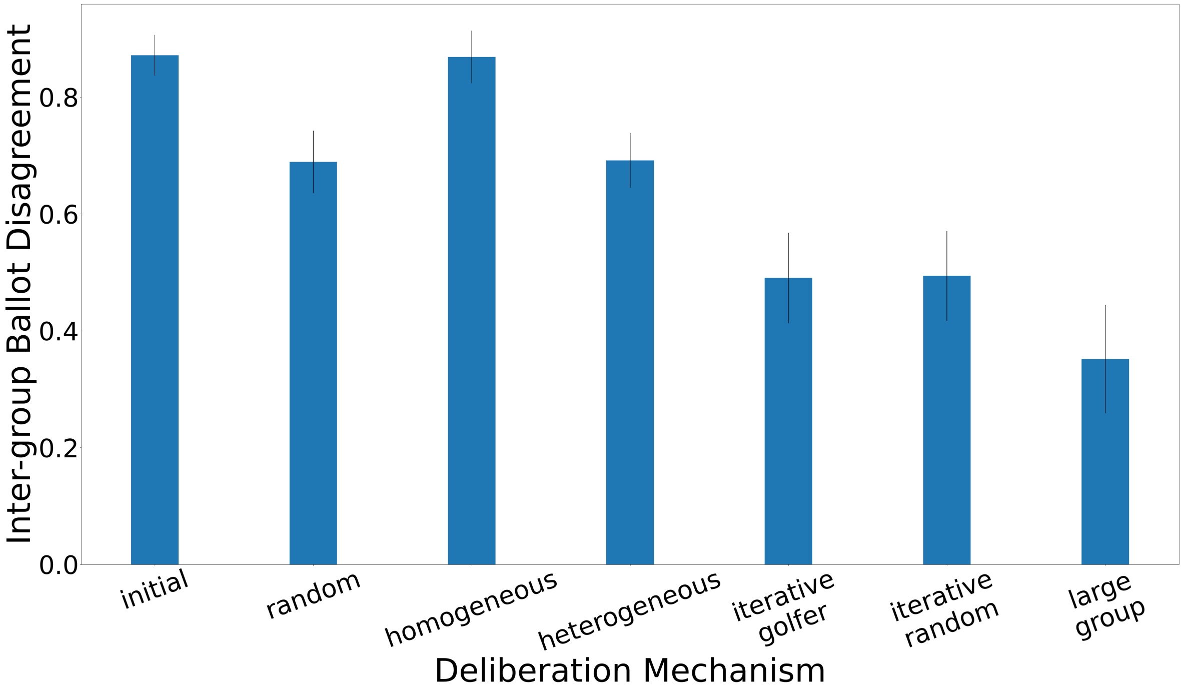

Figure 3.2: Inter-group Ballot Disagreement

$$\text{Ballot Disagreement Score} = 1 - \frac{|A_a \cap A_b|}{\min(|A_a|, |A_b|)} \tag{3.9}$$

A maximum disagreement score of 1 means the approval ballots are disjoint, *i.e.* the voters do not approve any candidates in common. This score is computed for every majority-minority voter pair in the population across all deliberation mechanisms and the average results are reported in Figure 3.2.

We observe a similar trend here as well (as seen in Figure 3.1). Deliberation significantly reduces disagreement between the two population groups and moves the overall population toward consensus. This positive effect is stronger in deliberation methods that increase exposure to more, diverse agents (*i.e.* the iterative versions and *large group*).

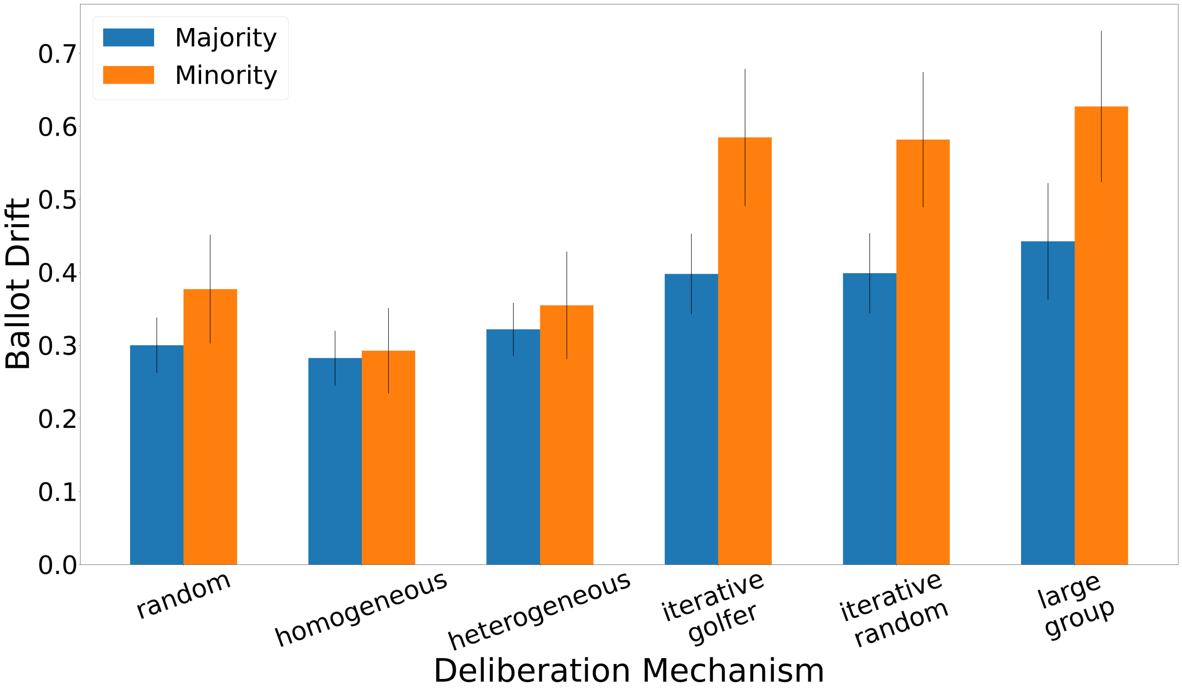

Figure 3.3: Ballot Drift

Ballot Drift

Trends observed from the preceding two measurements indicate that deliberation facilitates a shift towards consensus. Nevertheless, while attaining a higher level of agreement is preferable, it must not be pursued at the expense of disregarding minority opinions. Stated differently, consensus should not be solely attained by influencing the minority to conform to the majority opinions. In this section, we measure average ballot drifts compared to the pre-deliberation ballots for both population groups across all deliberation mechanisms.

In particular, for every voter, we compute the ballot disagreement score (defined in Equation 3.9) between their respective post- and pre-deliberation ballot for all deliberation mechanisms. This is computed for every voter and the average results for both population groups, minority and majority, are reported in Figure 3.3.

The results are positive and show that both groups undergo significant movement. While deliberation encourages agreement, the burden of consensus does not just fall on the minority population. However, it is worth noting that the minority ballot drift is higher

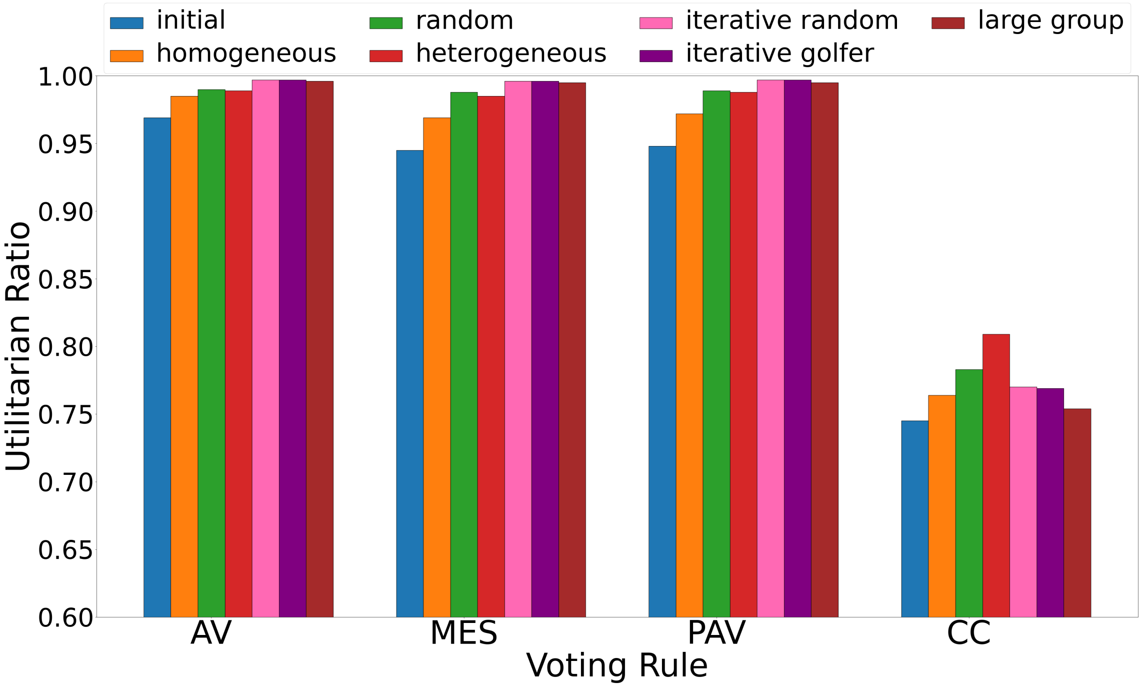

Figure 3.4: Utilitarian ratio across deliberation mechanisms.

than the majority drift in magnitude. This is expected since there are a greater number of majority agents in the population. Finally, we observe again that ballot drift is higher and stronger for deliberation methods encouraging exposure to more, diverse agents as seen in the previous measures.

3.5.2 Utilitarian Ratio

Figure 3.4 reports the impact of deliberation on utilitarian social welfare. First, we compare the voting rules where there is no deliberation (see *initial* case denoted by the blue bars). AV achieves the highest utilitarian ratio, *i.e.* the utilitarian social welfare provided by AV is closest to the optimal social welfare. Both proportional rules (MES and PAV) are similar and obtain utilitarian ratios that are only slightly lower than AV. Finally, CC performs the worst in terms of welfare. In general, our results match the trends reported in previous theoretical and experimental results [50].

We now address our main point of interest – the effect of deliberation. As seen in Figure

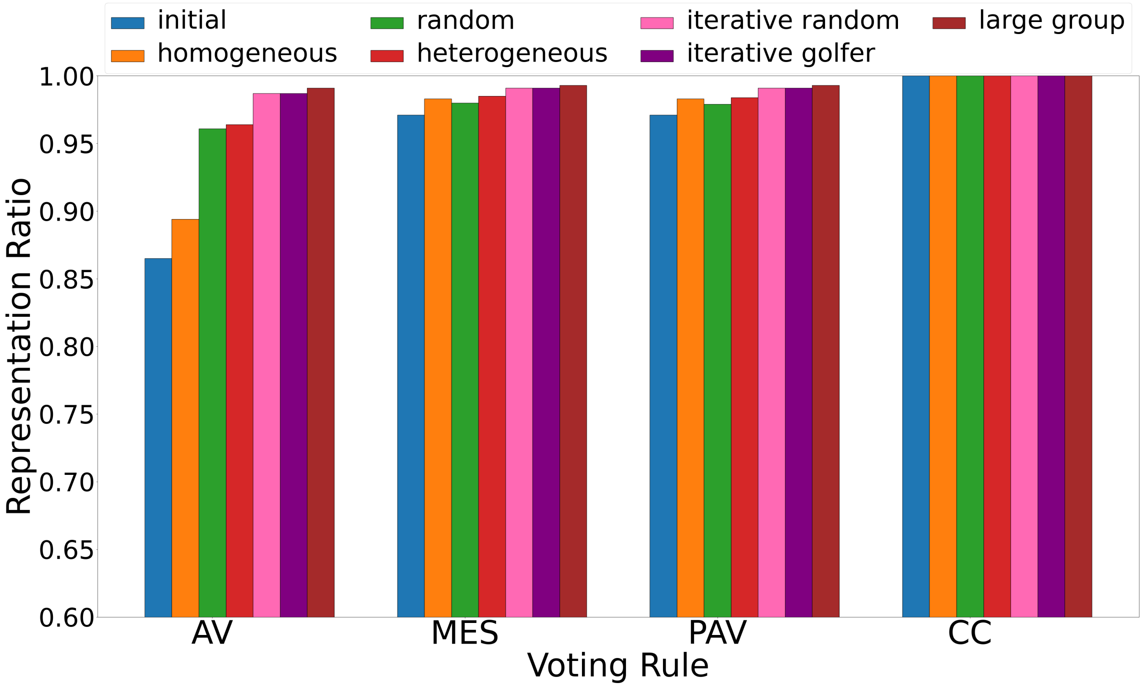

Figure 3.5: Representation ratio across deliberation mechanisms.

3.4, deliberation improves social welfare over the *initial* baseline (blue). In single-round deliberation, both *random* (green) and *heterogeneous* (red) methods show similar results and outperform *homogeneous* (orange) for AV, MES, and PAV. Iterative deliberation exhibits further improvement for these rules. It is worth noting that *iterative golfer* (purple) and *iterative random* (pink) perform similarly and match the *large group* benchmark (brown). For CC, social welfare is always improved with deliberation, however, iterative deliberation is not as powerful. This is due to the nature of the rule. We cover this in further detail later in the thesis.

3.5.3 Representation Ratio

Figure 3.5 shows the average representation ratio of the voting rules across different deliberation mechanisms. Since CC optimizes for diversity by design, $RR(CC) = 1.0$. We focus our analysis on the other rules so as to better understand if it is possible to achieve comparable representation when adding deliberation. Under no deliberation, AV has the

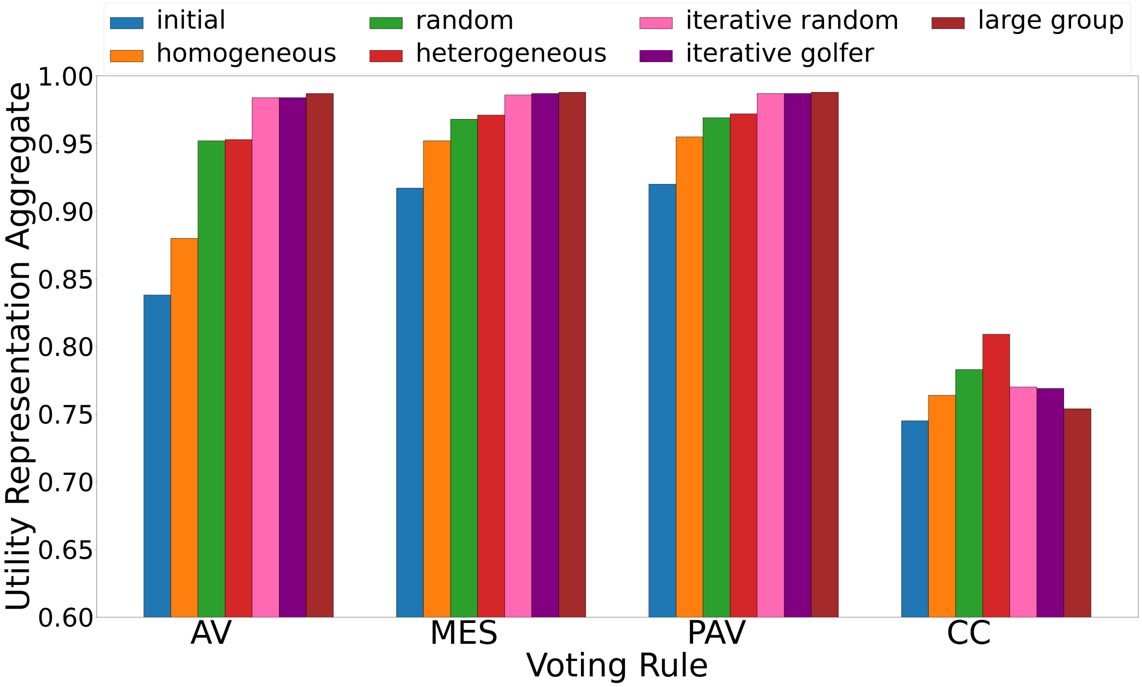

Figure 3.6: Utility-Representation aggregate score across deliberation mechanisms.

lowest representation ratio. Since AV simply picks the candidates with the highest approval scores, it does not care about minority preferences. The proportional rules (MES and PAV), however, achieve much higher representation as they are designed to maintain a balance between welfare and diversity.

The effect of deliberation is more pronounced here compared to the utilitarian ratio results, particularly for AV. Within the single-round mechanisms, *homogeneous* achieves a slight improvement over the *initial* setup for all rules. However, specifically for AV, both *heterogeneous* and *random* achieve much higher representation over both *initial* and *homogeneous* setups. Again, both iterative mechanisms achieve further improvements compared to the single-round setups and almost match the *large group* benchmark.

3.5.4 Utility-Representation Aggregate Score

This score captures how well a voting rule balances welfare and representation. Figure 3.6 shows the average results for this objective. Under no deliberation (*initial* baseline), we

Deliberation Strategy	**EJR%**		**PJR%**		**JR%**	
	AV	**CC**	**AV**	**CC**	**AV**	**CC**
Initial (no deliberation)	99.5	62.5	99.5	73.4	99.5	100
Homogeneous	96.4	69.9	96.4	75.1	96.4	100
Random	100	81.9	100	85.6	100	100
Heterogeneous	100	92.7	100	94.0	100	100
Iterative Random	100	31.4	100	53.6	100	100
Iterative Golfer	100	29.9	100	51.2	100	100
Large Group	100	6.10	100	23.4	100	100

Table 3.1: Proportionality Satisfaction (AV and CC).

see that the proportional rules (MES and PAV) perform the best, followed by AV, and then CC. This is consistent with earlier findings since the proportional rules are designed with this goal in mind. Both AV and CC perform poorly on this metric, pre-deliberation, since they do well on either welfare (AV) or representation (CC) but not both.

We observe a positive effect from deliberation and achieve a significant performance improvement over the *initial* baseline. Within the single-round mechanisms, *heterogeneous* and *random* perform similarly (except for CC where *heterogeneous* is better) and outperform the *homogeneous* setup. Iterative deliberation leads to further improvement as both iterative methods match the performance of the *large group*.

3.5.5 Proportionality Satisfaction

Table 3.1 covers proportionality satisfaction and shows the percentage of EJR-, PJR-, and JR-satisfying committees (out of 10,000 simulations) returned by AV and CC. We focus only on AV and CC since the proportional rules MES and PAV guarantee EJR. Even under no deliberation (*initial*), AV satisfies EJR in almost all profiles, which further improves to perfect satisfaction with deliberation (except *homogeneous*). This is interesting since AV is not guaranteed to satisfy EJR.[7] Proportionality satisfaction for CC also improves if single-round deliberation is supported, with *heterogeneous* achieving the best result. Iterative deliberation, however, does not perform well. We believe this arises due to CC's strong focus on representation and discuss this in further detail later.

[7] Since the minority and majority agents have highly correlated approval sets, T-cohesive groups may exist only for a small set of minority- and majority-supported candidates, thereby making the EJR requirement easy to satisfy. Furthermore, previous research [32, 13] shows that under many natural preference distributions (generated elections), there are many EJR-satisfying committees.

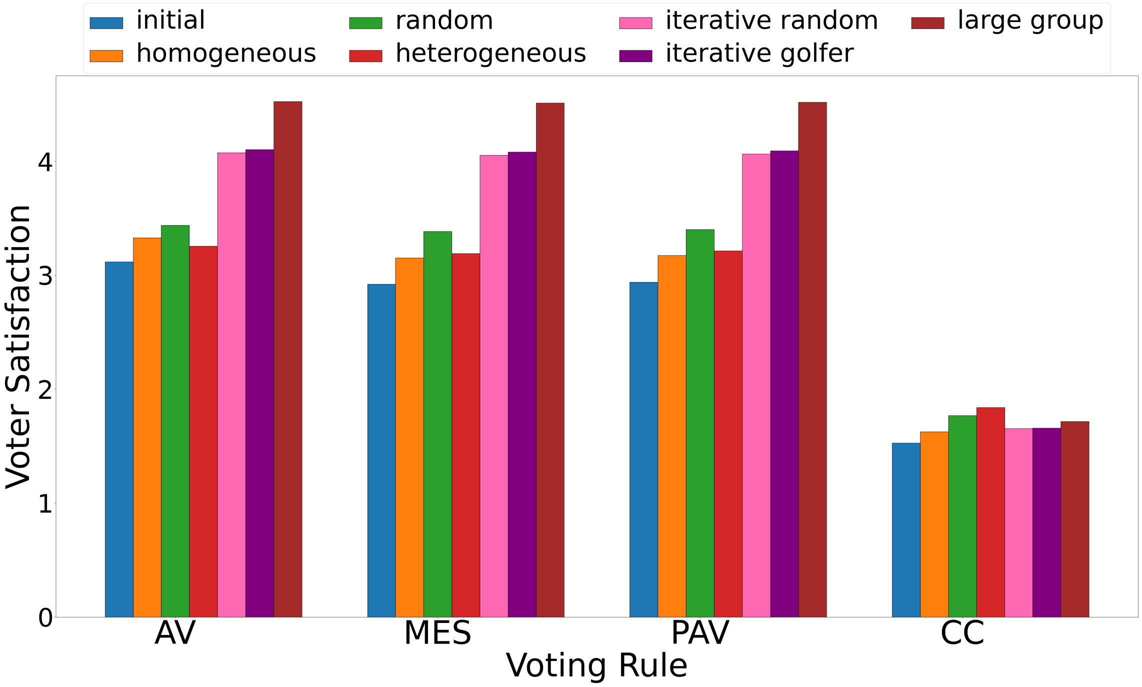

Figure 3.7: Voter satisfaction across deliberation mechanisms

3.5.6 Voter Satisfaction

Figure 3.7 shows the average voter satisfaction obtained by the voting rules.

AV is expected to achieve the highest satisfaction since it picks candidates with the highest support, i.e. the average number of candidates approved by a voter will be high. MES and PAV achieve comparable scores, just slightly lower than AV. Finally, CC achieves the lowest satisfaction of all rules. In an attempt to maximize voter coverage, CC might choose winning candidates that represent few voters, and as a result, have low approval scores. Due to this, it maximizes diversity but achieves low voter satisfaction.

Compared to the *initial* baseline, we observe an improvement in satisfaction scores under all deliberation mechanisms. In general, all single-round deliberation setups achieve comparable performance, with the exception of *random* performing the best in some cases. Moving on to the iterative methods, we notice a further increase in satisfaction scores for all rules except CC. While both iterative setups perform similarly and improve over the *initial* baseline, they are still outperformed by the *large group* benchmark.

3.6 Discussion

As we observed in the previous section, deliberation changes the quality of the outcomes produced by different multi-winner voting rules. In this section, we explore these observations in more detail.

3.6.1 Single-Round Deliberation

Even a single round of deliberation improved outcomes across all voting rules and all objectives. However, the choice of the deliberation structure was also important.

For all objectives, *random* and *heterogeneous* consistently outperformed *homogeneous*. We hypothesize that this improvement was due to these deliberation strategies maximizing exposure to diverse opinions. Under *homogeneous* deliberation, the population sub-groups become more inwardly focused, leading to the formation of distinct T-cohesive groups. This was particularly problematic when used with AV, which picks candidates with the highest approval support and fails to 'fairly' represent the cohesive minority agents in some cases, thereby failing EJR (Table 3.1). By allowing majority and minority agents to interact, there was an opportunity for minority agents to influence the majority population. This translated to higher welfare, representation, and proportionality guarantees (Figure 3.6 and Table 3.1).

3.6.2 Iterative Deliberation

In comparison to single-round methods, iterative deliberation further supports consensus (Figure 3.1) and improves all objectives for most voting rules. Furthermore, there was no statistical difference between the *iterative golfer* and *iterative random* methods. We view this as a positive result with practical design implications. While care does need to be taken in determining group sizes, a simple, computationally inexpensive mechanism is as effective as one that is more complex.

The exception to the observation is the CC rule. CC's strong focus on representation and coverage makes it unsuitable for deliberation methods that drive higher degrees of consensus (such as iterative methods and *large group*) since it fails to represent population groups proportionally.

Deliberation Strategy	**Minority Opinion Preservation**
Initial (no deliberation)	0
Homogeneous	0.20
Random	0.30
Heterogeneous	0.48
Iterative Random	0.65
Iterative Golfer	0.66
Large Group	0.92

Table 3.2: Minority Opinion Preservation: average number of initial (pre-deliberation) *minority-supported* candidates selected by AV in the final committee after deliberation.

3.6.3 Minority Opinion Preservation

While we have been extolling deliberation, there are caveats. In particular, it is important to ensure that deliberation processes are inclusive and encourage minority participation [37]. Care must be taken to ensure that when moving toward consensus, initial minority preferences are not ignored. While consensus would imply better voting outcomes, it could come at the cost of ignoring minority opinions. We measure whether this is a concern in our experiments by studying whether minority-supported candidates were selected by AV under different deliberation mechanisms.[8]

A candidate is either *minority-supported* or *majority-supported* based on the *initial* approval profile. We say that a candidate c is *minority-supported* if (pre-deliberation) the fraction of minority voters who include c in their approval ballot is greater than the fraction of majority voters who include c in their approval ballot. Formally, it can be defined as follows. The total approval of a candidate $c \in C$ can be written as $V(c) = V_{maj}(c) + V_{min}(c)$, where $V_{maj}(c)$ and $V_{min}(c)$ are the number of votes from the majority and minority agents, respectively. Candidate c is *minority-supported* if

$$V_{min}(c)/|N_{min}| \geq V_{maj}(c)/|N_{maj}|.$$

Table 3.2 reports the average number of pre-deliberation (initial) *minority-supported* candidates selected by AV (post-deliberation) across deliberation strategies. This serves as an indicator of whether minority preferences are *preserved*.

[8] This is not a concern for other rules since they are designed to achieve proportionality (MES and PAV) or diversity (CC).

Approval Voting	**MES** (initial) (0.917)	**PAV** (initial) (0.92)
Initial (0.838)	0.913	0.910
Homogeneous (0.88)	0.959	0.956
Random (0.952)	**1.038**	**1.034**
Heterogeneous (0.953)	**1.039**	**1.035**
Iterative Random (0.984)	**1.073**	**1.069**
Iterative Golfer (0.984)	**1.073**	**1.069**

Table 3.3: Average utility-representation aggregate score obtained by AV under different deliberation setups in comparison to the proportional rules under no deliberation.

In the *initial* setup (no deliberation), AV does not elect any *minority-supported* candidates. However, this improves as agents interact and deliberate with the broader population. Note that since the minority agents have similar preferences, and they constitute 20% of the population in our setup, a proportional committee would represent them with 1 (out of 5) candidate. As seen in Table 3.2, the *large group* setup comes close to the ideal outcome on average. Thus, with deliberation, AV can *preserve* and represent minority preferences.

3.6.4 "Simple" vs. "Complex" Voting Rules

We argue that the "complexity" of a voting rule can be measured along three axes. First, one can ask about the computational complexity of computing a winning outcome or committee (e.g. PAV is known to be NP-hard [4], whereas AV is polynomial). Second, there is growing work in better understanding the ramifications of ballot design and voting rules on the cognitive load of voters [11]. Finally, there is value in using simple explainable voting rules. Explainability engenders trust in the system (which in turn may impact engagement in participatory democratic processes).

While two of these dimensions are, somewhat subjective, we argue that AV can be viewed as being *simple* across all three, whereas CC, PAV, and MES are *complex* along at least one dimension. Specifically, the winning committee for AV can be computed in polynomial time, and as it comprises the candidates with the maximum support, it is straightforward to explain to the voters. On the other hand, determining the winning committee for PAV and CC is computationally hard, and we argue that both MES and PAV are complicated rules, making them hard to explain to the voters. Our hypothesis is that *simple* rules coupled with deliberation processes can do as well as more *complex* voting rules. To this end, we compare AV with deliberation to MES and PAV without deliberation,

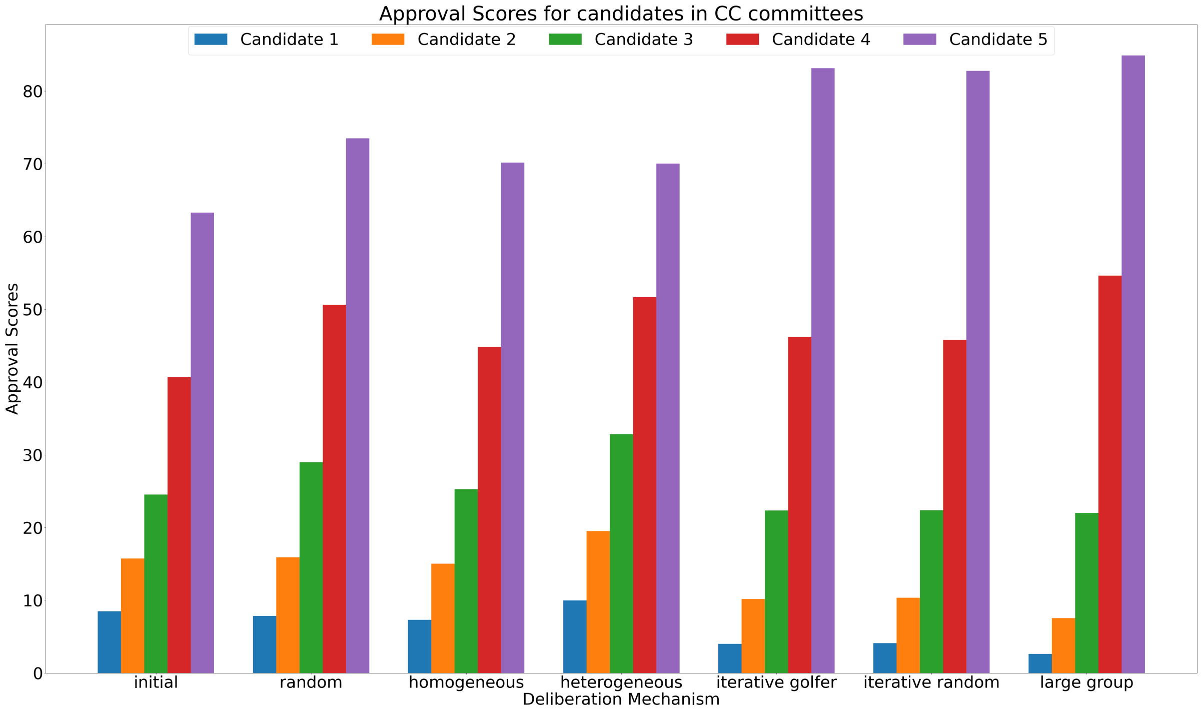

Figure 3.8: Average approval scores obtained by the 5 candidates in the winning committee chosen by CC across different deliberation mechanisms.

using the utility-representation aggregate score ($URagg(R)$) as our measure (Table 3.3). Values greater than 1.0 indicate that AV with the corresponding deliberation mechanism achieves a better $URagg$ score than MES/PAV without deliberation. These findings support our argument that "simple" rules coupled with effective deliberation strategies can be as effective as the "complex" rules.

3.6.5 Iterative deliberation with CC

In this section, we explain the odd drop in performance observed by CC in iterative deliberation and the *large group* setting (see Figures 3.4, 3.6, 3.7 and Table 3.1). Refer to Figure 3.8 for the average approval scores obtained by the winning candidates in the committees chosen by CC. The candidates (1 to 5) are ranked in increasing order of the number of approval votes they get (5 is highest).

We clearly observe that as we move from single round deliberation mechanisms to iterative methods (and *large group*), the approval votes for the highest supported candidate (5) increase and the same for the lowest supported candidate (1) decrease. For the iterative methods, approximately 80% of the agent population approves candidate_5 (≈ 90% for *large group*). This also reinforces the fact that iterative deliberation approaches consensus, as a major proportion of voters approve a single candidate. Accordingly, CC is able to represent approximately 80% of the voters with just one candidate. Since CC only cares about maximizing voter coverage, it chooses the rest of the candidates to represent the remaining voters. This leads to sub-optimal outcomes since instead of representing the population groups proportionally, CC optimizes for coverage and chooses candidates that might have very little support. This can be seen in Figure 3.8 as candidate_1 for the iterative methods and *large group* has less than 5% support. As a result, the almost 80% of the voter population that possibly gets only one representative in the final CC committee might be a cohesive voter group and thus, deserves more candidates for a fair and proportional outcome.

In conclusion, we see that with deliberation mechanisms that move towards consensus, CC exhibits a drop in welfare and proportionality guarantees since it is focused on maximizing representation. In general, other voting rules provide better overall performance than CC. However, if CC should ever be used with deliberation, we must pick an appropriate deliberation setup (single round) for the optimal outcome. This further shows that deliberation is not trivial and must be structured appropriately to obtain the best results.

3.7 Conclusion

In this chapter, we presented an empirical study of the relationship between deliberation and voting rules in approval-based multi-winner elections. In particular, we build a dynamic agent-based model of deliberation and investigate the performance of several standard voting rules under different deliberation strategies. Our results indicate that deliberation generally improves voting outcomes with respect to welfare, representation, and proportionality guarantees. Effectively designed mechanisms that increase exposure to diverse groups and opinions enhance the quality of deliberation, protect minority preferences, and in turn, achieve better outcomes. Importantly, we show that in the presence of effective deliberation, 'simpler' voting rules such as AV can be as powerful as more 'complex' rules without deliberation.

Our analysis provides encouraging insights to support the development of democratic research platforms such as Ethelo, Polis, and LiquidFeedback[9]. There are several promising directions for future work. First, our work could be extended to other democratic processes such as ranked-choice voting and participatory budgeting. Second, it would be interesting to conduct real-world user studies to support our results and understand the feasibility of our proposed deliberation methods. Another natural extension would be to explore other deliberation models and investigate if our results still hold true. Lack of transparency or complex voting procedures discourage participation in community-focused democratic procedures, especially from under-represented communities. Despite some calls for voting rules to revert to simplicity [67], there has not been considerable work towards improving simple voting rules for practical use cases. This work also suggests the need to empirically study the explainability of voting rules.

[9] https://ethelo.com/, https://pol.is/home, https://liquidfeedback.com/en/

Chapter 4

Fair Ranking through Proportional Voting

4.1 Introduction

The primary goal of an election or voting process is typically to aggregate individual preferences over a set of candidates and produce a (or a set of) winner(s). In the process of selecting a winner, the candidates are usually ranked using some metric based on the voting rule and the highest-ranked candidate(s) is/are chosen. Essentially, we can see how preference aggregation or voting is effectively a ranking task over the candidates. In the previous chapter, we established why fairness is desirable in election outcomes, when an outcome with multiple winners is deemed (proportionally) fair toward the voters, and how to achieve it. The same fairness objective translates to a ranking of candidates. For instance, consider a user population with approval preferences over a set of 30 candidates such that 50% of the users approve the first 10 candidates, 20% the next 10, and finally, 30% the last 10 candidates. Assuming the preferences are aggregated by Classical Approval Voting (AV), the final ranking of candidates is not proportionally fair to the voters because half of the population does not approve any of the first 10 candidates in the ranked list. Thus, it would be desirable to produce a fair ranking, and as seen in the previous chapter, proportional voting rules allow us to circumvent this issue by choosing candidates in a proportionally fair manner.

Participatory budgeting and liquid democracy are two election tasks that directly prioritize the need for fair ranking [5, 9]. In both scenarios, a ranked list of projects is generated to be reviewed and deliberated upon by citizens. Although, fair ranking is a concept that

	Male		**Female**	
White	A (10)	B (9)	C (7)	D (10)
Asian	E (9)	F (6)	G (7)	H (8)

Table 4.1: Example: Set of 4 candidates belonging to different groups. The quality scores are given in parentheses.

frequently arises outside of election domains, especially in the context of search engine and recommender system applications, where a ranked list of items must be generated based on some item relevance criteria and the user's needs. However, in the case of search or recommender systems, fairness becomes a multi-stakeholder concern [74]. Ideally, the ranked list should be fair to both the items being ranked and the users consuming the list (for example, artists and listeners on a music recommendation app, respectively). When considering item-side fairness, it is desirable to achieve a balanced and equitable distribution of exposure across the items presented in the list. This ensures that no specific group or individual is unfairly disadvantaged due to disproportionate representation in the list. An unfair ranking system would systematically and consistently assign lower rankings to items associated with a specific group, thereby perpetuating and replicating existing certain social biases. We explain this further with the following example.

Example 3 *Consider the situation in Table 4.1 where the objective is to rank 4 out of 8 candidates based on their relevance or quality scores. A purely relevance- or utility-focused ranking would generate a ranking of* $\{A, D, B, E\}$ *(B/E interchangeably), having a total utility (say, the sum of scores) of 38. However, we see that this ranking consists of only 1 Asian (E) and 1 female (D) candidate. A fair-ranking algorithm could instead generate* $\{A, D, E, H\}$*, consisting of 2 Asian (E, H), 2 White (A, D), 2 male (A, E), and 2 female candidates (D, H) and a total utility of 37. For only a slight drop in utility, this maintains fairness across all possible groups.*

The example above illustrates the challenge of creating a ranking of items that achieves both group fairness and high utility in terms of relevance. Often, there exists a genuine tradeoff between fairness and relevance. There are various reasons why a ranked list can exhibit bias or unfairness, such as biased annotations in the training data, bias in the relevance scores assigned to candidates, and potentially the use of sensitive features by a biased system to rank results. A fair-ranking system aims to address this issue by providing a ranked list that satisfies some pre-defined fairness criterion. However, the concept of fairness in ranking has been explored through various perspectives. Item-side fairness focuses on ensuring fairness to the items being ranked, while user-side fairness pertains to fairness for the consumers of the list. Moreover, methods for guaranteeing fairness in the ranking process can be categorized into three stages: pre-processing (mitigating bias in training data), in-processing (training bias-free models), and post-processing (re-ranking while adhering to fairness constraints). The fairness criteria also vary, including individual fairness, which aims to treat each candidate fairly regardless of their background, and group fairness, which aims to ensure proportional or equal representation for all groups. Group membership is often determined by sensitive attributes such as race, gender, and/or ethnicity. Informally, *statistical parity* refers to the notion that each group should be represented in the final ranked list in proportion to its population share. Additionally, it is important to consider that users typically pay more attention to the top items in a ranked list compared to lower-ranked items, and this factor should be taken into account when generating the final ranked list.

In this chapter, we are concerned with item-side post-processing group fairness. In the previous chapter, we extensively study the concept of proportional fairness with respect to voting and committee selection. Skowron et al. show that this can be naturally extended to produce a fair ranking instead of a committee [71]. However, their work explores proportional rankings that are fair towards voters (user-side), given approval preferences. General fair-ranking tasks either do not clearly have a voter-candidate relationship or do not have access to voter preferences. Taking inspiration from their work, we demonstrate that proportional voting rules can be used as strong post-processing item-side fair-ranking algorithms. We develop novel fair-ranking algorithms on top of the sequential version of proportional approval voting (SeqPAV) that achieve statistical parity among the groups of candidates being ranked. To utilize voting algorithms as ranking algorithms, we reframe the ranking task as a voting problem. In the previous chapter, we explored the notion that practical-use voting rules should be simple and explainable to ensure voter participation and transparency. However, complex voting rules like PAV (or SeqPAV) offer strong axiomatic fairness guarantees that can be advantageous for tasks where sacrificing some explainability is acceptable. Ranking in search presents an ideal use case where these voting algorithms can be effectively employed.

Contributions: In this chapter, we make the following contributions:

- We describe the post-processing item-side group fairness task and reformulate it as a voting problem to explore the use of proportional voting rules as fair-ranking algorithms.

- We develop four modified versions of SeqPAV as fair-ranking algorithms that can provide fairness across several sensitive attributes.

- Finally, we experimentally show that our SeqPAV-based ranking methods provide strong performance on both fairness and utility criteria and often match or beat the performance achieved by other standard fair-ranking algorithms.

4.1.1 Related Work

Fairness in ranking has recently gathered a lot of attention and has been extensively examined from several perspectives. Zehlike et al. present an extensive survey that comprehensively covers a range of fair-ranking constructs [86]. Within their study, they outline four major frameworks of fair-ranking algorithms and evaluate various methods within these frameworks. Additionally, they discuss the commonly utilized datasets in this domain.

Yang and Stoyanovich were one of the first to study fairness in rankings [82]. They focus on a single sensitive attribute and propose statistical parity-based fairness metrics measuring the relative representation of different (protected and non-protected) groups at different points in the list (top-10, top-20, etc.) while incorporating a position bias discount. In follow-up work, Zehlike at al. develop FA*IR, a fair-ranking method for the single sensitive attribute setting based on statistical tests ensuring a minimum proportion of protected candidates in every prefix of the ranked list [83]. This was later extended to account for multiple protected attributes per item [85]. Furthermore, Feldman et al. address the issue of eliminating disparate impact in datasets and propose a method to adjust the quality scores of candidates such that the resulting probability distribution of scores for protected and non-protected groups is similar [35]. CFAθ is another fair-ranking approach that works by modifying the score distributions for the protected candidates but works with multiple sensitive attributes in the dataset [84]. We use these three methods as benchmarks for comparison in the single- and multiple-attribute case.

Yang et al. later investigate the unintended reduction of *in-group* individual fairness that can occur when maximizing group fairness and utility in set selection and ranking algorithms [81]. They introduce metrics to measure in-group fairness in multiple-attribute settings and propose methods to mitigate this problem. Burke et al. provide a systematic overview of the multi-stakeholder nature of search and recommendation systems and also explore the use of social choice mechanisms for fair recommendation systems [16, 17]. Furthermore, Sapiezynski et al. introduce a new group fairness metric that accounts for user attention and position bias. We will use this metric for comparison in our experiments [66]. Singh and Joachims propose a set of group fairness metrics that evaluate statistical parity based on group exposure among ranking policies rather than a single ranked list [68]. Finally, given there are several fair-ranking metrics, Raj and Ekstrand develop a common notation to represent different metrics and enable a direct comparison among them [64].

Our work is situated within the context of addressing group fairness in ranked lists through post-processing techniques. However, we adopt a unique perspective by drawing inspiration from proportional fairness principles observed in voting or elections. By identifying shared fairness concepts in these two domains, we establish a connection and demonstrate the applicability of proportional voting methods to enhance fairness in general ranking tasks.

4.2 Background

4.2.1 Preliminaries and Notation

In this section, we define the notation necessary to understand the fair ranking task from the search system or information retrieval perspective. Following the general notation, let $C = \{c_1, c_2, ..., c_m\}$ be a set of m candidates (also referred to as items or documents) to rank. For each candidate $c_i \in C$, q_i denotes the "quality score" of candidate i, representing the overall quality or relevance of the candidate with respect to a specific search query or ranking task. This score is generally obtained by a pre-existing algorithm used to judge the quality of different candidates with respect to the search query. We assume the score is given to the ranking system. Finally, P denotes a set of protected or sensitive attributes and each attribute $p \in P$ takes one of a predefined set of values for that attribute. We use G to represent the set of all possible groups arising out of the protected attributes in P. For example, if the task is to rank job candidates, the protected attributes could be $P = \{$"race", "gender"$\}$, where "race" could be one of $\{$White, Person of Color (PoC)$\}$ and "gender" could be one of $\{$Male, Female, Non-Binary$\}$. In that case, G would be $\{$White, Person of Color, Male, Female, Non-Binary$\}$. Each candidate $c_i \in C$ belongs to one (or more) of the groups in G and the candidate's group association is represented by a group alignment vector $L_i \in [0, 1]^g$, where $g = |G|$. Extending this to the whole set of candidates, $L(C)$ is an $m \times g$ alignment matrix, where the rows correspond to candidates or items and the columns represent the groups.

Given a ranked list, user attention is expected to decrease for documents positioned at lower ranks in the list. This attention decay, referred to as *position bias*, is captured by a *position weight vector* a_R for ranked list R. Finally, fairness in a ranked list may be measured in terms of the exposure achieved by different groups in comparison to a target group distribution $\hat{p}$. This may be computed in several ways, depending on the fairness definition. Some examples include strict group equality, and proportional distribution according to the population demographics.

Problem Formulation: Given a list or set of candidates (or items) C belonging to different groups in G, produce a ranked list R of k candidates such that each group $g \in G$ is "fairly" represented in the ranking and the relevance utility of the ranking is maintained.

4.2.2 Fairness Measures

In this section, we cover different group fairness criteria and formally define the group fairness measure used in our experiments.

Fairness has been thoroughly studied in recent years and several metrics have been proposed to measure fairness in rankings, some borrowed from the machine learning fairness literature and others more specific to search or recommender systems and general ranking tasks. In this domain, two concepts that hold significant importance are *disparate treatment* and *disparate impact*. These concepts are frequently employed to address the idea of unfairness, aligning with the concepts of *direct* and *indirect* discrimination, respectively. Disparate treatment refers to the intentional differential treatment of various groups, either through the explicit use of sensitive attributes or other attributes deliberately causing discriminatory outcomes. On the other hand, disparate impact pertains to situations where the effects of a system differ among different groups, irrespective of intent [6].

Statistical parity [45] or demographic parity is a group fairness measure that refers to the idea of achieving comparable outcomes or proportional representation across groups. The goal is to ensure that all groups have equal or comparable rates of receiving positive outcomes with respect to a specific task (such as ranking). Proportional outcomes in this domain can be compared to proportionality in voting. We delve deeper into this idea later in this chapter. *Equality of opportunity* [42], on the other hand, is a fairness criterion that promotes *individual* fairness. It embodies the idea that qualified candidates should receive equal treatment regardless of their group membership.

Attention-Weighted Rank Fairness

In our work, we employ the commonly used fairness metric – *Attention-Weighted Rank Fairness* (AWRF). This metric, introduced by Sapiezynski et al. [66], measures the fairness of a ranked list by comparing the cumulative exposure across groups with a target distribution reflecting the population distribution.

Formally, the cumulative exposure across groups in a ranked list R is measured as $\epsilon_R = L(R)^T a_R$, where $L(R)$ is the group alignment matrix for candidates in R and a_R is the position weight vector. The resulting fairness metric, AWRF, measures the difference between the cumulative group exposure and the target distribution.

$$\text{AWRF(R)} = \Delta(\epsilon_R, \hat{p}) \tag{4.1}$$

The system's fairness is higher if the cumulative group exposure is closer to the target group distribution. This metric is suitable for evaluation because it allows soft association with respect to the groups, captures categorical protected attributes, accounts for position bias, and allows freedom to choose different decay methods for position bias. As per the TREC 2022 Fair Ranking Task [25], we define the distance function to be the Jenson-Shannon divergence, such that $\Delta(\epsilon_R, \hat{p}) = 1 - d_{JS}(\epsilon_R, \hat{p})$. The resulting metric is in the range $[0, 1]$, with 1 indicating a completely fair ranking as the distance between the target group distribution and the cumulative group exposure is minimized to be 0.

4.2.3 Relevance Measure

As defined in the problem formulation above, the ranking system should not only be fair but also produce a list that is of high relevance utility. A widely used measure of relevance in search rankings is the Normalized Discounted Cumulative Gain (NDCG). It is a normalized position-based performance metric that measures the cumulative gain of each item in the ranked list, where the relevance or quality of each item i is represented by q_i. The cumulative gain is normalized with respect to the ideal cumulative gain (IDCG), which is the highest possible gain achievable for the given set of items. For a ranked list R of length k, the metric is formally defined as:

$$\text{NDCG(R)} = \frac{1}{\text{IDCG}} . \sum_{i=1}^{k} \frac{q_i}{\log_2(i+1)}, \tag{4.2}$$

where q_i is the quality of candidate i, weighted by a logarithmic decay based on the position in the list.

The ideal ranking in terms of relevance utility, achieving an NDCG score of 1, would be the Top-k ranking (also known as "colorblind" ranking [83]). This ranking solely focuses on the qualifications of the candidates and arranges them based on their qualification scores, without taking fairness considerations into account. This results in a trade-off. A ranked list that demonstrates strong relevance performance might not fare well on fairness metrics due to the potential presence of inherent biases reflected in the relevance scores. Conversely, a list that prioritizes fairness alone would likely exhibit poor performance in terms of relevance. Therefore, it is essential to redefine the fair-ranking problem in a way that aims to generate a ranked list maximizing fair exposure according to AWRF while minimizing the loss of relevance as measured by NDCG. Finally, we employ the following aggregate metric to assess the performance of the ranked lists:

$$\mathbf{Metric_{agg}}(\mathrm{R}) = \mathrm{AWRF}(\mathrm{R}) * \mathrm{NDCG}(\mathrm{R}) \tag{4.3}$$

4.2.4 Proportional Rankings in Voting

Until now, our focus has primarily been on fairness and proportional representation within the context of search systems. Now, we shift our attention to a voting perspective and explore the concept of achieving a proportional ranking from this standpoint. The goal of our work is to bridge the gap between this literature and leverage the proportional characteristics of voting rules to provide fairness guarantees in a general ranking task.

In the previous chapter, we extensively study proportionality in multi-winner approval-based elections. However, the notion of proportionality, as defined for committee elections (refer to Section 2.3), does not directly apply to rankings. This distinction arises because the objective is not to select a committee but rather to create a ranking of candidates that embodies the principle of proportionality. Skowron et al. [71] extend the principle of proportional representation to rankings. Given approval ballots, the goal is to generate a ranking of candidates such that cohesive groups of voters are proportionally represented in each initial segment of the ranking. Drawing inspiration from Extended Justified Representation (EJR), the objective is to generate a ranking that guarantees each cohesive group of voters a proportional representation of their approved candidates in each initial segment of the ranking. We provide formal definitions below.

Given a profile P with a set of m candidates, C, and n voters, N, the average satisfaction of a group of voters $N' \subseteq N$ over a set of selected alternatives $S \subseteq C$ is defined as:

$$\mathrm{avg}(N', S) = \frac{1}{|N'|} \sum_{i \in N'} |A_i \cap S| \tag{4.4}$$

In the case of rankings, the subset of alternatives S is considered as the initial segment of a given ranked list R, i.e., $S = R_{\leq k}$ for some $k \in \{0, 1, .., m\}$. The objective is to ensure that a cohesive group of voters N' is represented in every segment of the ranking, with an average representation $\mathrm{avg}(N', R_{\leq k})$ that is proportional to the cohesiveness of the group. Cohesiveness of a group (similar to the idea of T-cohesive group) is defined formally as:

Definition 7 (Significant Group) *The cohesiveness of a voter group N' with proportion $\alpha(N') = |N'|/|N|$ is given by $\lambda(N') = |\cap_{i \in N'} A_i|$. Given $\alpha \in (0, 1]$ and $\lambda \in \{0, 1, ..., m\}$, a voter group $|N'|$ is (α, λ)-significant if $|N'| = \lceil \alpha n \rceil$ and $\lambda(N') \geq \lambda$.*

The proportional representation of a voter group in the ranked list depends on the group's significance and cohesiveness.

Definition 8 (Justifiable Demand) *The justified demand of a voter group $N' \subseteq N$ regarding the top-k positions in a ranked list is $jd(N', k) = \min(\lfloor \alpha(N') \cdot k \rfloor, \lambda(N'))$.*

The objective of a proportional voting rule is to ensure that every voter group gets an average representation that satisfies its justifiable demand, as defined above. However, it may not always be feasible to fulfill the demands of every group entirely. In such instances, the objective should be to ensure a substantial portion of the demand is met. Skowron et al. [71] provide theoretical guarantees for several voting rules and experimentally evaluate them in terms of average group representation. They record instances where the ranking produced by a voting rule *violates* the justified demand of a voting group, i.e., where the average representation for a voting group is less than its justified demand. Their results indicate that the sequential variant of PAV (Proportional Approval Voting) [76] is one of the best-suited rules to generate proportional rankings. The voting rule is defined below:

Sequential Proportional Approval Voting (SeqPAV). This rule represents the sequential version of PAV, where a ranking is established by progressively adding candidates to an initially empty ranking set $R = ()$. Recall, the PAV-score for a subset of alternatives $S \subseteq C$ is given by $sc_{pav}(A, S) = \sum_{i \in N} h(|S \cap A_i|)$, where $h(t) = \sum_{i=1}^{t} 1/i$. At step

$k \in \{0, 1, ..m\}$, it picks an unranked candidate c that maximizes $sc_{pav}(A, R_{\leq k-1} \cup \{c\})$, i.e., the candidate that improves the ranking's PAV-score the most. Ties are broken according to some specified order.

While these findings offer strong guarantees for proportional voting rules, they may not directly translate to general ranking tasks, which require considering both fairness and relevance as essential metrics. Drawing inspiration from this analysis, we expand upon this research in various ways. We develop fair-ranking algorithms that leverage proportional voting rules to ensure fairness guarantees while minimizing the loss of relevance utility.

4.3 Methodology

In this section, we define our proposed fair-ranking algorithms. We first describe how we reformulate the proportional rankings problem from the approval-based voting perspective to the general fair-ranking task and then define our proposed fair-ranking algorithms based on sequential PAV.

4.3.1 Fair Ranking as a Voting Task

Transforming the proportional rankings task into a suitable framework for the broader fair ranking objective is not straightforward. The first challenge is to incorporate or translate the voter-candidate relationship. Since the general fair ranking task is not modeled as an election, it focuses solely on candidates and does not explicitly mention voters or voter preferences. Second, for proportional rankings in voting, the objective is to produce a ranking of candidates that is fair to groups of voters. In contrast, the general fair ranking task seeks to create a ranked list that is fair towards groups of candidates with distinct protected attributes. Lastly, when constructing the final ranked list, we must also consider relevance criteria that account for the candidates' quality scores, in addition to fairness considerations.

To reformulate the general fair ranking task as a voting problem, we consider an election $E = (C, G)$, where the candidates to be ranked, C, are the alternatives, and their respective protected attributes, G, forms the set of all voters. The approval ballot A_{g_i} for each voter (attribute) $g_i \in G$ is the set of all candidates in C that belong to the group g_i. Given this approval-based election setup, we can now fit the problem as a proportional ranking task as per the definitions in Section 4.2.4. A proportional voting rule, say SeqPAV, would produce a ranked list of candidates (items), R, that proportionally represents cohesive groups of voters (in this case, protected attributes). The resulting list would ideally provide intersectional fairness across different protected attributes since each attribute denotes a voter. Finally, one of the ways we incorporate relevance in the resulting ranked list is by breaking ties according to the quality scores for candidates. The following example illustrates the reformulation.

Example 4 *Consider the task of ranking 5 job candidates given by $C = \{a, b, c, d, e\}$, with protected attributes "race" and "gender", where "race" could be one of {White, Person of Color (PoC)} and "gender" could be one of {Male, Female, Non-Binary}. In that case, G would be {White, PoC, Male, Female, Non-Binary}. Say, candidate a is (White, Male), b is (White, Female), c is (PoC, Male), d is (PoC, Non-Binary), and e is (White, Male).*

We would now translate this into an election such that $C = \{a, b, c, d, e\}$ is the set of candidates, and the set of attributes, G, is the set of voters. The voter ballots are given by $Male = \{a, c, e\}, Female = \{b\}$, Non-Binary $= \{d\}, White = \{a, b, e\}$, and PoC $= \{c, d\}$.

4.3.2 Algorithms

We now define our proposed algorithms based on SeqPAV. They take as input the transformed election task defined above and return a ranked list.

Default

We begin with the default variant of SeqPAV, which does not take into account the quality scores of the candidates explicitly. Recall that for each candidate $c_i \in C$, q_i denotes its "quality score", representing the overall quality or relevance of the candidate with respect to a specific search query or ranking task. The quality scores are used only to break ties between candidates. Essentially, the default variants only produce a ranked list that provides fair representation towards the groups of voters (attributes), ignoring the relevance utility of the ranking in the overall process. However, we consider two variations of the default rule: **Weighted** and **Equal** (Unweighted).

- **Weighted:** In the **weighted SeqPAV** approach, voters, who in this case represent protected groups or attributes, are assigned weights based on their respective voter group proportion. From the definitions above, the weight of a voter (attribute) $g_i \in G$ is the fraction of candidates belonging to that group, i.e., weight $w_{g_i} = |A_{g_i}|/|C|$. By assigning weights to voters based on their proportions, SeqPAV ensures that the ranked list provides proportional representation of voters (protected groups) according to their size or proportion. This approach generates a ranked list that reflects the distribution of the underlying population and produces better results when the target group distribution $\hat{p}$ is defined to maintain statistical parity.

- **Equal:** This method represents the default variation of SeqPAV, where all voters are assigned equal weights. In situations where the distribution of the underlying population is uneven, but there is a desire to generate a ranked list that treats all attributes (voters) equally, this method is suitable.

Scored

In order to generate a fair ranked list that explicitly takes into account the quality scores of the candidates, we make modifications to SeqPAV accordingly. Recall, at step $k \in \{0, 1, ..m\}$, it picks an unranked candidate c that maximizes $sc_{pav}(A, R_{\leq k-1} \cup \{c\})$, i.e., the candidate that provides the maximum marginal contribution to the ranking. Specifically, we modify the utility function such that for each candidate, the marginal utility gained by adding it to the ranked set also accounts for its quality score q_c:

$$\text{marginalUtility(c)} = \sum_{i \in N} (h(A_i \cap (R_{\leq k-1} \cup \{c\})) - h(A_i \cap R_{\leq k-1})) \cdot q_c \tag{4.5}$$

We consider both weighted and equal versions of the **scored SeqPAV** method for our experiments. In conclusion, we get a total of four variants of SeqPAV for fair ranking. A general algorithm pseudocode for the four methods is given in Algorithm 1.

In the following section, we will conduct experimental evaluations of the SeqPAV-based ranking methods on various widely-used datasets. We will compare their performance, in terms of the aggregate relevance and fairness metric (as defined in equation 4.3), with the baseline ranking methods.

4.4 Experimental Setup

In this section, we describe our experimental setup in detail. We first describe the datasets used in our experiments and then list the baseline ranking methods chosen for comparison.

Algorithm 1: General SeqPAV fair-ranking algorithm

```
Input: Candidate Set $C$, Voter Set or Attribute Set $G$, quality scores for all
       candidates: $q_c, \forall c \in C$, weighted $\in \{0, 1\}$ and method $\in \{default, scored\}$
       to indicate the SeqPAV method, and Ranked List Size $k \in \{0, 1, ..., m\}$.
Output: Ranked list $R$ of size $k$ satisfying fairness condition while minimizing
        loss in relevance utility.
// creating approval ballots for voters
$A_g \leftarrow$ set of all candidates in $C$ that belong to group $g, \forall g \in G$
if weighted is 0 then
    $w_g \leftarrow 1, \forall g \in G$
else
    $w_g = |A_g|/|C|, \forall g \in G$
Ranking Set $R \leftarrow []$
$j \leftarrow 0$
while $j < k$ do
    $C' \leftarrow$ unranked candidates
    if method is default then
        // Pick unranked candidate with highest PAV-score improvement.
        // Break ties by higher quality score $q_c$.
        $c_{max} \leftarrow \arg\max_{c \in C'} sc_{pav}(A, R_{\leq j-1} \cup \{c\})$
    else if method is scored then
        // Pick unranked candidate with highest marginal utility.
        // Accounts for candidate's quality score $q_c$.
        $c_{max} \leftarrow \arg\max_{c \in C'}$ marginalUtility(c) as per equation 4.5
    Ranking Set $R[j] \leftarrow c_{max}$
    Remove $c_{max}$ from $C'$
    $j \leftarrow j + 1$
return $R$
```

4.4.1 Datasets

We use multiple publicly available datasets for our experiments, with each dataset representing a group of individuals characterized by specific demographic attributes. These datasets also include a quality "score" attribute for each individual. In our experiments, we conduct tests for various values of k on each dataset. Additionally, we test multiple versions of each dataset, where the protected attributes are varied. However, it is important to note that the datasets (and their permutations) utilized in this study are directly based on those defined for the baseline methods [83, 85]. Based on prior work, we use various versions of these datasets that encompass both single and multiple protected attributes. This choice enables a direct comparison between SeqPAV and the baseline methods. Additionally, we include two more datasets that represent more challenging ranking tasks. A brief description of each dataset is given below. Table 4.2 and 4.3 present an overall summary of these datasets.

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) [1] is an assessment tool designed to predict the likelihood of recidivism for individuals who have been convicted. It employs a comprehensive set of over a hundred items or questions to score the probability of reoffending and is currently used by several jurisdictions in the United States. However, a recent study has provided evidence that COMPAS exhibits racial discrimination by having a higher false positive rate specifically for African Americans [1]. The objective is to produce a fair ranked list of top-k individuals in the dataset who are least likely to recidivate given their scores. Candidate scores are calculated as a weighted aggregate of the columns “recidivism, “violent recidivism” and “prior arrests” in the original dataset [85]. For single attribute methods, the attributes considered are “race” and “sex”. In the multiple attribute case, the groups considered are a result of different combinations of the attributes “race”, “age”, and “sex”. “Race” can be categorized as either “white” or “PoC”, “age” is either “younger than 25”, “between 25 and 45”, or “older than 45”, and finally, “sex” is classified as either “male” or “female”.

German Credit dataset [46] is the Statlog German Credit Dataset that contains credit ratings for individuals given by the German agency *Schufa*. It reports a credit-worthiness score, which refers to the quality score for each candidate. This score is calculated as a weighted sum of credit duration, credit amount, and employment length. “Sex” and “age” are used separately and in combination as the attributes.

LSAT dataset [79] was compiled to investigate potential disparities in admission criteria for law schools in the United States. The “qualification” score for a candidate is given by the score obtained in the US Law School Admission Test (LSAT). Group attributes in this dataset are “sex” and “race”.

CS Rankings dataset contains information about computer science departments in the United States[1]. Assembled and used by [81], this dataset considers “publication count”

[1] https://csrankings.org/

Dataset	# Items	Ranking Size (k)	Quality Score	Sensitive Attribute	Protected Group (%)
D1 - COMPAS	6173	1000	recidivism score	race	African American (51.2%)
D2 - COMPAS	6173	1000	recidivism score	sex	Female (19.3%)
D3 - German Credit	1000	100	credit rating	sex	Female (31%)
D4 - German Credit	1000	100	credit rating	age	<25 yr. (14.9%)
D5 - German Credit	1000	100	credit rating	age	<35 yr. (54.8%)

Table 4.2: Datasets and experimental settings for single sensitive attributes

Dataset	# Items	Ranking Size (k)	Quality Score	Sensitive Attributes	# Groups
D1 - COMPAS	6173	1500	recidivism score	race	2
D2 - COMPAS	6173	500	recidivism score	age	3
D3 - COMPAS	6173	300	recidivism score	race, sex, age	4
D4 - German Credit	1000	50	credit rating	sex, age	6
D5 - LSAT	21,792	300	LSAT score	sex, race	4
D6 - CS Rankings	51	16	publication count	department size, geographic area	10
D7 - MEPS	960	20	utilization	race, age	16

Table 4.3: Datasets and experimental settings for multiple categorical sensitive attributes

as the quality score of a candidate (department). The group attributes are “department size”, categorized as “large” or “small”, and “geographic area”, classified as “North East”, “West”, “Middle West”, “South Center”, and “South Atlantic”.

MEPS (Medical Expenditure Panel Survey) serves as an extensive dataset providing individual and household-level information on health expenditures made by individuals belonging to various demographic or socioeconomic groups. A candidate’s quality score corresponds to the “utilization” feature in the dataset, which represents the total number of trips requiring medical care. It is calculated by summing the number of office-based visits, outpatient visits, ER visits, inpatient nights, and home health visits. Our analysis is conducted on a specific subset of the dataset, which is the same portion utilized by [81] in their study. Specifically, it is the data from Panel 20 of the calendar year 2016, consisting of the top 960 individuals with a utilization score exceeding 5. The sensitive attributes are “race” (with values “White”, “Black”, “Multiple races”, “Native Hawaiian”, “Asian Indian”, “Filipino”, “Chinese”, and “American Indian”) and age (“Middle” and “Young”).

4.4.2 Baseline Methods

Below, we list the baselines utilized for comparison and briefly describe each method. We generate top-k rankings for all the datasets mentioned earlier and evaluate the performance of these baselines compared to the ranking methods based on SeqPAV.

Single Attribute Methods

- **Colorblind:** This approach generates a top-k ranking solely based on the quality scores of the candidates, without incorporating any group fairness constraints. As a result, this method maximizes the relevance utility of the ranking.
- **Feldman et al.** [35]**:** Motivated by the problem of removing *disparate impact*, this ranking method aims to align the probability distribution of the protected candidates with that of the non-protected candidates. Specifically, this is achieved by replacing the quality score of a candidate i in the protected group with a candidate j in the non-protected group: $q_i \leftarrow q_j$, such that $F_n(j) = F_p(i)$, where $F_n(\cdot)$ and $F_p(\cdot)$ are the quantiles of a candidate among the non-protected and protected candidates, respectively.
- **FA*IR** [83]**:** Designed for the binary group setting, this approach works on the assumption that a ranking is fair when candidates are chosen based on a Bernoulli distribution (coin tosses) that remains unaffected by the candidate's sensitive attributes. In particular, this method ensures that the proportion of protected candidates in the ranking does not significantly deviate below a predetermined minimum percentage p. It accomplishes this by formulating a fairness condition that tests the statistical significance of whether the generated ranking is likely to have been produced by a Bernoulli process. The group fairness condition dictates that a ranking prefix of length k having τ_p protected candidates fairly represents the protected group with minimum proportion p and significance α if $F(\tau_p; k, p) > \alpha$, where F corresponds to the cumulative distribution for a binomial distribution.

Multiple Attribute Methods

- **Multinomial FA*IR** [85]**:** The multinomial version of the FA*IR algorithm extends the same fairness framework to account for multiple attributes or groups by replacing the Bernoulli distribution with a dice roll. Particularly, the process of choosing candidates for each position in the ranked list is equivalent to rolling a $|G|$-sided dice, where each side of the dice represents one group $g \in G$. For each group g, the minimum proportion is given by p_g. This method guarantees a minimum percentage of representation for each group in the ranking. The multinomial FA*IR algorithm is tested with two variants. The first variant, p_{stat}, corresponds to the statistical parity setting where the minimum proportion for each group (p-value) is the same as its respective proportion in the dataset. The second variant, p_{eq}, considers all p-values to be equal.

- **Categorical sampling ("dice roll") [85]:** This method follows the same approach as multinomial FA*IR but does not provide any guarantees with respect to the minimum proportions. In other words, this procedure generates a ranking based on the dice roll procedure described above. Repeating this process many times would result in an approximation of the multinomial FA*IR method. This is used for comparison and the mean and standard deviation for 10,000 rankings created through this method are reported. The dice roll baseline is also tested with the same two variants (p_{eq} and p_{stat}) as the multinomial FA*IR algorithm.

- **Continuous Fairness Algorithm (CFAθ) [84]:** This method introduces a framework where the score distributions of the protected candidates are modified such that they align with the Wasserstein-barycenter of all group distributions. This new score distribution for each group is obtained by interpolating between the barycenter and the group distribution, within the constraint of a predefined fairness parameter $\theta \in [0, 1]$. They consider the new score distribution to be the fair score representation for each group. θ dictates the fairness emphasis as a high value corresponds to more group fairness and a low value corresponds to more individual fairness. At $\theta = 1$, the algorithm achieves statistical parity. Thus, this method is compared only against the ranking methods based on achieving statistical parity.

4.5 Results

In this section, we present and analyze the outcomes of our experiments. We generate ranked lists for all baseline approaches mentioned in Section 4.4.2, as well as the SeqPAV ranking methods defined in Section 4.3.2, across all datasets with the experimental configurations described in Section 4.4.1. We then measure the performance of all generated ranked lists on the relevance (NDCG), fairness (AWRF), and aggregate (NDCG*AWRF) metrics as defined in Sections 4.2.2 and 4.2.3, respectively. For calculating AWRF, we set the target group distribution $\hat{p}$ to the statistical parity vector, i.e., it matches the group proportions in the whole dataset. Accordingly, we expect the weighted SeqPAV methods and the statistical parity-based baselines to perform better than their corresponding equal versions.

Our experiments are driven by several objectives. Firstly, we aim to investigate the performance of various SeqPAV ranking methods on both single and multiple attribute datasets and compare their effectiveness against the baselines. Additionally, we seek to gain insights into the relative performance of different SeqPAV methods.

4.5.1 Single Attribute

Results for the single attribute experiments are shown in Table 4.4. First, we observe that SeqPAV's scored methods perform the best on all datasets as they achieve the maximum aggregate relevance-fairness scores. It is worth noting that the weighted SeqPAV version consistently beats the FA*IR baseline on the aggregate metric across all datasets.

Within the SeqPAV methods, we expect the weighted method to achieve higher scores since it is designed to achieve statistical parity and result in a cumulative exposure that is closer to the target distribution. However, in cases where the target group distribution is nearly equal across groups (D1 and D5), the SeqPAV equal methods demonstrate comparable performance. Finally, we observe that the scored method consistently outperforms the corresponding default method on the aggregate metric as it manages to improve the relevance utility (NDCG) by explicitly accounting for the marginal utility gained by candidate quality scores in the ranking, with minimal loss of fairness. However, if fairness is the only criterion, the default methods usually achieve a better AWRF score compared to the corresponding scored method. This is expected since the default methods do not account for relevance utility and manage to maximize fairness by following the default proportional approval voting principles.

While these results are promising for the use of SeqPAV ranking methods, we must note that the experimental setup might not be challenging enough. This is evident from the fact that the relevance and fairness values across all methods are quite comparable, exhibiting only minor differences. Furthermore, except for datasets D1 and D5, even the baseline colorblind ranking produces quite high fairness scores, indicating that the top-k ranking itself is quite fair. However, SeqPAV does show a promising direction in terms of performance and utility. In [83], the authors highlight FA*IR's versatility in producing rankings for multiple values of p, enabling flexibility in controlling fairness guarantees for different groups. We emphasize that SeqPAV offers a similar level of flexibility by adjusting the weights assigned to voters (groups) in the election setup.

In the following section, we explore the more challenging multiple-attribute setup.

4.5.2 Multiple Attributes

We now report results for the datasets with multiple sensitive attributes. To allow for a meaningful comparison, the results in Tables 4.5 and 4.6 are reported such that comparable methods are grouped together. The weighted SeqPAV ranking methods are comparable to the p_{stat} methods for multinomial FA*IR and the dice roll baseline, and CFAθ. The equal SeqPAV versions, on the other hand, are comparable to the p_{eq} methods.

The general trend of results follows from the single attribute methods. SeqPAV's scored and weighted methods perform the best on almost all datasets in terms of the aggregate fairness-relevance metric. Both SeqPAV-weighted methods consistently outperform the multi FA*IR p_{stat} method on the fairness metric (AWRF) while achieving comparable relevance (NDCG) scores. This improvement is highest for dataset D4, where the weighted SeqPAV methods achieve an AWRF score that is 2.7% higher than that for FA*IR p_{stat}. CFAθ exhibits strong fairness performance (except D2) but is outperformed by the SeqPAV's scored weighted method in all cases, with the default weighted method achieving comparable performance. Since the dice roll p_{stat} baseline follows the same process as FA*IR p_{stat} but without minimum fairness guarantees, it usually performs the worst among the statistical parity methods (except in D5)[2]. The benefit of providing fairness guarantees in FA*IR p_{stat} is also highlighted by the authors in [85].

We now move our attention to the p_{eq} methods. Following the experimental setup, they are expected to achieve worse performance on AWRF compared to the p_{stat} methods since the target distribution is based on statistical parity. However, we focus on the comparison between the equal SeqPAV versions and the p_{eq} FA*IR baseline. We notice that the scored SeqPAV equal method beats FA*IR p_{eq} on the aggregate metric across all datasets except D1 where the performance is comparable. Again, the highest improvement is observed for D4 where SeqPAV scored equal is higher on the aggregate metric by 20.3%. SeqPAV default equal also achieves comparable performance and even beats FA*IR p_{eq} on D5.

Datasets D6 & D7

Observing the results for datasets D1 to D5, we take the benchmark baseline method, FA*IR, and compare its performance with SeqPAV methods on more challenging datasets D6 and D7. These datasets present a greater challenge due to their inclusion of a significant number of intersectional groups (10 and 16, respectively). Results are shown in Table 4.7.

[2] The results for dice roll baseline were not available for D2 in the codebase shared by the authors.

For the CSRankings dataset (D6), we observe that the SeqPAV default equal method achieves the highest fairness (AWRF) score, comfortably beating the other methods. This emphasizes the strength of the default voting method in terms of fairness guarantees, especially when the statistical parity target distribution is close to equal. However, this comes at the cost of lower relevance scores (NDCG). The scored versions, on the other hand, achieve the desirable balance between relevance and fairness and outperform the FA*IR baselines on the aggregate metric. A similar comparison between the SeqPAV methods can be made for D7, where the default versions achieve higher fairness but the scored variants perform much better on the aggregate fairness and relevance metric. However, in this case, since the statistical parity distribution is close to equal, the overall best performance on the aggregate metric is observed by SeqPAV scored equal method, closely followed by the corresponding weighted version. Finally, we again observe that SeqPAV-based methods outperform the corresponding FA*IR baselines.

4.6 Discussion

Based on the findings in the preceding section, it is evident that ranking methods based on SeqPAV exhibit robust fairness and relevance performance when compared to the baselines. In this section, we will emphasize the key points of comparison and delve into a more detailed analysis of their performance.

4.6.1 SeqPAV vs. FA*IR

Starting with the simple single attribute tasks, Table 4.4 clearly demonstrates that SeqPAV-scored methods surpass FA*IR in all datasets on the aggregate fairness-relevance metric. However, it is evident that the performance between the two is highly comparable, displaying only marginal variances. The purpose of our analysis is to showcase the adaptability of a voting rule designed for proportional outcomes in elections, which can be effectively repurposed as a robust ranking algorithm with minimal adjustments. The positive results noticed in single attribute setups allow us to extend our analysis to the more challenging multiple attribute scenarios.

The multiple attribute setup poses a more challenging requirement to provide intersectional fairness across groups. Through Tables 4.5, 4.6, and 4.7 we notice that SeqPAV replicates the strong performance exhibited in the single attribute setup. For all multiple attribute datasets (D1 to D7), SeqPAV-scored methods either match or outperform the multinomial FA*IR baseline.

In conclusion, our experiments show that SeqPAV exhibits strong fair-ranking properties that match and often outperform the baselines on both single and multiple attribute setups. Additionally, SeqPAV provides the same flexible advantage as FA*IR as it also allows the creation of rankings with fairness guarantees beyond statistical parity. However, SeqPAV holds an additional advantage over FA*IR. While FA*IR requires modifying the input dataset to consolidate multiple categorical group attributes into a single group attribute using combinations of individual values, SeqPAV can handle input datasets with multiple categorical attributes without the need for preprocessing the data. Finally, it is worth noting the computational advantages of SeqPAV-based ranking methods as they are polynomial time computable. These methods prove to be lightweight in computation compared to the overwhelming computational burden observed in the baselines. Our aim is to further develop upon these findings by carrying out a comprehensive analysis of the computational complexity, comparing SeqPAV-based ranking methods with the baselines.

4.6.2 Comparison among SeqPAV methods

Now, we delve deeper into the variations in performance among the various SeqPAV methods. It is worth noting that for most datasets, even the SeqPAV default equal method exhibits strong performance, especially on fairness guarantees. It achieves a significantly higher fairness score compared to the corresponding multinomial FA*IR p_{eq} (5.88%) and SeqPAV scored equal methods (9.26%) on D6. We now explore the comparison in more detail.

First, we focus on the comparison between the weighted and equal versions. We repeat that our experimental setup is designed to judge fairness according to statistical parity. Thus, unless the target group distribution is close to equal, we expect the weighted methods to outperform the equal methods on fairness. However, we notice that when the target distribution is close to equal across groups (see D1 and D5 for single attribute setup and D6 and D7 for multiple attribute setup), the SeqPAV-equal methods match or even outperform the corresponding weighted method. We believe this arises due to the strong proportional guarantees provided by the default SeqPAV voting rule. In situations where it is desirable to maintain an equal distribution across groups, the off-the-shelf (equal) SeqPAV rule can serve as a useful fair-ranking algorithm.

The comparison between the default and scored SeqPAV versions is more straightforward. For all datasets, we observe that the scored variant performs better than its corresponding default variant on the aggregate fairness-relevance metric. However, we do notice that the default version provides better fairness scores when the fairness criteria are more challenging (see AWRF for D6 and D7, Table 4.7). This is because the default SeqPAV version creates a ranking driven solely by the fairness requirements and does not take into account the qualification scores of candidates. However, this results in high fairness scores but lower relevance scores compared to the scored SeqPAV variants. Contrarily, the scored SeqPAV methods offer a favorable trade-off between fairness and relevance utility. These scored methods consistently outperform their respective default counterparts in the aggregate metric by effectively enhancing relevance utility (NDCG). This improvement is achieved by explicitly considering the marginal utility gained from candidate quality scores in the ranking process while maintaining a minimal loss in fairness.

4.7 Conclusion

In this chapter, we reformulate the fair-ranking problem as a voting task and present a novel use case for sequential proportional approval voting rules. We illustrate the applicability of committee voting rules designed for proportional outcomes as post-processing fair-ranking algorithms, using SeqPAV as a case study. We develop four modified versions of SeqPAV that can explicitly account for both relevance and fairness demands.

Through experimentation on diverse datasets covering scenarios with a single sensitive attribute as well as more complex scenarios involving multiple sensitive attributes, we empirically evaluate the performance of SeqPAV-based ranking methods against various baselines. Our findings demonstrate that SeqPAV-based algorithms consistently achieve strong performance on the aggregate fairness-relevance metric, often matching or surpassing the performance of FA*IR and other baseline methods.

Importantly, these modified algorithms can cater to a range of use cases, accommodating different fairness-relevance trade-off requirements specified by the ranking task. The degree of fairness guarantees can be controlled by assigning appropriate weights to voters, while the relevance criterion can be adjusted by selecting the suitable method (default or scored). The flexibility offered by these methods proves highly valuable for regulatory agencies in tackling a wide array of fair-ranking tasks. Additionally, it is important to highlight that the default SeqPAV method offers robust axiomatic fairness guarantees. Although the modified SeqPAV methods may sacrifice some of these properties, there is a possibility that certain theoretical guarantees could still be maintained. Exploring the changes in these theoretical guarantees resulting from the modifications would be an intriguing direction for future research.

Dataset	**Method**	**NDCG**	**AWRF**	**NDCG*AWRF**	**% Prot. Output**
D1 (51.2%) - COMPAS, k=1000, race	SeqPAV scored equal	0.9852	0.9988	**0.984**	0.461
	FA*IR	0.9858	0.9979	0.9837	0.463
	SeqPAV scored weighted	0.9837	0.9993	0.983	0.472
	SeqPAV default equal	0.9799	0.9999	0.9798	0.5
	SeqPAV default weighted	0.978	1	0.978	0.512
	Feldman et al.	0.9779	1	0.9779	0.512
	Colorblind	1	0.9586	0.9586	0.252
D2 (19.3%) - COMPAS, k=1000, gender	SeqPAV scored weighted	0.998	1	**0.9979**	0.203
	SeqPAV default weighted	0.9973	1	0.9973	0.192
	Feldman et al.	0.9972	1	0.9972	0.194
	Colorblind	1	0.9966	0.9966	0.278
	FA*IR	1	0.9966	0.9966	0.278
	SeqPAV scored equal	0.9877	0.9576	0.9459	0.461
	SeqPAV default equal	0.9825	0.9469	0.9303	0.5
D3 (31%) - German Credit, k=100, gender	SeqPAV scored weighted	0.9993	0.9997	**0.999**	0.3
	SeqPAV default weighted	0.9991	0.9999	0.9989	0.31
	FA*IR	0.9994	0.9994	0.9989	0.3
	Colorblind	1	0.9976	0.9976	0.26
	Feldman et al.	0.9976	1	0.9975	0.31
	SeqPAV scored equal	0.9909	0.9876	0.9786	0.46
	SeqPAV default equal	0.9881	0.9824	0.9707	0.5
D4 (14.9%) - German Credit, k=100, age <25	SeqPAV scored weighted	0.9992	0.9992	**0.9984**	0.13
	SeqPAV default weighted	0.9982	0.9998	0.998	0.15
	FA*IR	0.9983	0.9997	0.998	0.15
	Feldman et al.	0.9953	0.9996	0.9949	0.15
	Colorblind	1	0.9948	0.9948	0.09
	SeqPAV scored equal	0.9677	0.9486	0.9179	0.43
	SeqPAV default equal	0.9575	0.9292	0.8897	0.5
D5 (54.8%) - German Credit, k=100, age <35	SeqPAV scored equal	0.9927	0.997	**0.9897**	0.47
	SeqPAV default equal	0.9908	0.9985	0.9893	0.5
	SeqPAV scored weighted	0.9883	0.9995	0.9878	0.51
	FA*IR	0.9914	0.9961	0.9876	0.5
	Feldman et al.	0.9854	1	0.9853	0.55
	SeqPAV default weighted	0.9853	1	0.9853	0.55
	Colorblind	1	0.9523	0.9523	0.24

Table 4.4: Results for single attribute datasets and methods on fairness, relevance, and aggregate metrics with the percentage of the protected group in the output ranked list. The best-performing method on the aggregate fairness-relevance metric is bolded.

Dataset	Method	NDCG	AWRF	NDCG*AWRF
D1 - COMPAS, k=1500, race, 2 groups	multi FA*IR p_{stat}	0.996	0.9993	**0.9954**
	SeqPAV scored weighted	0.9954	0.9999	0.9953
	CFAθ	0.995	0.9998	0.9948
	SeqPAV default weighted	0.9944	1	0.9944
	dice roll p_{stat}	0.993	0.9999	0.9929
	multi FA*IR p_{eq}	1	0.9886	0.9886
	colorblind	1	0.9884	0.9884
	dice roll p_{eq}	1	0.9878	0.9878
	SeqPAV scored equal	1	0.9871	0.9871
	SeqPAV default equal	0.9999	0.9868	0.9867
D2 – COMPAS, k=500 age, 3 groups	SeqPAV scored weighted	0.9694	0.9987	**0.9682**
	multi FA*IR p_{stat}	0.9646	0.9979	0.9626
	SeqPAV default weighted	0.9616	1	0.9616
	CFAθ	0.962	0.9244	0.8892
	SeqPAV scored equal	0.9484	0.976	0.9256
	multi FA*IR p_{eq}	0.942	0.9666	0.9106
	SeqPAV default equal	0.9352	0.9716	0.9087
	colorblind	1	0.8748	0.8748
D3 - COMPAS, k=300, race, age, sex, 4 groups	multi FA*IR p_{stat}	0.982	0.9959	**0.9779**
	SeqPAV scored weighted	0.9784	0.9983	0.9768
	CFAθ	0.9764	0.9986	0.975
	SeqPAV default weighted	0.9727	0.9997	0.9724
	dice roll p_{stat}	0.9647	0.9989	0.9636
	colorblind	1	0.9356	0.9356
	SeqPAV scored equal	0.8349	0.8524	0.7116
	multi FA*IR p_{eq}	0.8191	0.8532	0.6989
	SeqPAV default equal	0.7913	0.8004	0.6333
	dice roll p_{eq}	0.7917	0.7947	0.6291

Table 4.5: Results for multiple attribute datasets (D1 to D3) and methods on fairness, relevance, and aggregate metrics. The best-performing method (non-colorblind) on the aggregate fairness-relevance metric is bolded. Comparable results are grouped together.

Dataset	**Method**	**NDCG**	**AWRF**	**NDCG*AWRF**
D4 — German Credit, k=50 sex, age, 6 groups	SeqPAV scored weighted	0.9732	0.9961	**0.9694**
	CFAθ	0.9574	0.9989	0.9563
	multi FA*IR p_{stat}	0.9815	0.9698	0.9518
	SeqPAV default weighted	0.946	0.996	0.9422
	dice roll p_{stat}	0.944	0.9749	0.9204
	colorblind	1	0.9806	0.9806
	SeqPAV scored equal	0.9361	0.9445	0.8842
	multi FA*IR p_{eq}	0.8315	0.8838	0.7348
	SeqPAV default equal	0.8295	0.883	0.7325
	dice roll p_{eq}	0.7685	0.8395	0.6452
D5 — LSAT, k=300 sex, race, 4 groups	SeqPAV scored weighted	0.9979	0.9998	**0.9977**
	SeqPAV default weighted	0.9977	0.9999	0.9976
	dice roll p_{stat}	0.9972	0.9992	0.9964
	CFAθ	0.9991	0.9964	0.9955
	multi FA*IR p_{stat}	0.9947	0.9807	0.9755
	colorblind	1	0.9822	0.9822
	SeqPAV scored equal	0.9804	0.9405	0.9221
	SeqPAV default equal	0.979	0.935	0.9154
	dice roll p_{eq}	0.9736	0.9116	0.8875
	multi FA*IR p_{eq}	0.9793	0.9013	0.8827

Table 4.6: Results for multiple attribute datasets (D4 and D5) and methods on fairness, relevance, and aggregate metrics. The best-performing method (non-colorblind) on the aggregate fairness-relevance metric is bolded. Comparable results are grouped together.

Dataset	**Method**	**NDCG**	**AWRF**	**NDCG*AWRF**
D6 — CSRankings, k=16 size, area, 10 groups	SeqPAV scored weighted	0.9746	0.8942	**0.8715**
	multi FA*IR p_{stat}	0.9689	0.8783	0.851
	SeqPAV default weighted	0.8371	0.8952	0.7494
	SeqPAV scored equal	0.9652	0.8961	0.865
	multi FA*IR p_{eq}	0.9316	0.9247	0.8615
	SeqPAV default equal	0.836	0.9791	0.8186
	colorblind	1	0.807	0.807
D7 — MEPS, k=20 race, age, 16 groups	SeqPAV scored weighted	0.9118	0.8566	0.781
	multi FA*IR p_{stat}	0.9357	0.816	0.7635
	SeqPAV default weighted	0.8048	0.9253	0.7447
	SeqPAV scored equal	0.8639	0.9055	**0.7823**
	multi FA*IR p_{eq}	0.8433	0.8761	0.7388
	colorblind	1	0.718	0.718
	SeqPAV default equal	0.6836	0.9738	0.6657

Table 4.7: Results for multiple attribute datasets (D6 and D7) and methods on fairness, relevance, and aggregate metrics. The best-performing method (non-colorblind) on the aggregate fairness-relevance metric is bolded. Comparable results are grouped together.

Chapter 5

Conclusion

Proportional representation fairness emphasizes that sizable groups of individuals sharing similar preferences should be adequately represented. It serves as a crucial objective that should be embraced in various election or voting scenarios to ensure fair outcomes in any democratic setting, regardless of its scale. This thesis works within the diverse definitions and frameworks of proportionality in multi-winner elections. Our primary objective is to examine the influence of deliberative democracy on voting and develop inclusive deliberation strategies that enhance fairness in election outcomes while preserving simplicity in the overall process. Subsequently, we broaden our scope to leverage the robust fairness guarantees offered by intricate proportional voting rules and explore their potential as fair-ranking algorithms. In this section, we present a summary of our findings, acknowledge the limitations of our work, and discuss potential avenues for impact and further research.

5.1 Deliberation and Voting

5.1.1 Summary of Results

In Chapter 3, we empirically study the impact of deliberation on voting outcomes, specifically in approval-based multi-winner (or committee) elections. We design an abstract agent-based model of deliberation to simulate several deliberation scenarios involving different voter types. Using this model, we explore several simple, heuristic-driven group deliberation strategies and study the performance of various standard voting rules under these deliberation methods. Our findings reveal that deliberation generally enhances voting outcomes in terms of welfare, representation, and proportionality assurances. Deliberation mechanisms designed to promote exposure to diverse groups and opinions elevate the quality of deliberation, protect minority preferences, and consequently achieve better voting outcomes. Additionally, we demonstrate that, in the presence of effective deliberation, "simpler" voting rules like classical Approval Voting (AV) can achieve proportionality guarantees comparable to those promised by "complex" voting rules in the absence of deliberation. Thus, by placing greater emphasis on deliberation, we can achieve higher welfare and representation while maintaining a straightforward and transparent voting procedure.

5.1.2 Limitations & Future Work

This work gave us valuable insights into the significance of deliberation in participatory democracy and the key factors contributing to the design of effective deliberation mechanisms capable of facilitating favorable voting outcomes. However, we must acknowledge the limitations of our study and highlight the areas that require further investigation.

First, our work measures proportional representation outcomes solely through popular qualitative, axiomatic guarantees. However, due to the binary nature of these axioms, which categorize voting rules as either satisfying or failing the proportionality conditions, it becomes challenging to gauge the degree to which a voting rule may yield proportional outcomes without consistently meeting the axiomatic criteria. Investigating the extent of proportionality provided by these voting rules from a quantitative perspective is an interesting direction for future research [69]. Second, although bounded confidence models offer a suitable representation of small-scale deliberation by incorporating dynamic agent types, it is valuable to consider alternative opinion dynamics models. Conducting empirical comparisons among these models to determine which ones effectively capture real-world deliberation dynamics would enhance our analysis of the influence of deliberation on voting outcomes. Barrett et al. [8] present a comparison of various opinion dynamics models concerning the attainment of consensus in citizens' assemblies. It would be intriguing to leverage the insights from their work to develop more robust deliberation models and explore their impact on voting outcomes. In addition to empirical work, it is essential to carry out real-user studies to validate whether the positive effects of deliberation and our proposed deliberation strategies yield similar outcomes when implemented with human participants. Additionally, we only consider simple, heuristic-based group division strategies for deliberation mechanisms. While it might come at the cost of increased complexity in the democratic process, experimenting with more advanced group division mechanisms could provide interesting insights. Barrett et al. explore this idea in citizens' assemblies where repeated interactions between participants lead to diminishing returns [7].

In our experiments, we split the voting population into two sets – a majority and a minority, to experimentally verify how favorable the voting outcomes are towards each population set. Since this is an initial step in the direction of fully understanding the impact of deliberation on voting, we restrict our analysis to a simple scenario of just two group types. Incorporating multiple minority/majority groups would increase complexity, thereby making it difficult to clearly understand the impact of deliberation on voting outcomes. However, our model is designed to accommodate different population splits and it would be interesting to explore this in future work. Our intuition is that since our model is designed to provide proportional results, it would scale proportionally to multiple minority groups. Moreover, there is potential for extending our research to other dimensions of democratic processes. Exploring diverse ballot types or investigating election domains like ranked-choice voting and participatory budgeting present intriguing avenues for future research. Finally, throughout Chapter 3, we consistently emphasize the importance of promoting simplicity and explainability in democratic processes to ensure citizen participation and trust. However, there is a noticeable absence of well-defined, quantifiable metrics to assess these aspects. Consequently, it is imperative to conduct a systematic examination of the explainability of standard voting rules in order to draw more concrete conclusions in this regard.

5.1.3 Broader Impact

The findings presented in this chapter contribute to the overarching objective of leveraging computer science techniques to create improved, inclusive, fair, and user-friendly participatory democracy platforms. The advancement of computer science research in this direction can greatly support the endeavors of practitioners aiming to enhance democratic institutions for the future. There is substantial potential for delivering algorithmic expertise, both from theoretical and experimental standpoints. The diverse tasks within the democratic process necessitate different types and degrees of algorithms, depending on the desired levels of complexity, explainability, and simplicity. By comprehending the specific requirements and continuously refining and implementing these algorithms, practitioners can make significant strides forward.

As highlighted earlier in the thesis, traditionally, the focus of social choice literature has been the aggregation of preferences. However, the results observed in this thesis and other related work have demonstrated the remarkable effectiveness of deliberative democracy in producing favorable election results. Considering deliberation as the pillar of success, further research should aim to develop better models of deliberation tailored to both small- and large-scale democratic processes. These models and mechanisms should be designed with the overarching objectives of fostering broad societal participation and facilitating meaningful conversations. Moreover, these methods hold value in wider research domains, including deliberative alignment. The core concept behind deliberative alignment is to drive progress in governance, diplomacy, technology, and various other fields by leveraging the collective opinions and desires of society as a whole. By employing the methods discussed, advancements can be made in these areas, fostering a more inclusive and democratic approach to decision-making and problem-solving. Since language is at the core of deliberation, some recent research has already investigated the potential of using large language models in scaling deliberation through effective moderation [73]. Similarly, Polis, a real-time system designed and maintained by the Computational Democracy Project has been used for facilitating and scaling deliberation forums [72]. These examples exemplify the ongoing efforts aimed at leveraging technology to construct improved democratic systems, while simultaneously employing democratic principles to enhance technological advancements. We aspire that our work can contribute to the development of these platforms and aid in the construction of better democracies for the future.

5.2 Fair Ranking through Proportional Voting

5.2.1 Summary of Results

In Chapter 4, we explore the use of proportional voting rules as fair-ranking algorithms. We reframe the fair-ranking problem as a voting task and demonstrate that committee voting rules designed for proportional outcomes in elections exhibit robust performance when applied to general fair-ranking tasks. We introduce four modified versions of SeqPAV, which explicitly consider both relevance and fairness requirements. Through extensive experimentation on diverse datasets, encompassing scenarios with single or multiple sensitive attributes, we consistently observe that algorithms based on SeqPAV deliver impressive results in terms of the aggregate fairness-relevance metric. In fact, they often match or even surpass the performance of FA*IR and other baseline methods.

5.2.2 Limitations & Future Work

While this chapter establishes a potential new application domain for proportional voting rules, we must be aware of the shortcomings or limitations of our experimental setup. First, we note that the scores achieved on most datasets for all methods are considerably high, and the performance advantage of SeqPAV-based ranking methods over baselines is relatively modest. For the purpose of this thesis, we limit our analysis to these datasets to enable a direct comparison with the baselines. Although consistent outcomes are observed across datasets, it is essential to augment our findings by conducting experiments on more demanding datasets. Such experiments would provide insights into the scalability of our SeqPAV-based ranking methods when confronted with more challenging ranking tasks. Moreover, we emphasize the necessity for a thorough examination and comparison with a broader range of baseline methods, including those that consider ranking policies rather than solely focusing on individual ranked lists.

In our experiments, we only implement simple modifications to the SeqPAV rule for developing ranking algorithms. Although the results have shown promise, future research could explore more nuanced modifications based on the foundational principles and properties of proportional voting methods. Finally, an intriguing avenue for future work involves expanding beyond post-processing fairness and reranking techniques. It would be valuable to explore the integration of these methods with retrieval techniques, considering a holistic approach that encompasses both stages of the ranking process.

5.2.3 Broader Impact

This chapter emphasizes the advantageous axiomatic fairness guarantees of proportional voting rules and discusses the possibility of using these rules outside of election domains, presenting fair ranking as a use case. The importance of algorithmic fairness has significantly increased, becoming essential in various applications involving algorithmic decision-making. The broader scope of this work includes the application of voting algorithms outside of election domains and the further advancement of these voting mechanisms for diverse purposes. As demonstrated in Chapter 4, the guarantees and flexibility offered by voting methods could be highly valuable for regulatory agencies in addressing a wide array of fairness-specific tasks. Further research in this direction would bridge the gap between proportionality and fairness, leading to a shared mathematical framework that can be applied to diverse problem domains.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. 2016.

[2] Haris Aziz, Markus Brill, Vincent Conitzer, Edith Elkind, Rupert Freeman, and Toby Walsh. Justified representation in approval-based committee voting. *Social Choice and Welfare*, 48(2):461–485, 2017.

[3] Haris Aziz, Edith Elkind, Shenwei Huang, Martin Lackner, Luis Sánchez-Fernández, and Piotr Skowron. On the complexity of extended and proportional justified representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[4] Haris Aziz, Serge Gaspers, Joachim Gudmundsson, Simon Mackenzie, Nicholas Mattei, and Toby Walsh. Computational aspects of multi-winner approval voting. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[5] Haris Aziz and Nisarg Shah. Participatory budgeting: Models and approaches. In *Pathways Between Social Science and Computational Social Science*, pages 215–236. Springer, 2021.

[6] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California law review*, pages 671–732, 2016.

[7] Jake Barrett, Kobi Gal, Paul Gölz, Rose M Hong, and Ariel D Procaccia. Now we're talking: Better deliberation groups through submodular optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5490–5498, 2023.

[8] Jake Barrett, Kobi Gal, and Loizos Michael. Beyond the echo chamber: Modelling opinion changes in citizens' assemblies. https://sites.google.com/view/cmas23/schedule, 2023. Accessed: 2023-05-30.

[9] Jan Behrens, Axel Kistner, Andreas Nitsche, and Björn Swierczek. *The principles of LiquidFeedback*. Interacktive Demokratie, 2014.

[10] Gerdus Benadè, Nevo Itzhak, Nisarg Shah, Ariel D. Procaccia, and Ya'akov (Kobi) Gal. Efficiency and usability of participatory budgeting methods. Working Paper, 2018.

[11] Gerdus Benadè, Swaprava Nath, Ariel D Procaccia, and Nisarg Shah. Preference elicitation for participatory budgeting. *Management Science*, 67(5):2813–2827, 2021.

[12] Steven Brams and Peter C Fishburn. *Approval voting*. Springer Science & Business Media, 2007.

[13] Robert Bredereck, Piotr Faliszewski, Andrzej Kaczmarczyk, and Rolf Niedermeier. An experimental view on committees providing justified representation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, page 109–115. AAAI Press, 2019.

[14] Markus Brill, Rupert Freeman, Svante Janson, and Martin Lackner. Phragmén's voting methods and justified representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[15] B Buchanan. Good-enough golfers. https://github.com/islemaster/good-enough-golfers, 2017.

[16] Robin Burke, Nicholas Mattei, Vladislav Grozin, Amy Voida, and Nasim Sonboli. Multi-agent social choice for dynamic fairness-aware recommendation. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 234–244, 2022.

[17] Robin D Burke, Himan Abdollahpouri, Bamshad Mobasher, and Trinadh Gupta. Towards multi-stakeholder utility evaluation of recommender systems. *UMAP (Extended Proceedings)*, 750, 2016.

[18] Yves Cabannes. Participatory budgeting: a significant contribution to participatory democracy. *Environment and Urbanization*, 16(1):27–46, 2004.

[19] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Rev. Mod. Phys.*, 81:591–646, May 2009.

[20] John R Chamberlin and Paul N Courant. Representative deliberations and representative decisions: Proportional representation and the borda rule. *American Political Science Review*, 77(3):718–733, 1983.

[21] Hun Chung and John Duggan. A formal theory of democratic deliberation. *American Political Science Review*, 114(1):14–35, 2020.

[22] Morris H. Degroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

[23] John S Dryzek, André Bächtiger, Simone Chambers, Joshua Cohen, James N Druckman, Andrea Felicetti, James S Fishkin, David M Farrell, Archon Fung, Amy Gutmann, et al. The crisis of democracy and the science of deliberation. *Science*, 363(6432):1144–1146, 2019.

[24] John S. Dryzek and Christian List. Social choice theory and deliberative democracy: A reconciliation. *British Journal of Political Science*, 33(1):1–28, 2003.

[25] Michael D. Ekstrand, Graham McDonald, Amifa Raj, and Isaac Johnson. Overview of the trec 2022 fair ranking track. *arXiv preprint arXiv: Arxiv-2302.05558*, 2023.

[26] Edith Elkind, Piotr Faliszewski, Piotr Skowron, and Arkadii Slinko. Properties of multiwinner voting rules. *Social Choice and Welfare*, 48(3):599–632, 2017.

[27] Edith Elkind, Davide Grossi, Ehud Shapiro, and Nimrod Talmon. United for change: deliberative coalition formation to change the status quo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5339–5346, 2021.

[28] Stephen Elstub and Oliver Escobar. *Handbook of democratic innovation and governance*. Edward Elgar Publishing, 2019.

[29] Stephen Elstub, Oliver Escobar, Ailsa Henderson, Tamara Thorne, Nick Bland, and Evelyn Bowes. *Citizens' Assembly of Scotland: Research Report*. Scottish Government Social Research, January 2022.

[30] Brandon Fain, Ashish Goel, Kamesh Munagala, and Sukolsak Sakshuwong. Sequential deliberation for social choice. In *International Conference on Web and Internet Economics*, pages 177–190. Springer, 2017.

[31] Brandon Fain, Kamesh Munagala, and Nisarg Shah. Fair allocation of indivisible public goods. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, EC '18, page 575–592, New York, NY, USA, 2018. Association for Computing Machinery.

[32] Roy Fairstein, Dan Vilenchik, Reshef Meir, and Kobi Gal. Welfare vs. representation in participatory budgeting. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, page 409–417, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems.

[33] Piotr Faliszewski, Piotr Skowron, Arkadii Slinko, and Nimrod Talmon. Multiwinner voting: A new challenge for social choice theory. *Trends in Computational Social Choice*, 74(2017):27–47, 2017.

[34] Reza Zanjirani Farahani and Masoud Hekmatfar. *Facility location: concepts, models, algorithms and case studies.* Springer Science & Business Media, 2009.

[35] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[36] James Fishkin. *When the people speak: Deliberative democracy and public consultation.* Oxford University Press, 2009.

[37] Sergiu Gherghina, Monika Mokre, and Sergiu Miscoiu. Introduction: Democratic deliberation and under-represented groups. *Political Studies Review*, 19(2):159–163, 2021.

[38] Ashish Goel and David T Lee. Towards large-scale deliberative decision-making: Small groups and the importance of triads. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 287–303, 2016.

[39] Paul Gölz. *Social Choice for Social Good: Proposals for Democratic Innovation from Computer Science.* PhD thesis, Carnegie Mellon University, 2022.

[40] Jürgen Habermas. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy.* Polity, 1996.

[41] Xiao Han, Leye Wang, Soochang Park, Angel Cuevas, and Noël Crespi. Alike people, alike interests? a large-scale study on interest similarity in social networks. In *2014*

IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), pages 491–496, 2014.

[42] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[43] Warwick Harvey. CSPLib problem 010: Social golfers problem. http://www.csplib.org/Problems/prob010, 2002. Accessed: 2022-12-14.

[44] Rainer Hegselmann and Ulrich Krause. Opinion Dynamics and Bounded Confidence Models, Analysis and Simulation. *Journal of Artificial Societies and Social Simulation*, 5(3):1–2, 2002.

[45] Corinna Hertweck, Christoph Heitz, and Michele Loi. On the moral justification of statistical parity. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 747–757, 2021.

[46] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[47] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[48] Martin Lackner, Peter Regner, Benjamin Krenn, Elvi Cela, Jonas Kompauer, Florian Lackner, Stanisław Szufa, and Stefan Schlomo Forster. abcvoting: A Python library of approval-based committee voting rules, 2021. Current version: https://github.com/martinlackner/abcvoting.

[49] Martin Lackner and Piotr Skowron. Consistent approval-based multi-winner rules. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 47–48, 2018.

[50] Martin Lackner and Piotr Skowron. Utilitarian welfare and representation guarantees of approval-based multiwinner rules. *Artificial Intelligence*, 288:103366, 2020.

[51] Annick Laruelle. Voting to select projects in participatory budgeting. *European Journal of Operational Research*, 288(2):598–604, January 2021.

[52] Ke Liu, Sven Löffler, and Petra Hofstedt. Social golfer problem revisited. In Jaap van den Herik, Ana Paula Rocha, and Luc Steels, editors, *Agents and Artificial Intelligence*, pages 72–99, Cham, 2019. Springer International Publishing.

[53] Jan Lorenz. Continuous opinion dynamics under bounded confidence: A survey. *International Journal of Modern Physics C*, 18(12):1819–1838, 2007.

[54] D M Mackie, L T Worth, and A G Asuncion. Processing of persuasive in-group messages. *J Pers Soc Psychol*, 58(5):812–822, May 1990.

[55] C. L. Mallows. Non-null Ranking Models. I. *Biometrika*, 44(1-2):114–130, 06 1957.

[56] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[57] Kanav Mehra, Nanda Kishore Sreenivas, and Kate Larson. Deliberation and voting in approval-based multi-winner elections. *arXiv preprint arXiv: Arxiv-2305.08970*, 2023.

[58] Ismael Peña-López et al. Innovative citizen participation and new democratic institutions: Catching the deliberative wave. 2020.

[59] Juan Perote-Peña and Ashley Piggins. A model of deliberative and aggregative democracy. *Economics & Philosophy*, 31(1):93–121, 2015.

[60] Dominik Peters, Grzegorz Pierczyński, and Piotr Skowron. Proportional participatory budgeting with additive utilities. *Advances in Neural Information Processing Systems*, 34:12726–12737, 2021.

[61] Dominik Peters and Piotr Skowron. Proportionality and the limits of welfarism. In *Proceedings of the 21st ACM Conference on Economics and Computation*, EC '20, page 793–794, New York, NY, USA, 2020. Association for Computing Machinery. Extended version at https://arxiv.org/abs/1911.11747.

[62] Ariel D Procaccia and Jeffrey S Rosenschein. The distortion of cardinal preferences in voting. In *International Workshop on Cooperative Information Agents*, pages 317–331. Springer, 2006.

[63] Soroush Rafiee Rad and Olivier Roy. Deliberation, single-peakedness, and coherent aggregation. *American Political Science Review*, 115(2):629–648, 2021.

[64] Amifa Raj and Michael D Ekstrand. Measuring fairness in ranked results: An analytical and empirical comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 726–736, 2022.

[65] Luis Sánchez-Fernández, Edith Elkind, Martin Lackner, Norberto Fernández, Jesús Fisteus, Pablo Basanta Val, and Piotr Skowron. Proportional justified representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[66] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. Quantifying the impact of user attentionon fair group representation in ranked lists. In *Companion proceedings of the 2019 world wide web conference*, pages 553–562, 2019.

[67] Nisarg Shah. Reverting to simplicity in social choice. In *The Future of Economic Design*, pages 39–44. Springer, 2019.

[68] Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2219–2228, 2018.

[69] Piotr Skowron. Proportionality degree of multiwinner rules. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 820–840, 2021.

[70] Piotr Skowron, Piotr Faliszewski, and Jérôme Lang. Finding a collective set of items: From proportional multirepresentation to group recommendation. *Artificial Intelligence*, 241:191–216, 2016.

[71] Piotr Skowron, Martin Lackner, Markus Brill, Dominik Peters, and Edith Elkind. Proportional rankings. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 409–415, 2017.

[72] Christopher Small, Michael Bjorkegren, Timo Erkkilä, Lynette Shaw, and Colin Megill. Polis: Scaling deliberation by mapping high dimensional opinion spaces. *Recerca: revista de pensament i anàlisi*, 26(2), 2021.

[73] Christopher T. Small, Ivan Vendrov, Esin Durmus, Hadjar Homaei, Elizabeth Barry, Julien Cornebise, Ted Suzman, Deep Ganguli, and Colin Megill. Opportunities and risks of llms for scalable deliberation with polis. *arXiv preprint arXiv: 2306.11932*, 2023.

[74] Nasim Sonboli, Robin Burke, Michael Ekstrand, and Rishabh Mehrotra. The multi-sided complexity of fairness in recommender systems. *AI magazine*, 43(2):164–176, 2022.

[75] Dariusz Stolicki, Stanisław Szufa, and Nimrod Talmon. Pabulib: A participatory budgeting library. *arXiv preprint arXiv:2012.06539*, 2020.

[76] Thorvald N Thiele. Om flerfoldsvalg. *Oversigt over det Kongelige Danske Videnskabernes Selskabs Forhandlinger*, 1895:415–441, 1895.

[77] Philipp Verpoort. GroupSelect by Sortition Foundation. https://github.com/sortitionfoundation/groupselect-app, 2020. Accessed: 2022-12-12.

[78] Gérard Weisbuch, Guillaume Deffuant, Frédéric Amblard, and Jean-Pierre Nadal. Meet, discuss, and segregate! *Complexity*, 7(3):55–63, 2002.

[79] Linda F Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

[80] D A. Wilder. Some determinants of the persuasive power of in-groups and out-groups: Organization of information and attribution of independence. *Journal of Personality and Social Psychology*, 59(6):1202–1213, 1990.

[81] Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. Balanced ranking with diversity constraints. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6035–6042, 2019.

[82] Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th international conference on scientific and statistical database management*, pages 1–6, 2017.

[83] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578, 2017.

[84] Meike Zehlike, Philipp Hacker, and Emil Wiedemann. Matching code and law: achieving algorithmic fairness with optimal transport. *Data Mining and Knowledge Discovery*, 34(1):163–200, 2020.

[85] Meike Zehlike, Tom Sühr, Ricardo Baeza-Yates, Francesco Bonchi, Carlos Castillo, and Sara Hajian. Fair top-k ranking with multiple protected groups. *Information processing & management*, 59(1):102707, 2022.

[86] Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking: A survey. *arXiv preprint arXiv: Arxiv-2103.14000*, 2021.

[87] Gil Ben Zvi, Eyal Leizerovich, and Nimrod Talmon. Iterative deliberation via metric aggregation. In *International Conference on Algorithmic Decision Theory*, pages 162–176. Springer, 2021.

APPENDICES

Appendix A

Iterative Golfer

Iterative golfer strategy is a weaker version of the popular social golfer problem [43, 52] in combinatorial optimization.

> *Social golfer problem*: n golfers must be repeatedly assigned to g groups of size s. Find the maximum number of rounds (and the corresponding schedule) such that no two golfers play in the same group more than once.

Social golfer problem maximizes the number of rounds with a hard constraint that no two golfers should meet again. The iterative golfer strategy is a weaker version of this where we fix the number of rounds R, and minimize the number of occurrences where any pair of agents meet more than once. Given some group assignment $G^r = \{G_1^r, G_2^r, \ldots, G_g^r\}$ at round r, we introduce a cost given by:

$$cost(G^r) = \sum_{G_x \in G^r} \sum_{a,b \in G_x} f^2(a, b) \tag{A.1}$$

where $f(a, b)$ is the number of times a and b have been in the same group in the previous rounds G^1 through G^{r-1}. The number of prior meetings is squared to ensure an even number of conflicts among all possible pairings (as opposed to one specific pair meeting repeatedly). We use an existing approximate solution [15] that creates group assignment for each round such that the cost given by (A.1) is minimized. The iterative golfer can thus be seen as a more efficient strategy than iterative random if the objective is to ensure each agent has the highest possible exposure to others' preferences.