PreFAIR: Combining Partial Preferences for Fair Consensus Decision-making

Kathleen Cachel
Worcester Polytechnic Institute
Worcester, MA, USA
kcachel@wpi.edu

Elke Rundensteiner
Worcester Polytechnic Institute
Worcester, MA, USA
rundenst@wpi.edu

ABSTRACT

Preference aggregation mechanisms help decision-makers combine diverse preference rankings produced by multiple voters into a single consensus ranking. Prior work has developed methods for aggregating multiple rankings into a fair consensus over the *same* set of candidates. Yet few real-world problems present themselves as such precisely formulated aggregation tasks with each voter fully ranking all candidates. Instead, preferences are often expressed as rankings over partial and even disjoint subsets of candidates. For instance, hiring committee members typically opt to rank their top choices instead of exhaustively ordering every single job applicant. However, the existing literature does not offer a framework for characterizing nor ensuring group fairness in such partial preference aggregation tasks. Unlike fully ranked settings, partial preferences imply both a *selection decision* of whom to rank plus an *ordering decision* of how to rank the selected candidates. Our work fills this gap by conceptualizing the open problem of fair partial preference aggregation. We introduce an impossibility result for fair selection from partial preferences and design a computational framework showing how we can navigate this obstacle. Inspired by Single Transferable Voting, our proposed solution PreFair produces consensus rankings that are fair in the selection of candidates and also in their relative ordering. Our experimental study demonstrates that PreFair achieves the best performance in this dual fairness objective compared to state-of-the-art alternatives adapted to this new problem while still satisfying voter preferences.

CCS CONCEPTS

• Computing methodologies; **• Social and professional topics → User characteristics**;

KEYWORDS

Preference Aggregation, Fair Consensus Decision-making, Group Fairness, Algorithmic Fairness

ACM Reference Format:
Kathleen Cachel and Elke Rundensteiner. 2024. PreFAIR: Combining Partial Preferences for Fair Consensus Decision-making. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3630106.3658961

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658961

1 INTRODUCTION

Preference data is ubiquitous in our world: employees evaluate job applicants, reviewers rate funding proposals, admission committees rank applicants, and so forth. These decisions are often collective, meaning many voters (reviewers, committee members, etc.) express their preferences about candidates, and these preferences are then combined into a representative decision attempting to satisfy all voters. The task of combining voter preference rankings into a single consensus ranking is referred to as preference aggregation [9]. Unfortunately, recent work has shown that bias towards marginalized candidate groups, such as gender or race, in individual voter preferences gets aggregated into the final consensus ranking [39]. As preference aggregation is performed in a number of contexts affecting people's livelihoods [17, 28, 37], it is essential to design aggregation mechanisms that mitigate discriminatory bias in consensus decision-making. Various strategies have been proposed for group fair preference aggregation, e.g., methods for producing consensus rankings fairly ordering candidate groups [11, 12, 39, 60]. However, these approaches overlook a critical context that is prevalent in the real world, namely, *partial preferences.*

In real-world preference aggregation tasks, due to time constraints and the cognitive load of ranking potentially hundreds of candidates [41], voters often provide only *partial preference rankings.* Specifically, each voter ranks only a subset of the candidates (often their top choices), and the final consensus ranking orders only k candidates. Consequently, it is critical to develop *fair partial preference aggregation* systems that consider the unique challenges of this preference setting. To date, however, all fairness-enhanced preference aggregation mechanisms assume that all voters provide preferences ranking all considered candidates [11, 12, 39, 60] and thus characterize fairness as a one-sided concern. They do not address the challenges inherent to the partial preference setting. The foremost challenge, as illustrated in Figure 1, is that *fairness in partial preference contexts is two dimensional.* Specifically, aggregation mechanisms must consider which k out of m candidates they rank in the first place (selection fairness) and then how they rank these k selected candidates (ordering fairness). This challenge is derived from the fact that in partial preference settings, voters may rank potentially disjoint subsets of the larger candidate pool. For this reason, candidate groups may be under-selected or entirely absent in a voter's preferences. This, in turn, impacts the final consensus. Fair preference aggregation methods, however, inherently do not address *selection bias* in voters' choices of whom they choose to rank – as these methods assume both voters and the final consensus ranking order all the candidates [11, 12, 39, 60].

The second challenge is that *both fairness objectives must be addressed during aggregation.* Simple interventions using fair ranking

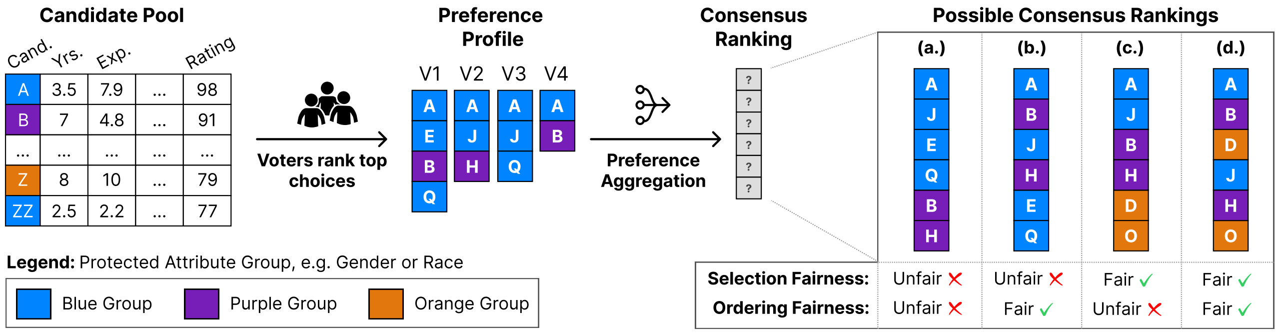


Figure 1: Fair partial preference aggregation problem. The goal is to combine voter preference rankings into a consensus ranking of k candidates that is both a fair selection with respect to the larger candidate pool and a fair ordering of the final candidate groups being ranked, e.g., consensus ranking (d.). The result must incorporate as many voter preferences as possible to maintain voter satisfaction.

methods [14, 27, 31, 63] to pre- or post-process the input or output of aggregation do not address both fairness concerns. Re-ranking a biased selection will not mitigate selection bias. Thus, using a traditional partial preference aggregation mechanism [9, 25, 57] and then either fairness re-ranking its consensus or the voter preferences prior to using the mechanism will not ensure the consensus is a fair selection. The third challenge is that *fairness and voter satisfaction are often conflicting objectives.* Conventional approaches prioritize the latter, without considering the former [9, 25, 57]. Yet, fair solutions, to be practical, must continue to appropriately capture voter preferences in the consensus. This is exacerbated by the fact that complete voter preferences are unobservable and thus unavailable.

Addressing the above challenges, our work formally characterizes the fair partial preference aggregation problem. We design the PreFair methodology for solving this open problem. PreFair takes as input both partial voter preference rankings and features (attributes) of candidates (including group membership) and outputs a consensus ranking of k candidates that is both a fair ordering and selection of the larger candidate pool. Advancing beyond prior methods [11, 12, 39, 60], PreFair consensus rankings satisfy the unique fairness concerns inherent to partial preferences while maximizing the satisfaction of possibly conflicting voter preferences. Our primary contributions are as follows.

- We define the new fair partial preference aggregation problem. We provide the first conceptualization of modern fairness notions in the task of aggregating partial voter preference rankings.
- We introduce an impossibility result for fair selection from partial preferences and design a computational framework showing how we can navigate this obstacle.
- We utilize this framework in PreFair, our proposed solution to addressing fairness concerns in partial preferences during aggregation. Inspired by Single Transferable Voting, PreFair's novel aggregation mechanism produces consensus rankings that are fair in the selection of k candidates and also in their relative ordering.
- We demonstrate experimentally that PreFair outperforms state-of-the-art techniques [9, 12, 25, 57, 60] in the fair partial preference aggregation task on four real-world datasets. We also utilize controlled simulations to reveal how the degrees of "partial-ness" in voter preference rankings contribute to downstream consensus ranking unfairness. We observe that PreFair successfully handles diverse preference ranking conditions.

2 RELATED WORK

Traditional Partial Preference Aggregation. To the best of our knowledge, our work is the first to study fairness for marginalized groups of candidates when combining partial voter preference rankings. While other works have proposed aggregation methods specifically for partial voter preference rankings, both in the information retrieval [1, 3, 5, 18, 25, 29, 42, 45, 52, 62] and social choice literature [4, 6, 7, 22, 58], none of them consider fairness for candidate groups. These methods can be summarized as the Single Transferable Voting family [4, 6, 7], score-based algorithms [1, 3, 5, 18, 29, 42, 45, 62], and Markov chain based algorithms [25, 52]. Additionally, work by Chakraborty et al. [15] and Aird et al. [2] explores utilizing these aggregation rules, as is, as part of larger fair recommender systems.

Fairness Related to Rankings. Fairness-enhanced ranking is an active research area. Broadly, this line of work encompasses several problem settings described below. However, these settings do not address the fairness concerns specific to combining *partial* voter preferences into a consensus ranking.

Ranking and Learning-to-Rank order candidates (e.g., items, objects, or alternate entities) by either using a set of given relevance scores or by learning relevance scores, respectively. Some works focus on producing a fair ordering of candidates via re-ranking [14, 20, 31, 53, 63] while other do so through the learning-to-rank process [8, 49, 54, 64]. However, unlike our work, this setting does not consider multiple rankings nor address the explicit aggregation of preferences. For a survey of fair ranking algorithms, see Patro et al. [46].

Multi-winner Voting constructs an *unranked set* of the most preferred candidates from voter preference rankings or yes/no approvals of candidates. Recent work has designed voting rules and integer programs [10, 13, 44] enforcing constraints on the number of candidates chosen per group. As experimentally shown in Section 5, fair multi-winner voting has two shortcomings when applied to our target problem. First, while this methodology ensures selection fairness, it is only guaranteed if voter preferences contain enough candidates from every group – which is not always the case in partial preferences. Second, akin to consensus ranking (c.) in Figure 1, this approach neglects the equally important goal of fairly ordering candidates.

Aggregating Full Preferences is, as mentioned earlier, the line of work most closely related to our problem. Our work departs from this literature [11, 12, 39, 60] by focusing on the distinct fairness concerns of partial voter preference rankings. Moreover, all prior fair preference aggregation work [11, 12, 39, 60] has modeled the task as fairness enhancing the Kemeny rule [62]. Kemeny is a distance-based method that determines the consensus ranking by minimizing the average Kendall tau distance between it and the voter preference rankings. Thus, it must assume that all rankings order an identical set of candidates. One key difference in our work is we do not rely on the Kemeny rule. Nonetheless, since the Epira method from Cachel and Rundensteiner [11] and Rapf from Wei et al. [60] aim to approximate the Kemeny rule, we are able to observe their performance on partial preferences in our experimental Section. As neither approach contains explicit mechanics to mitigate the selection bias introduced by partial voter preferences, we find that they underperform our proposed solution and generate results similar to the consensus ranking (b.) in Figure 1. Our method addresses this drawback by introducing a strategy for mitigating voter selection bias.

3 PROBLEM FORMULATION

Our setting considers a pool of m candidates $C = \{c_1, c_2, ..., c_m\}$, for instance, people who applied for a job, and n voters who prioritize their top choices among all the candidates by each providing a ranked list. This collection of n voter preference rankings, denoted by $R = \{r_1, ..., r_n\}$, is conventionally referred to as a preference profile [9]. We assume this preference profile contains partial orderings, that is, voters provide rankings of less than the full set of m candidates. For example, if a hiring committee has hundreds of applicants, committee members may only rank roughly their top-20 candidates each. Since voters only rank some of the candidates, many candidates may go completely unranked and do not appear in any of the voters' preference rankings. Consequently, the set of candidates appearing in the preference profile R, denoted as S, may be a much smaller subset of the entire candidate pool C. Our partial preference aggregation scenario does not assume each voter ranks the same number of candidates. Thus, by convention [9, 57], the final consensus ranking orders k-candidates, where $k \leq |S|$ is decided upon ahead of time.

Specific to our fair variant of partial preference aggregation, candidates have an associated categorical protected attribute, such as gender or race. The set of candidates in C that share a value (e.g., woman) in the protected attribute (e.g., gender) are referred to as a group g_i. We use $G = \{g_1, ..., g_v\}$ to denote the set of v groups in the candidate pool C. Additionally, we assume we have access to non-protected numeric features for each candidate in C, denoted as dataset $\mathcal{D}$. For instance, in the job applicant scenario, $\mathcal{D}$ might be attributes such as assessment scores, skill counts, and years of experience. It is apparent that such data is made available to rankers in applications from hiring, healthcare, and education, as without such information, it would be impossible to ask voters to provide their preferences.

Our *fair partial preference aggregation* problem is to produce a consensus ranking τ of k candidates that guarantees all groups are fairly represented while concurrently ensuring that groups are also fairly ordered in τ. We also want the consensus ranking to satisfy the voter preferences, which may have significant disagreements. In other words, we seek to produce a k-consensus ranking meeting two fairness objectives, namely, (1) a fair selection and (2) a fair ordering, while delivering as much voter satisfaction as possible. The *fair selection objective* ensures that all groups are *fairly represented* in the final consensus decision. While the *fair ordering objective* ensures the groups receive similar *favorable consensus ranking positions*. If only one of the two objectives is satisfied, the resulting consensus ranking may not be holistically fair to all candidate groups. These dual fairness objectives in our fair partial preference aggregation problem mitigate the harmful effects of both selection and ordering biases. In the next section, we formulate our formal definitions of the fair selection and the fair ordering criteria. Table 1 describes the notation used in this work.

4 OUR METHODOLOGY: PREFAIR

In this section, we introduce a framework for the fair partial preference aggregation task. We provide our formal fairness objectives in Section 4.1, and present PreFair, the first computational solution to this problem in Section 4.2.

4.1 Dual Fairness Objectives for Partial Preference Aggregation

As discussed in Section 3, there are two important potential sources of bias in the partial preference aggregation task. Namely, *selection bias*, whereby voters ignore or under-select certain groups in their stated preferences, and *ordering bias* whereby voters relegate certain groups to unfavorable preference ranking positions. Below, we define key fairness objectives for mitigating both types of bias.

4.1.1 Selection fairness in fair partial preference aggregation. Selection fairness is the first objective we formalize. In particular, our methodology supports two of the popular and standard conceptualizations of fair selection [13, 56].

First, *proportional representation* (also known as statistical or demographic parity [24, 47]) requires the consensus ranking to contain the same percentage of candidates from all groups G. Eq. 1 defines proportional representation for candidates with respect to their groups in consensus ranking τ.

$$|g_i \cap \text{candidates} \in \tau| = |\tau| \times \frac{|g_i \cap C|}{|C|}, \forall g_i \in G \quad (1)$$

Second, *equal representation* requires the consensus ranking to contain the same number of candidates from every group in G.

Table 1: Overview of commonly used notation.

Notation	Meaning
$C = \{c_1, ..., c_m\}$	The total set of m candidates, called a "pool'
$\mathcal{D}$	Dataset of d non-protected features $f_1, ..., f_d$ $\forall c_i \in C$
$R = \{r_1, ..., r_n\}$	Preference profile of n partial rankings of C
$S = \{c_1, ..., c_l\}$	Chosen candidates from C ranked in R
$G = \{g_1, ..., g_v\}$	Candidates belong to one of v disjoint groups
k	Desired count of candidates in consensus ranking
$\Theta = \{\theta_1, ..., \theta_v\}$	Per-group counts required for fair selection

Eq. 2 defines equal representation in terms of the candidates in consensus ranking τ.

$$|g_i \cap \text{candidates} \in \tau| = |\tau| \times |G|^{-1}, \forall g_i \in G \tag{2}$$

Selection fairness can be measured based on the observation that both of the above fair selection notions ultimately enforce a certain number of candidates per group to be in τ. We propose to measure selection fairness as *selection fairness divergence (SFD)*, with lower SFD values being more fair. Following group fairness metrics contrasting realized and ideal per-group values [31, 32, 59, 61], we formulate SFD using KL-divergence [40].

$$SFD(\tau, G_\tau) = d_{KL}(\mathbf{D}_\tau || \mathbf{D}) \text{ with } \mathbf{D} = \mathbf{D_p} \text{ (proportional per Eq. 1) or } \mathbf{D} = \mathbf{D_e} \text{ (equal per Eq. 2)} \tag{3}$$

where $d_{KL}(\mathbf{D}_\tau || \mathbf{D})$ is the KL-divergence score of $\mathbf{D}_\tau$, the proportions of each group in τ, and the desired proportion $\mathbf{D}$. For quantifying proportional representation per Eq. 1, $\mathbf{D}$ is $\mathbf{D_p} = [|g_1|\backslash|C|, ..., |g_v|\backslash|C|]$ and for quantifying equal representation per Eq. 2, $\mathbf{D}$ is $\mathbf{D_e} = [|G|^{-1}, ..., |G|^{-1}]_{1x|G|}$.

4.1.2 *Ordering fairness in fair partial preference aggregation.* From a rich line of fairness measurement in rankings [48, 51], we employ a representation-based fair ordering objective in our aggregation problem. In a nutshell, this ensures that no matter how many candidates there are, the resulting consensus ranking τ will have all groups represented in the first handful of positions. Specifically, we use the NDKL measurement approach introduced by Yang and Stoyanovich [61] and Geyik et al. [31]. Its main idea is that all groups be represented comparably throughout the prefixes of the consensus ranking τ, with more emphasis placed on higher prefixes of τ. We measure *ordering fairness* by NDKL, as in Eq. 4, checking that groups are equally represented, with lower values being more fair than higher values.

$$NDKL(\tau, G_\tau) = \frac{1}{Z} \sum_{i=|G_\tau|, 2*|G_\tau|, ...}^{|\tau|} \frac{1}{log_2(i+1)} d_{KL}(\mathbf{D}_{\tau_i} || \mathbf{D}_{\mathbf{e}\tau}) \tag{4}$$

where $d_{KL}(\mathbf{D}_{\tau_i} || \mathbf{D}_{\mathbf{e}\tau})$ corresponds to the KL-divergence score of the proportions of each group in the first i positions in τ, denoted by $\mathbf{D}_{\tau_i}$, and $\mathbf{D}_{\mathbf{e}\tau} = [|G_\tau|^{-1}, ..., |G_\tau|^{-1}]_{1x|G_\tau|}$. Then $Z = \sum_{i=|G_\tau|, 2*|G_\tau|, ...}^{|\tau|} \frac{1}{log_2(i+1)}$ [31]. Eq. 4, adopts the formulation from Yang and Stoyanovich [61], measuring representation over salient prefixes, specifically prefixes that correspond to the number of groups ranked in τ, i.e., $|G_\tau|$. Eq. 4 measures ordering fairness with respect to the *groups ordered in ranking* τ. Thus, if a group from candidate pool C was not included in τ, then NDKL would not surface this unfairness. This allows us to disambiguate performance on two fairness objectives in a targeted fashion, namely, selection and ranking.

In our experiments, we use a second variation of NDKL, which we call pool NDKL (pNDKL). It measures both selection and ordering fairness by utilizing all groups in candidate pool C, denoted by G, instead of G_τ (groups in τ).

$$pNDKL(\tau, G) = \frac{1}{Z} \sum_{i=|G_\tau|, 2*|G_\tau|, ...}^{|\tau|} \frac{1}{log_2(i+1)} d_{KL}(\mathbf{D}_{\tau_i} || \mathbf{D_e}) \tag{5}$$

Our PreFair methodology strives for lower NDKL and pNDKL values per Eq. 4 and Eq. 5 respectively.

4.2 The PreFair Method

We now describe PreFair, which features two main steps Configuration and Aggregation as illustrated in Algorithm 1.

4.2.1 *PreFair: Configuration step sets the stage for fairness.* The Configuration step makes it possible for the generated consensus ranking τ to satisfy selection fairness, regardless of how severe voter selection bias is in the given preference profile. The Configuration step, depicted in blue in Algorithm 1, first calibrates how many candidates per group need to be in the consensus ranking. The user's desired notion of fair selection, $\mathcal{F}$ (equal or proportional representation), is translated into the necessary per-group counts, denoted as Θ. Θ is determined from the desired consensus length k and Eq. 1 or Eq. 2, respectively. At this point, we observe that if certain groups are under-represented in the preference profile, such that we cannot meet the per-group candidate counts Θ we cannot create a fair consensus ranking from the preference profile R. This implies a fair consensus ranking is unattainable from the voter preferences at hand. Formally, proposition 4.1 establishes a fair selection impossibility result, the proof of which can be found in Appendix A.

Proposition 4.1 (Impossibility of Fair Selection from Partial Preferences). *If for any group $g_i \in G$, $|S \cap g_i| < \theta_{g_i}$ then any τ of length k produced from the preference profile R cannot satisfy selection fairness relative to candidate pool C.*

To move past the challenge presented by this result, we propose the conceptual idea of *pulling up* additional candidates. In other words, including candidates that were not explicitly ranked by voters. This ensures PreFair guarantees the constructed consensus ranking τ orders a fair selection of candidates. Specifically, in

PreFair, we devise a simple and customizable scheme for pulling up additional candidates when required.

The high-level idea of PreFair's pulling up approach is to take each voter's provided (partial) preference ranking and augment it into an extended ranking that now also includes some candidates that they had not explicitly ranked. To do this, we assume that voters' unstated preferences, i.e., the ordering of candidates they did not spend time ranking, resembles the same preference scheme that they had utilized for ordering the candidates in their provided preferences. Then we can apply similarity measures to the non-protected features $\mathcal{D}$ of candidates to infer how each voter would have ranked their unranked candidates.

More precisely, if S does not contain enough candidates per group to satisfy Θ PreFair does the following. To preference ranking r_i provided by each voter (i.e., $\forall\, r_i \in R$) PreFair appends a second ranking of the remaining candidates in C that were not ranked by the respective voter in r_i. To do so, we compute the centroid μ_i of the non-protected features of the candidates ranked in r_i. We construct the second ranking of the remaining candidates by ordering them by decreasing similarity between their (non-protected) features $\mathcal{D}$ and the previously computed centroid μ_i. In our experiments, we use unweighted cosine similarity, which, in additional experiments in Appendix E, we show works well for this task, but our software implementation can support alternate distance functions. Moreover, this step could be further personalized by asking voters to provide weights indicating the relative importance of the features associated with candidates.

An advantage of this pulling-up approach is that it infers preferences per voter, allowing human-in-the-loop interaction where voters can adjust these inferred preferences to their liking. Additionally, this strategy does not require complex architectures such as training preference learners [19]. Nonetheless, there are many other possible design choices for pulling up candidates. They may include requiring different inputs such as historical data or making use of potentially unstructured data such as candidate resumes or applications. This is an interesting area for future research that is beyond the scope of this paper.

4.2.2 *PreFair: Aggregation Step using a novel Group-Aware STV mechanism.*

While the above Configuration step makes it possible for PreFair to produce a fair selection of candidates, the second step of PreFair creates the fair consensus ranking τ. This Aggregation step, shown in orange in Algorithm 1, has two tasks corresponding to the dual fairness criteria of our problem. Its first task is to ensure the required number of candidates per group, denoted by Θ, are included in τ. Its second task is to ensure these candidates are ordered fairly.

The Aggregation component of PreFair is inspired by the Single Transferable Voting (STV) family [57] of preference aggregation methods. Single Transferable Voting is a desirable fair preference aggregation backbone since it operates on partial voter preference rankings. Note that PreFair only pulls up additional candidates when needed. At a high level, STV performs a series of round-based iterations until it determines the top k winners. First, a quota is calculated. Typically, this is the Droop quota [23], determined as $\lfloor \frac{n}{k+1} \rfloor + 1$. Then, in each iteration, candidates are "elected" if they have enough votes, either first-place votes or transferred votes to satisfy the quota. When as many candidates remain as there are open consensus spots, the remaining candidates are automatically elected. Transferred votes are accumulated when elected candidates have a surplus of votes on top of the Droop quota. That surplus is then split among subsequent preferences. Also, when no candidates are elected, the last placed candidate is eliminated from consideration, and votes are again transferred to subsequent preferences. The consensus ranking produced ranks candidates in the order in which they were elected. Appendix B visually illustrates traditional STV, along with providing more mechanistic details.

Algorithm 1 PreFair

Input: Preference profile R, dataset $\mathcal{D}$, candidate pool C with each candidate's group membership in G, fair selection objective $\mathcal{F}$ (`equal` or `proportional`), and consensus ranking length k.
Output: Consensus ranking τ fairly ordering k candidates that are a $\mathcal{F}$-fair selection from C.

```
Θ ← Eq. 1 if F == proportional or Θ ← Eq. 2 if F == equal          Configuration step
if Θ cannot be satisfied by candidates in R then // Need to pull up additional candidates
    for each voter's preference ranking r_i ∈ R do
        μ_i ← centroid of features in D ∀ candidate ranked in r_i   // Centroid of features for ranked candidates
        for each candidate c_j ∈ C ∉ r_i do
            cos_{c_j} ← (μ_i · D_{c_j}) / (‖μ_i‖‖D_{c_j}‖)
        updating R append to r_i all c_j ∈ C ∉ r_i ordered by decreasing cos_{c_j}   // Order by similarity to centroid
Create b ← ⌈∑Θ/|G|⌉ bins          Aggregation step
Set a per-group candidate count for each bin using Eq. 6
Assign candidates selected via Group-aware STV, as in Figure 2, to bins
τ ← flattens bins such that candidates are ordered by their arrival into the bin
```

As part of PreFair's Aggregation strategy, we propose a group-aware STV mechanism that both selects a fair representation of candidates and orders these candidates fairly. Unlike traditional STV, PreFair does not rank candidates strictly by their selection order. Instead, consensus ranking τ is produced by first creating "bins". We set the number of bins $b = \lceil \sum \Theta / |G| \rceil$. The number of candidates from a given group g that are included in a given bin i is:

$$\beta_g^i = |\theta_g| \div b + (1 \text{ if } i < (|\theta_g| \mod b) \text{ else } 0) \tag{6}$$

where θ_g is the required number of candidates for group g in τ. As the fair ordering criteria (Section 4.1.2) requires bins to have the same number of candidates per group, with more emphasis placed on higher bins, we sort all b β_g^i values by ascending counts to create the β_g vector. Each value in β_g denotes how many candidates from group g should be in each bin. For instance, $\Theta = \{\theta_1 = 8, \theta_2 = 6, \theta_3 = 10\}$ means from the first group we need 8 candidates, 6 from the second, and 10 from the third. Then we would have 8 bins where the

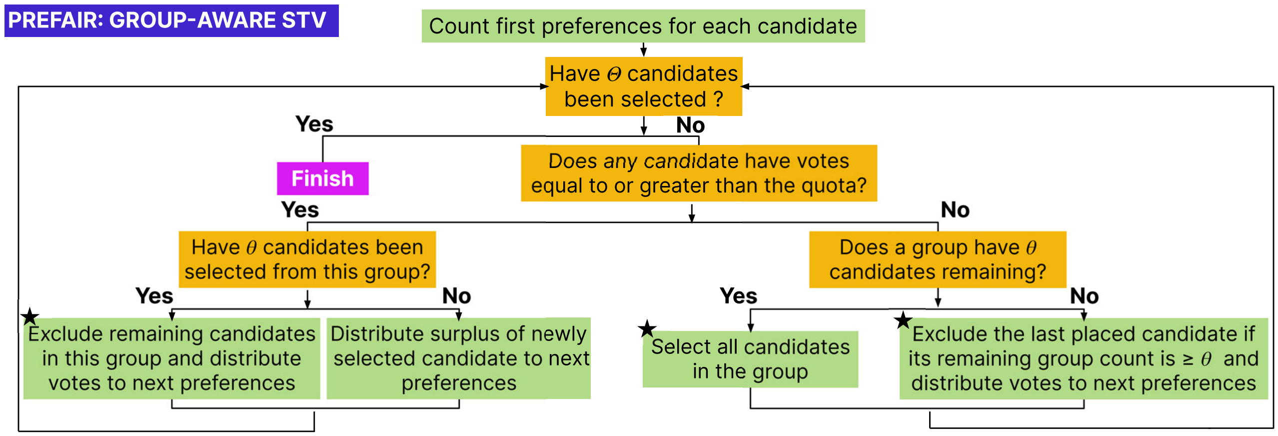


Figure 2: PreFair: Group-Aware Single Transferable Vote Diagram. For visual comparison, this chart resembles the traditional (candidate fairness-unaware) Single Transferable Vote diagram in Appendix B and at https://www.stv.govt.nz/countingdiagram.shtml.

first 6 contained one candidate from all groups, and the last two bins each contain 1 candidate from group g_2 and two candidates from group g_3 (i.e, $\beta_1 = [1,1,1,1,1,1,0,0]$, $\beta_2 = [1,1,1,1,1,1,1,1]$, and $\beta_3 = [1,1,1,1,1,1,2,2]$). Then, as candidates are selected through iterations of Group-aware STV, they are placed in the first bin that has no members of their group. The final consensus ranking τ flattens each bin, with candidates ordered by their arrival into the bin.

While our proposed binning approach to crafting τ ensures candidates are fairly ordered, PreFair's actual selection of these k candidates is performed by the Group-aware STV mechanism. This round-based selection process, referenced in Algorithm 1, is illustrated in Figure 2. Instead of simply finding k candidates belonging to *any group*, it seeks to find θ_i candidates $\forall\, g_i \in G$ (represented by Θ in Figure 2). PreFair's Group-aware STV has three key invariants, starred in Figure 2, that characterize a mechanistic difference compared to traditional STV. First, it only eliminates candidates from consideration if their corresponding group g_i contains at least θ_i candidates still under consideration. Second, if a group g_i has exactly θ_i candidates left, then all are selected. Third, if a group has candidates remaining after θ_i of them have already been selected for τ, then all remaining candidates are eliminated, and their votes are transferred to their subsequent preferences. Once all the candidates are selected by the Group-aware STV mechanism, PreFair's Aggregation component flattens the bins into τ as mentioned above.

4.2.3 Example of PreFair: Consider the four preference rankings: $r_1 = c_1 \prec c_2 \prec c_3 \prec c_4 \prec c_5 \prec c_6 \prec c_7, r_2 = c_2 \prec c_4 \prec c_7 \prec c_3 \prec c_1, r_3 = c_1 \prec c_2 \prec c_4 \prec c_3$, and $r_4 = c_5 \prec c_4 \prec c_7 \prec c_1 \prec c_3 \prec c_6$ (note $c_1 \prec c_2$ means c_1 is ranked higher than c_2), and $c_1, c_2, c_3, c_4 \in g_1$ and $c_5, c_6, c_7 \in g_2$. The goal is equal representation and $k = 4$, thus two members of each group must be in τ. The profile has enough candidates per group, so no additional candidates are pulled up during Configuration. Next, during Aggregation, the droop quota is $1 = \lfloor \frac{4}{4+1} \rfloor + 1$. In the first iteration of Group-aware STV, c_1, c_2 and c_5 are elected with 2, 1, and 1 first preference votes, respectively. c_1 and c_5 are placed in the first bin (in that order) and c_2 is in the second bin. The surplus of c_1 is transferred to its next preferences c_3 and c_4, who each receive .33 and .66 surplus, respectively[1]. In the next iteration, since group g_1 has its two members, this group is eliminated, and its surplus is transferred. From c_3 the next preference is c_6 (r_1, r_2) which now receives .33 votes. Then from c_4 the next preferences are c_6 $(r1)$ and c_7 (r_2, r_3) so c_6 now has .55 votes and c_7 has 0.44 votes. In the next iteration, no remaining candidate (c_6, c_7) has more votes than the quota. So c_7 is eliminated, since it has the least votes and c_6 is chosen. Then τ is $c_1 \prec c_5 \prec c_2 \prec c_6$.

4.2.4 Guaranteeing fair selection and ordering by PreFair. By construction, a consensus ranking produced by PreFair satisfies the chosen fair selection objective (either Eq. 2 or Eq. 1). It also satisfies ordering fairness (Eq. 4) because each prefix of size $|G|$ contains one candidate from every group, when possible. Below, we present Proposition 4.2 which captures these fairness properties of PreFair. The proof can be found in Appendix A.

Proposition 4.2 (Fairness Guarantees of PreFair). *Consensus ranking τ produced by PreFair is guaranteed to satisfy selection fairness (either Eq. 2 or Eq. 1) and ordering fairness with the minimum NDKL value (Eq. 4) of any ranking with the same number of candidates per group.*

5 EXPERIMENTAL STUDY

5.1 Datasets and Metrics

We evaluate PreFair on the following *datasets using four metrics.* Appendix C provides additional details about the processing of these datasets and their licenses, while Table 2 summarizes their characteristics. *Our software and experimental study implementation* are available at https://github.com/KCachel/prefair.

[1] Appendix B provides more details on vote transferring.

World Happiness [33]: Annual rankings of the top 20 countries by happiness over 17 years and $\mathcal{D}$ with 8 related features. Preference rankings have a strong bias toward European countries, with African countries entirely unranked.

IBM HR [34]: Rankings of top 500 employees by four performance indicators and $\mathcal{D}$ with 25 employment-related features. Groups are five distinct age categories; employees in their 30s are most preferred.

Econ Freedom [35]: Annual rankings of the top 40 countries by economic freedom over 10 years and $\mathcal{D}$ with 52 features. Groups are World Bank Regions. All regions are ranked, but preferences contain a strong European bias.

GSCI [55]: Annual rankings of the top 70 countries by global sustainability competitiveness index over 6 years and $\mathcal{D}$ with 6 features. Groups are continents, with a strong preference for European countries.

ACSEmp-Mallows: Is a partially synthetic dataset we create. We sample 25 candidates from four groups and their corresponding features $\mathcal{D}$, from the ACSEmployement dataset in Folktables [21]. Then we use the Mallows model [36, 43] to create 20 different preference profiles ranking these candidates, based on two parameters, α, a dispersion parameter in the Mallows model controlling agreement among voters in the preference profile, and preference completeness, a parameter controlling the partial-ness of the voter preference rankings. See Appendix C for further details.

For fairness metrics, we utilize the previously defined SFD (Eq. 3), NDKL (Eq. 4), and pNDKL (Eq. 5). To measure voter satisfaction, we average among voters the proportion of candidates shared between the top 10% of their preference rankings and the candidates in the top of the k-consensus ranking τ. We call this Average Sat and measure it at various τ *depths* (.1, .2, .3, .4, .5). For example, Average Sat with depth = .2 tells us the average proportion of candidates shared between the top 10% of candidates in the voter preference rankings and the top 20% of τ. The intuition is that higher Average Sat values at smaller depths show that τ is satisfying the voters' top preferences.

5.2 Compared Methods

We compare PreFair against the following *six preference aggregation methods.* They include both traditional fairness-unaware partial preference aggregators and state-of-the-art fairness-enhanced mechanisms that we can reasonably adapt to handle partial preferences. Table 4 compares the characteristics of these methods.

i) Borda [9]: A fairness-unaware positional scoring-based aggregator which handles partial voter preference rankings.

ii) STV [57]: The fairness-unaware single transferable voting method as described in Section 4.2.2.

iii) MC4 [25]: A fairness-unaware aggregator designed explicitly for partial voter preferences that approximates the Kemeny method [38]. While Kemeny is the gold-standard in aggregation [16, 50], its restricted to full preferences.

iv) Epira [11]: A post-processing fair rank aggregation method. While explicitly designed for full voter preferences [11], by using Borda as the aggregation method, we study Epira in our partial setting. Epira ensures the consensus ranking satisfies fairness of exposure [53], a form of ordering group fairness. As in [11], we set $\gamma = .9$ and use the highest exposure returned by the method.

v) Rapf [60]: A fair preference aggregation method also explicitly designed for full preferences. Yet, since it randomly selects a single voter's preference ranking and re-ranks it to be fair, we employ it in partial settings. It ensures the consensus fairly orders groups using the p-fairness notion proposed by Wei et al. [60]. Rapf randomly picks a voter's ranking to be the consensus, thus we report the average of ten runs.

vi) Fmwv [13]: A fair multi-winner voting method. Utilizing Borda scoring, Fmwv produces a *unranked consensus set* that satisfies constraints on the number of candidates chosen per group. We then rank this returned set using the scores returned by the method. If the profile contains the necessary candidates, this method ensures selection fairness.

STV and Fmwv have a parameter k for consensus size; for other methods, we utilize the top k candidates as τ.

5.3 Experimental Results

We compare PreFair with the above methods in regard to the fairness and voter satisfaction of the consensus ranking. We do so for four datasets, a controlled study on the impact of preference partial-ness, and an ablation study of PreFair's process. The best-performing methods ensure both selection and ordering fairness exhibited by low SFD and NDKL, and do so with high voter satisfaction, i.e., higher Average Sat values. *Across experiments, we show PreFair consistently has the best selection and ordering fairness performance and does not drastically degrade voter satisfaction.*

5.3.1 Real-World Bias Mitigations. Figures 3 and 4 present the results of all methods for equal and proportional selection fairness, respectively. Each figure is broken into two sub-figures, where (a.) displays performance on both fairness objectives and (b.) displays Average Sat at increasing consensus depths.

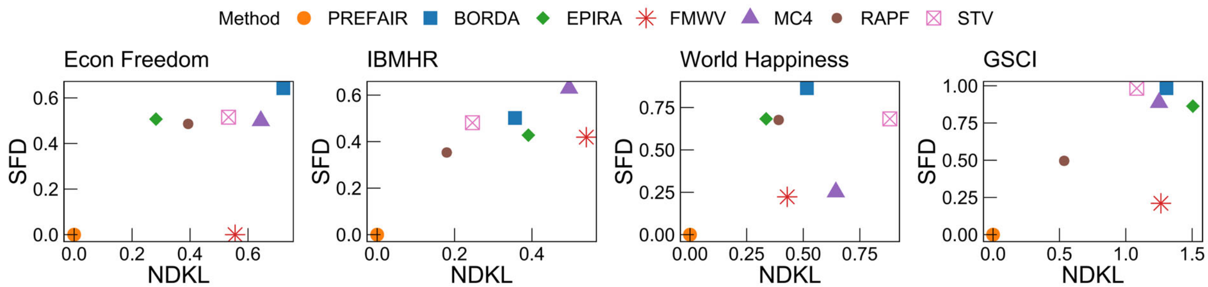
Notably, PreFair is the only method that produces a consensus that satisfies both selection and ordering fairness. Across datasets in Figures 3a and 4a, PreFair is always in the most desirable bottom left corner, representing low SFD and NDKL values. Moreover, in Figures 3b and 4b we see that on the whole PreFair has lower Average Sat values, but Average Sat does not drastically fall. Even in some datasets (e.g. *Econ Freedom* in Figure 3b), PreFair is not the worst performing method in terms of Average Sat scores. This decrease is expected because PreFair has to pull in additional candidates compared to the original preference profile.

The next best performing methods are Fmwv, Rapf, and Epira depending on the fairness metric. Fmwv tends to occupy the bottom right side of Figures 3a and 4a, meaning it does well in terms of selection fairness (SFD) and not in terms of ordering fairness (NDKL). However, even in terms of selection fairness, which Fmwv is optimized for, when the profile does not contain enough candidates from every group in order to make the fair selection, Fmwv does poorly. A good example of this is in *IBM HR* in Figure 3a. The profile had no members of the youngest group and since Fmwv only uses the profile, unlike PreFair, the final consensus is not a fair selection of the entire candidate pool.

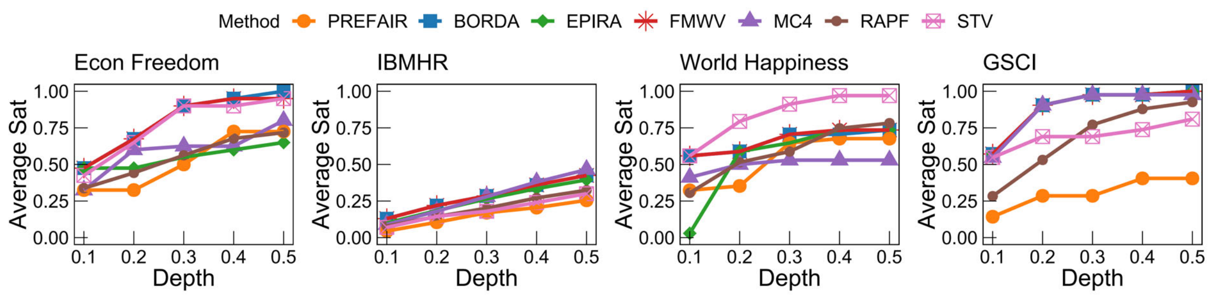
On the other hand, Rapf and Epira tend to have lower or middle of the road NDKL values and high SFD values. This indicates that

Table 2: Overview of datasets.

Dataset	Features d	Cands. in pool $\|C\|$	Cands. in profile $\|S\|$	Voters $\|R\|$	Groups in C	Groups in R	τ length k
World Happiness	8	153	50	17	Regions, 5	4	20
IBM HR	25	1462	754	4	Age, 5	5	100
Econ Freedom	52	157	70	10	World Bank Region, 5	5	20
GSCI	6	185	94	6	Continents, 6	6	36
ACSEmp-Mallows	16	100	varies.	50	Race, 4	varies	20



(a) All methods evaluated for selection fairness (lower SFD values) and ordering fairness (lower NDKL values).



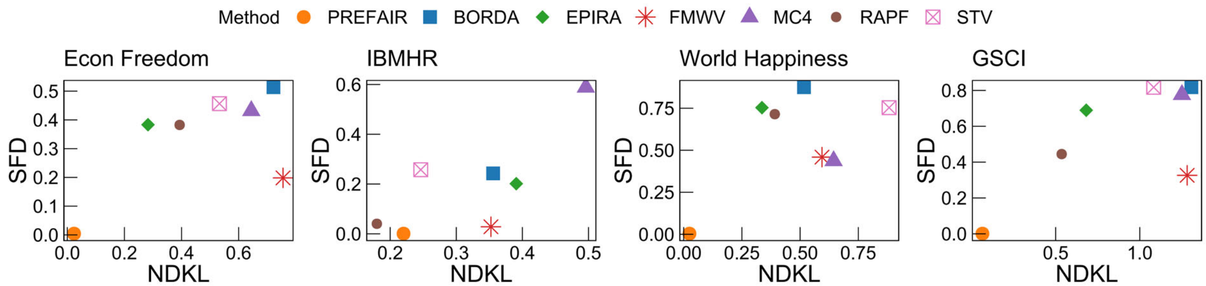
(b) All methods evaluated for voter satisfaction (higher Average Sat values) at different depths of the k-consensus ranking.

Figure 3: Results for equal representation based (Eq. 2) selection fairness in consensus rankings. Across all datasets, only PreFair has the best selection and ordering fairness performance – i.e., is plotted in the bottom left corner of plots in Figure 3a. Moreover, PreFair is not always ranked the lowest in terms of voter satisfaction in Figure 3b.

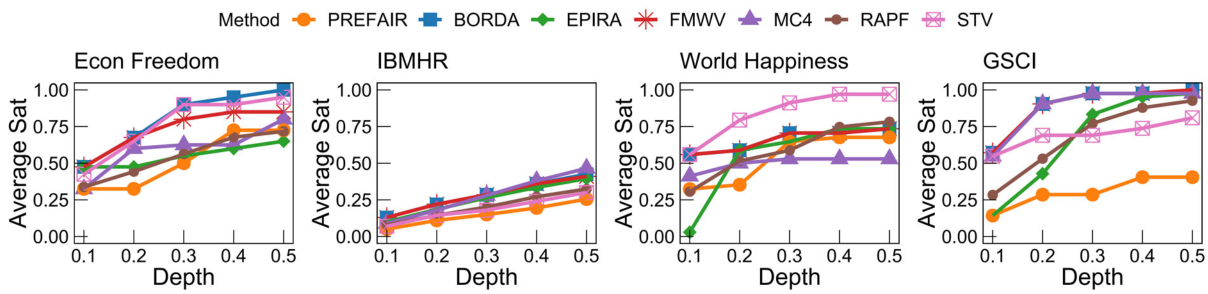
these methods do poorly on the selection fairness objective but help the ordering fairness goal. This is anticipated since neither method is concerned with selection fairness nor has access to candidates outside the profile. Nonetheless, they fairly rank candidates in the consensus ranking. Thus, we see Rapf and Epira tend to have the second and third best NDKL values after PreFair, e.g., *World Happiness* in Figures 3a and 4a. The primary drawback of Rapf and Epira in partial preference settings is that they do not ensure selection fairness. Moreover, for the *Econ Freedom* and *World Happiness* datasets, using a completely fairness-unaware method, such as STV or MC4, provides comparable or better selection fairness than Rapf and Epira.

Lastly, as expected, since Borda, MC4, and STV do not utilize candidate selection or ordering fairness interventions, they are, on the whole, the least fair. In Figures 3a and 4a, Borda, MC4, and STV occupy the top right corner, indicating little fairness of either kind. In Figures 3b and 4b, they tend to have the highest Average Sat values since their sole goal is voter satisfaction. It's hard to discern a clear order among them. It appears that STV offers slightly more selection fairness - e.g., *Econ Freedom* and *IBM HR* in Figure 3a. This likely stems from STV's proportionally property [9], meaning that if voters have diverse first preferences, this diversity translates into the consensus. Nonetheless, while these methods can aggregate partial preferences, they do not ensure fairness towards candidate groups.

5.3.2 Studying performance with varying conditions of partial-ness in voter preferences. Figure 5 presents a heatmap of pNDKL values for each method on the preference profiles in our controlled setting of the *ACSEmp-Mallows* dataset. For ease of visualization, we use pNDKL, since this metric combines ordering and selection fairness. Other metrics are in Appendix E. Moving up the y-axis, voter agreement with a biased central ranking increases. Moving left to right on the x-axis, the "partial-ness" of voter preferences decreases. For example, preference completeness of 1 and agreement $\alpha = 1$ means that all voters ranked all candidates and voters were in strong agreement with a central biased ranking. Whereas, preference completeness of 0.2 and $\alpha = 0.2$ means voters ranked



(a) All methods evaluated for selection fairness (lower SFD values) and ordering fairness (lower NDKL values).



(b) All methods evaluated for voter satisfaction (higher Average Sat values) at different depths of the k-consensus ranking.

Figure 4: Results for proportional representation based (Eq. 1) selection fairness in consensus rankings. Across all datasets, only PreFair has the best selection and ordering fairness performance – i.e., is plotted in the bottom left corner of plots in Figure 4a. Moreover, PreFair is not always ranked the lowest in terms of voter satisfaction in Figure 4b.

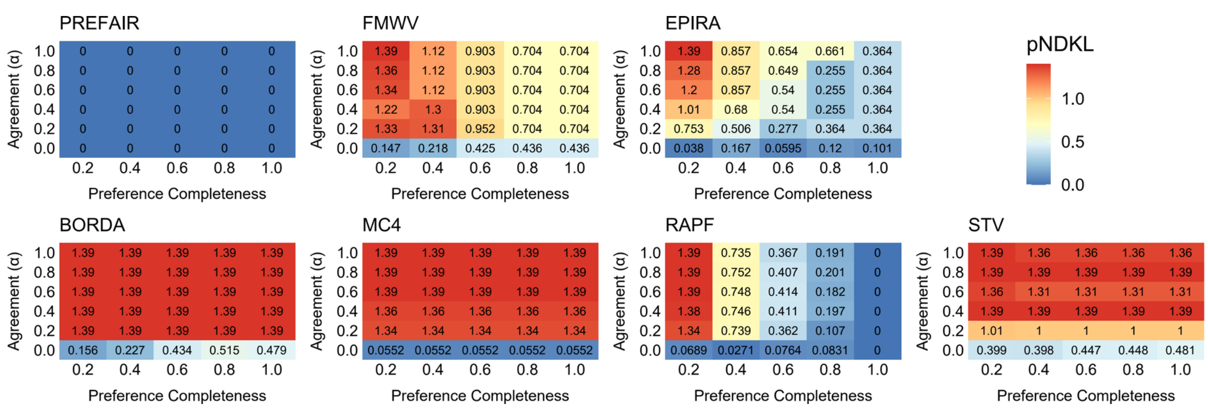


Figure 5: Only PreFair is fair under all voter agreement and preference profile partial-ness conditions. We see all methods are fair when there is no voter agreement, Fmwv, Epira, and Rapf are strongly affected by the partial-ness of voter preferences, and Epira, and STV are slightly affected by voter agreement conditions.

only 20 candidates each and had relatively little agreement with a central biased ranking.

We again observe the expected behavior that Borda, MC4, and STV do not produce fair consensus rankings under almost any voter agreement or preference completeness conditions. It is interesting to note, though, that all methods appear fair when $\alpha = 0$. This is because voters disagree with the biased central ranking, and, in turn, the profile and its resulting consensus ranking are both fair. Next, we see Fmwv, Epira, and Rapf are predominately impacted by preference completeness, i.e., consensus rankings are fairer on the right side of each heatmap since preferences are fuller. These methods perform the worst when preferences are more partial. Then, Epira, and STV are slightly affected by the levels of voter agreement. Their consensus rankings get fairer moving *down* each heatmap since the profile itself is fairer because there is less voter agreement with the biased central ranking. Only, PreFair is fair

Table 3: Ablation results for equal representation selection fairness (proportional in Appendix E). Best performance indicated in bold.

Dataset	Method		SFD ↓	NDKL ↓	pNDKL ↓	Average Sat ↑				
	Config.	Agg.				0.10	0.20	0.30	0.40	0.50
Econ Freedom	PreFair	PreFair	**0.000**	**0.000**	**0.000**	**0.325**	**0.325**	**0.500**	**0.725**	0.725
	PreFair	Epira	0.276	0.252	0.252	0.000	0.250	0.650	0.725	**0.775**
	PreFair	Rapf	0.055	0.155	0.155	0.133	0.320	0.393	0.567	0.595
GSCI	PreFair	PreFair	**0.000**	**0.000**	**0.000**	**0.143**	**0.285**	**0.285**	0.405	0.405
	PreFair	Epira	0.381	0.474	0.474	0.000	0.000	0.000	0.143	0.429
	PreFair	Rapf	0.165	0.167	0.167	0.143	0.226	0.283	**0.426**	**0.560**
IBM HR	PreFair	PreFair	**0.000**	**0.000**	**0.000**	0.045	0.105	0.170	0.205	0.255
	PreFair	Epira	0.431	0.341	0.341	**0.085**	**0.170**	**0.260**	**0.330**	**0.385**
	PreFair	Rapf	0.278	0.238	0.238	0.069	0.122	0.173	0.230	0.286
World Happiness	PreFair	PreFair	**0.000**	**0.000**	**0.000**	**0.324**	**0.353**	**0.647**	**0.676**	**0.676**
	PreFair	Epira	0.165	0.435	0.435	0.029	0.029	0.029	0.324	0.382
	PreFair	Rapf	0.156	0.106	0.106	0.053	0.326	0.365	0.524	0.556

under all agreement and partial-ness conditions in voter preference rankings.

5.3.3 Ablation study. To ensure that PreFair's Aggregation step contributes to its success, we conduct an ablation study. We compare PreFair with two similar approaches that first complete the Configuration step of our PreFair, specifically, its mechanism of pulling up additional candidates. Then, instead of using PreFair's Aggregation mechanism on the updated preferences, we use the Epira and Rapf methods [12, 60]. Appendix D details these implementation adjustments. Table 3 presents the results.

By using PreFair in its entirety, we do much better in all fairness objectives - i.e., SFD, NDKL, and pNDKL. PreFair also has the highest voter satisfaction in half of all Average Sat comparisons. Table 3 shows that if PreFair's Configuration step is used as a pre-processing step, Epira and Rapf exhibit increased performance compared to their standard versions – as can be seen in in Sections 5.3.1 and 5.3.2. This is expected as these approaches do not consider the selection fairness objective. Thus, by using PreFair to pull up additional candidates, these methods do better than they would have otherwise with only the partial preferences. Nonetheless, even with this extra support, using PreFair's Aggregation step with either Epira and Rapf underperforms the full PreFair approach in both fairness objectives of fair partial preference aggregation and risks decreased voter satisfaction.

6 LIMITATIONS AND FUTURE WORK

As this is the first work studying fairness in combining partial voter preference rankings, our approach has limitations. First, while PreFair produces consensus rankings that fairly represent and order groups, we are unable to make statements that PreFair minimally decreases voter satisfaction to meet these fairness objectives. This is because the tradeoff between fairness and voter satisfaction is controlled by the problem at hand, such as the preference profile, voter agreement, how well inference maps to unknown voter preferences, etc. Future work may explore ways to bound potential decreases in voter satisfaction in the PreFair methodology.

Second, as PreFair incorporates ranked candidate group fairness into the Single Transferable Voting mechanics, additional work could examine which social choice axioms are preserved between traditional STV and PreFair's Group-aware STV. For instance, anonymity [9] is preserved since voters are not weighted differently, and likewise, non-dictatorship [9] is still satisfied. However, future work might address additional axiomatic properties such as proportionality [26], and evaluate Group-aware STV as a standalone voting mechanism. Third, in this work, we conceptualize ordering fairness to align with the NDKL-based fair ordering framework with a single multi-valued attribute. Future work could study the incorporation of additional fairness notions and intersectional fairness concerns.

7 CONCLUSION

Our work introduces the fair partial preference aggregation problem for contexts where voters provide partial rankings of a candidate pool. To solve this problem, we introduce a novel strategy called PreFair. It features a unique preference inference mechanism with a novel Group-aware STV aggregation method. PreFair produces consensus rankings guaranteed to satisfy both selection fairness and ordering fairness. We demonstrate that, compared to existing alternatives from the literature, PreFair achieves the best performance in this dual fairness objective while ensuring voter satisfaction across a wide range of datasets and scenarios.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their feedback. This research was supported in part by the NSF-IIS 2007932 grant.

RESEARCHER STATEMENTS

Ethical Considerations. This work does not involve human subjects or participants. All datasets are publicly available and do not contain identifiable user data. In most cases, "candidates" are countries or synthetic entities.

We provide details on dataset download procedures, processing, creation in the case of the partially synthetic *ACSEmp-Mallows* dataset, and license information in Appendix C. We have made available all source code, datasets, and code for setting up and running the experimental study. This can be found in the corresponding PreFair repository (https://github.com/KCachel/prefair) which contains the GPL 3.0 license. Our experimental setup is described in Section 5 and Appendix D.

Adverse Impacts. While the PreFair methodology of this work facilitates performing fair preference aggregation of partial preferences; we would like to emphasize that not all *preference aggregation* tasks are *fair*[2] *preference aggregation* tasks, i.e., appropriate for the use of PreFair. A good example of this is democratic elections, such as those using ranked-choice ballots to determine political representatives. Using PreFair and, even more broadly, enforcing any additional criteria in such situations risks vote tampering. Algorithms like PreFair provide practitioners with an alternative to traditional preference aggregation mechanisms, yet their "fit" should be assessed by individual stakeholders.

Researcher Positionality. The authors are influenced by their perspectives centered on computing research and developing computational solutions to human-centric data problems. The authors draw from several years of experience in collective decision-making scenarios for tasks such as hiring and admissions in academic and industry settings.

REFERENCES

[1] Nir Ailon. 2007. Aggregation of Partial Rankings, p-Ratings and Top-m Lists. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (New Orleans, Louisiana) *(SODA '07)*. Society for Industrial and Applied Mathematics, USA, 415–424.

[2] Amanda A. Aird, Cassidy All, Paresha Farastu, Elena Stefancova, Joshua Sun, Nicholas Mattei, and Robin Burke. 2023. Exploring Social Choice Mechanisms for Recommendation Fairness in SCRUF. *ArXiv* abs/2309.08621 (2023).

[3] Javed A. Aslam and Mark Montague. 2001. Models for Metasearch. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, Louisiana, USA) *(SIGIR '01)*. Association for Computing Machinery, New York, NY, USA, 276–284. https://doi.org/10.1145/383952.384007

[4] Manel Ayadi, Nahla Ben Amor, Jérôme Lang, and Dominik Peters. 2019. Single Transferable Vote: Incomplete Knowledge and Communication Issues. In *Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1288–1296.

[5] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2017. Retrieval Consistency in the Presence of Query Variations. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 395–404. https://doi.org/10.1145/3077136.3080839

[6] John J. Bartholdi and James B. Orlin. 2015. Single transferable vote resists strategic voting. *Social Choice and Welfare* 8 (2015), 341–354.

[7] Gerdus Benade, Ruth Krebs Buck, Moon Duchin, Dara Gold, and Thomas Weighill. 2021. Ranked Choice Voting and Minority Representation. *Social Science Research Network* (2021).

[8] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2212–2220. https://doi.org/10.1145/3292500.3330745

[9] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. 2016. *Handbook of Computational Social Choice*. Cambridge University Press. https://doi.org/10.1017/CBO9781107446984

[10] Robert Bredereck, Piotr Faliszewski, Ayumi Igarashi, Martin Lackner, and Piotr Skowron. 2018. Multiwinner Elections with Diversity Constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[2]In this work, fairness is defined as fair with respect to candidate groups.

[11] Kathleen Cachel and Elke Rundensteiner. 2023. Fairer Together: Mitigating Disparate Exposure in Kemeny Rank Aggregation. (2023), 1347–1357. https://doi.org/10.1145/3593013.3594085

[12] Kathleen Cachel, Elke Rundensteiner, and Lane Harrison. 2022. MANI-Rank: Multiple Attribute and Intersectional Group Fairness for Consensus Ranking. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. 1124–1137. https://doi.org/10.1109/ICDE53745.2022.00089

[13] L. Elisa Celis, Lingxiao Huang, and Nisheeth K. Vishnoi. 2018. Multiwinner Voting with Fairness Constraints. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (Stockholm, Sweden) *(IJCAI'18)*. AAAI Press, 144–151.

[14] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2018. Ranking with Fairness Constraints. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 107)*, Ioannis Chatzigiannakis, Christos Kaklamanis, Dániel Marx, and Donald Sannella (Eds.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 28:1–28:15. https://doi.org/10.4230/LIPIcs.ICALP.2018.28

[15] Abhijnan Chakraborty, Gourab K. Patro, Niloy Ganguly, Krishna P. Gummadi, and Patrick Loiseau. 2019. Equality of Voice: Towards Fair Representation in Crowdsourced Top-K Recommendations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 129–138. https://doi.org/10.1145/3287560.3287570

[16] Vincent Conitzer, Andrew Davenport, and Jayant Kalagnanam. 2006. Improved Bounds for Computing Kemeny Rankings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 6. 620–626.

[17] Wade D. Cook, Boaz Golany, Michal Penn, and Tal Raviv. 2007. Creating a consensus ranking of proposals from reviewers' partial ordinal rankings. *Computers & Operations Research* 34, 4 (2007), 954–965. https://doi.org/10.1016/j.cor.2005.05.030

[18] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 758–759. https://doi.org/10.1145/1571941.1572114

[19] Salvatore Corrente, Salvatore Greco, Miłosz Kadziński, and Roman Słowiński. 2013. Robust Ordinal Regression in Preference Learning and Ranking. *Mach. Learn.* 93, 2–3 (Nov 2013), 381–422. https://doi.org/10.1007/s10994-013-5365-4

[20] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/3340531.3411962

[21] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 6478–6490. https://proceedings.neurips.cc/paper_files/paper/2021/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf

[22] John Doucette, Kate Larson, and Robin Cohen. 2015. Conventional Machine Learning for Social Choice. *Proceedings of the AAAI Conference on Artificial Intelligence* 29, 1 (Feb. 2015). https://doi.org/10.1609/aaai.v29i1.9294

[23] Henry Richmond Droop. 1881. On methods of electing representatives. *Journal of the Statistical Society of London* 44, 2 (1881), 141–202.

[24] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[25] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. 2001. Rank Aggregation Methods for the Web. In *Proceedings of the 10th International Conference on World Wide Web* (Hong Kong, Hong Kong) *(WWW '01)*. Association for Computing Machinery, New York, NY, USA, 613–622. https://doi.org/10.1145/371920.372165

[26] Piotr Faliszewski, Piotr Skowron, Stanisław Szufa, and Nimrod Talmon. 2019. Proportional Representation in Elections: STV vs PAV. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1946–1948.

[27] Yunhe Feng and Chirag Shah. 2022. Has CEO Gender Bias Really Been Fixed? Adversarial Attacking and Improving Gender Fairness in Image Search. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 11 (Jun. 2022), 11882–11890. https://doi.org/10.1609/aaai.v36i11.21445

[28] Erica B. Fields, GüL E. Okudan, and Omar M. Ashour. 2013. Rank Aggregation Methods Comparison: A Case for Triage Prioritization. *Expert Syst. Appl.* 40, 4 (mar 2013), 1305–1311. https://doi.org/10.1016/j.eswa.2012.08.060

[29] Edward A. Fox and Joseph A. Shaw. 1993. Combination of Multiple Searches. In *Text Retrieval Conference*.
[30] Chris Geller. 2005. Single transferable vote with Borda elimination: proportional representation, moderation, quasi-chaos and stability. *Electoral Studies* 24, 2 (2005), 265–280. https://doi.org/10.1016/j.electstud.2004.06.004
[31] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2221–2231. https://doi.org/10.1145/3292500.3330691
[32] Avijit Ghosh, Lea Genuit, and Mary Reagan. 2021. Characterizing Intersectional Group Fairness with Worst-Case Comparisons. In *Proceedings of 2nd Workshop on Diversity in Artificial Intelligence (AIDBEI) (Proceedings of Machine Learning Research, Vol. 142)*, Deepti Lamba and William H. Hsu (Eds.). PMLR, 22–34.
[33] JF Helliwell, H Huang, M Norton, L Goff, and S Wang. 2023. World happiness, trust and social connections in times of crisis. *World Happiness Report* (2023).
[34] IBM. 2016. *IBM HR Analytics Employee Attrition & Performance.* Technical Report. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset
[35] The Fraser Institute. 2023. *Economic Freedom of the World 2023 Annual Report.* Technical Report. https://www.fraserinstitute.org/economic-freedom/dataset
[36] Ekhine Irurozki, Borja Calvo, and Jose A Lozano. 2016. PerMallows: An R package for Mallows and generalized Mallows models. *Journal of Statistical Software* 71, 1 (2016), 1–30.
[37] Anson Kahng, Yasmine Kotturi, Chinmay Kulkarni, David Kurokawa, and Ariel Procaccia. 2018. Ranking Wily People Who Rank Each Other. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1. https://doi.org/10.1609/aaai.v32i1.11467
[38] John G Kemeny. 1959. Mathematics without numbers. *Daedalus* 88, 4 (1959), 577–591.
[39] Caitlin Kuhlman and Elke Rundensteiner. 2020. Rank Aggregation Algorithms for Fair Consensus. *Proc. VLDB Endow.* 13, 12 (jul 2020), 2706–2719. https://doi.org/10.14778/3407790.3407855
[40] Solomon Kullback. 1997. *Information theory and statistics.* Courier Corporation.
[41] Jérôme Lang. 2020. Collective Decision Making under Incomplete Knowledge: Possible and Necessary Solutions. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, 4885–4891. https://doi.org/10.24963/ijcai.2020/680 Survey track.
[42] David Lillis, Fergus Toolan, Rem Collier, and John Dunnion. 2006. ProbFuse: A Probabilistic Approach to Data Fusion. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Seattle, Washington, USA) *(SIGIR '06)*. Association for Computing Machinery, New York, NY, USA, 139–146. https://doi.org/10.1145/1148170.1148197
[43] Colin L Mallows. 1957. Non-null ranking models. *Biometrika* 44, 1/2 (1957), 114–130.
[44] Farhad Mohsin, Ao Liu, Pin-Yu Chen, Francesca Rossi, and Lirong Xia. 2022. Learning to Design Fair and Private Voting Rules. *J. Artif. Int. Res.* 75 (dec 2022), 38 pages. https://doi.org/10.1613/jair.1.13734
[45] André Mourão, Flávio Martins, and João Magalhães. 2013. NovaSearch at TREC 2013 Federated Web Search Track: Experiments with rank fusion. In *TREC*.
[46] Gourab K. Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. 2022. Fair Ranking: A Critical Review, Challenges, and Future Directions. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1929–1942. https://doi.org/10.1145/3531146.3533238
[47] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. (2017), 5684–5693.
[48] Amifa Raj and Michael D. Ekstrand. 2022. 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* (, Madrid, Spain,) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 726–736. https://doi.org/10.1145/3477495.3532018
[49] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation of BERT Rankers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (, Virtual Event, Canada,) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 306–316. https://doi.org/10.1145/3404835.3462949
[50] Frans Schalekamp and Anke van Zuylen. 2009. Rank Aggregation: Together We're Strong. In *Proceedings of the Meeting on Algorithm Engineering & Expermiments* (New York, New York). Society for Industrial and Applied Mathematics, USA, 38–51.
[51] Tobias Schumacher, Marlene Lutz, Sandipan Sikdar, and Markus Strohmaier. 2022. Properties of Group Fairness Metrics for Rankings. *ArXiv* abs/2212.14351 (2022).
[52] D. Sculley. [n. d.]. Rank Aggregation for Similar Items. In *Proceedings of the 2007 SIAM International Conference on Data Mining (SDM)*. 587–592. https://doi.org/10.1137/1.9781611972771.66
[53] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2219–2228. https://doi.org/10.1145/3219819.3220088
[54] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for Fairness in Ranking. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, Article 487, 11 pages.
[55] Solability. 2023. *The Global Sustainability Competitiveness Index.* Technical Report. https://solability.com/the-global-sustainable-competitiveness-index/the-index/downloads
[56] Julia Stoyanovich, Ke Yang, and HV Jagadish. 2018. Online set selection with fairness and diversity constraints. In *Proceedings of the EDBT Conference*.
[57] Nicolaus Tideman. 1995. The Single Transferable Vote. *Journal of Economic Perspectives* 9 (1995), 27–38.
[58] Rohith Dwarakanath Vallam, Ramasuri Narayanam, Srikanth G. Tamilselvam, Nicholas Mattei, Sudhanshu S. Singh, Shweta Garg, and Gyana R. Parija. 2019. DeepAggregation: A New Approach for Aggregating Incomplete Ranked Lists Using Multi-Layer Graph Embedding. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2235–2237.
[59] Sriram Vasudevan and Krishnaram Kenthapadi. 2020. LiFT: A Scalable Framework for Measuring Fairness in ML Applications. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 2773–2780. https://doi.org/10.1145/3340531.3412705
[60] Dong Wei, Md Mouinul Islam, Baruch Schieber, and Senjuti Basu Roy. 2022. Rank Aggregation with Proportionate Fairness. In *Proceedings of the 2022 International Conference on Management of Data* (Philadelphia, PA, USA) *(SIGMOD '22)*. Association for Computing Machinery, New York, NY, USA, 262–275. https://doi.org/10.1145/3514221.3517865
[61] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management* (Chicago, IL, USA) *(SSDBM '17)*. Association for Computing Machinery, New York, NY, USA, Article 22, 6 pages. https://doi.org/10.1145/3085504.3085526
[62] H. Peyton Young. 1974. An axiomatization of Borda's rule. *Journal of Economic Theory* 9 (1974), 43–52.
[63] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, Singapore) *(CIKM '17)*. Association for Computing Machinery, New York, NY, USA, 1569–1578. https://doi.org/10.1145/3132847.3132938
[64] Meike Zehlike and Carlos Castillo. 2020. Reducing Disparate Exposure in Ranking: A Learning To Rank Approach. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 2849–2855. https://doi.org/10.1145/3366424.3380048

A ADDITIONAL TECHNICAL DETAILS FOR FAIR PARTIAL PREFERENCE AGGREGATION

This section presents the proofs of technical results appearing in Section 4.

A.1 Impossibility of Selection Fairness for Partial Preferences

Proof of Proposition 4.1. Fair selection of candidates from the candidate pool C, in terms of either proportional or equal representation, i.e., as in Eq. 1 and Eq. 2, requires a certain number of candidates per group to be in the consensus ranking τ. In section 4 we model the number of candidates per group g_i that need to be in τ as $\theta_g i$, where Θ is the set of values for all group G. Given a preference profile R, R contains a set of candidates $S \cap C$. By definition of θ_{g_i}, when $\S \cap g_i$ is less than θ_{g_i} we cannot produce a consensus ranking τ that is fair selection of candidates form C. □

A.2 Fairness Gurantees of PreFair

Proof of Proposition 4.2. First, we observe that the PreFair algorithm includes exactly as many candidates per group as needed to translate either equal (Eq. 2) or proportional representation (Eq. 1) into per-group candidate counts modeled as Θ. It follows that PreFair satisfies selection fairness for the desired fair selection objective (equal or proportional representation) and consensus ranking length k. Thus, we need to show that consensus ranking τ produced by PreFair minimizes the NDKL (Eq. 4) value for the number of candidates per group in τ. Proceeding by contradiction, assume ranking τ' orders the same number of candidates per group as ranking τ produced by PreFair, and $\text{NDKL}(\tau', G) < \text{NDKL}(\tau, G)$. By the definition of NDKL (Eq. 4), ranking τ' must contain a multiple of $|G|$ sized prefix with stronger equal representation of groups than τ. This corresponds to a contradiction, since our PreFair would have equally represented each group at each "bin" (i.e., prefix) according to Eq. 6, and when each bin was flattened, groups are represented equally at such prefixes. □

B ADDITIONAL DETAILS FOR SINGLE TRANSFERABLE VOTING AND GROUP-AWARE STV

This section provides additional background on STV as used in Section 4.

B.1 Quota Calculation, Vote Transfer, and Tie-breaking

Single Transferable Voting is largely a family of round-based preference aggregation algorithms; however, specific implementations vary in how the quota is calculated, how votes are transferred amongst candidates, and how ties are broken. In this work, and in our implementation of STV and PreFair we utilize the following standard strategies. For the quota, we utilize the Droop quota [23], $\lfloor \frac{n}{k+1} \rfloor + 1.$, alternatives include the Hare quota [57]. Vote transferring when a candidate (or group) is eliminated and after election is done the same way. Specifically, we use the standard formula[3]

$$\frac{Surplus\ of\ original\ candidate}{Total\ transferable\ next\ preferences\ of\ original\ candidate} \times$$

$$(next\ preference\ for\ the\ transferred\ candidate).$$

For example, if candidate c_1 has 3 surplus votes and its next preferences are c_2, c_2, and c_3 then candidate c_2 receives $\frac{3}{3} \times 2 = 2$ transferred votes and candidate c_3 receives $\frac{3}{3} \times 1 = 1$ transferred votes. For tie-breaking in elimination and election we utilize Borda scores of candidates [30].

B.2 Graphical Overview of Conventional STV

C ADDITIONAL DATASET DETAILS

This section presents processing details for each of the datasets used in Section 5. The scripts for performing this processing and their corresponding experiments can be found at https://github.com/KCachel/prefair.

[3]See also https://en.wikipedia.org/wiki/Single_transferable_vote for an intuitive explanation.

Econ Freedom. We use the data available at https://www.fraserinstitute.org/economic-freedom/dataset; specifically, the file presenting the entire economic freedom dataset.[4] Candidates are countries, and groups are the provided World Bank regions. The voters represent preference rankings produced by utilizing each year (data collected every five years between 1975 – 2000 and then annually between 2017 – 2022) as a voter, and ranking by the Economic Freedom Summary Index. Note that incompleteness arises in the preference profile since not all countries had data collected over all years. For dataset $\mathcal{D}$, we average each country's values in 52 numeric features over all the years provided in the datsets. These features include scores such as judicial independence, gender disparity, and inflation.

GSCI. We use the data provided at https://solability.com/the-global-sustainable-competitiveness-index/the-index/downloads provided under the Creative Commons License (CC BY-NC-SA 4.0). We use the GSCI Score <year> files for 2017 – 2022, which must be downloaded individually and then combined. Countries act as candidates, and groups are regions that are created using the pycountry Python package[5] to convert country names to regions. The voters represent preference rankings produced by utilizing each year as a voter, and ranking by the Sustainable Competitiveness. Note that incompleteness arises in the preference profile since not all countries were evaluated for every year. For dataset $\mathcal{D}$, we average each country's values in 6 features over all the years provided in the data. These features include scores such as social capital and governance.

IBM HR. We use the IBM HR Analytics Employee Attrition & Performance dataset hosted on Kaggle[6] provided with the DbCL license. Candidates are employees, and groups are age brackets created by binning the age variable to $10s, 20s, 30s, 40s, \geq 50$. The voters represent preference rankings produced by utilizing the features YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager as voters, and ranking by their scores. Note that this has the intended effect of creating an overall age-biased preference profile. For dataset $\mathcal{D}$, we use the additional 25 numeric features (e.g. distance from home and hourly rate) associated with employees.

World Happiness. We use the data provided at https://happiness-report.s3.amazonaws.com/2023/DataForTable2.1WHR2023.xls, as part of the 2023 World Happiness Report published by the Sustainable Development Solutions Network [33][7]. Countries act as candidates, and groups are regions which are created using the pycountry Python package[5] to convert country names to regions. The voters represent preference rankings produced by utilizing each year from 2006 to 2023 as a voter, and ranking by the overall happiness Life Ladder features. Note that incompleteness arises in the preference profile since not all countries were evaluated for every year. For dataset $\mathcal{D}$, we average each country's values in

[4]Per the Fraser Institute terms, our repository cannot directly contain the Economic Freedom data. The Fraser Institute makes the data publicly available, but users wishing to use it must download it themselves.

[5] https://pypi.org/project/pycountry/.

[6]https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset.

[7]No license is provided.

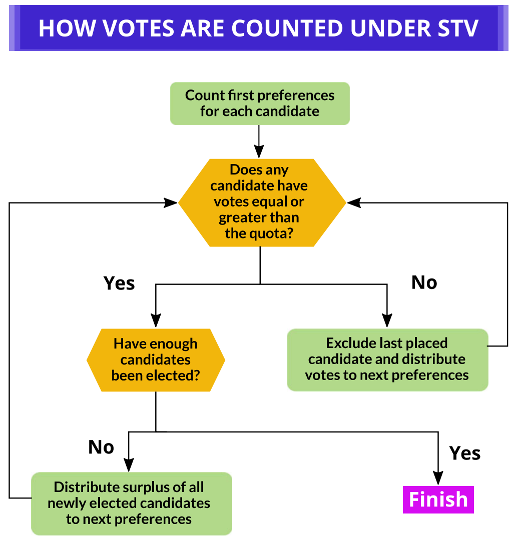


Figure 6: Single Transferable Vote Counting Diagram. Accessed from https://www.stv.govt.nz/countingdiagram.shtml.

8 features over all the years provided in the data. These features include scores such as social support, perceptions of corruption, and healthy life expectancy at birth.

ACSEmp-Mallows. We create a partially synthetic dataset for our controlled study using the *ACSEmployment* dataset in Folktables, provided with the MIT license, and the Mallows model [36, 43]. To create a candidate pool of 100 candidates with four equal-sized groups, we sample four groups in the *ACSEmployment* dataset, specifically for the state of Alabama in survey year 2018. We use the 16 features associated with candidates as dataset $\mathcal{D}$. Then, in order to create a preference profile, we utilize the Mallows model [36, 43]. The model utilizes a central ranking of candidates and a dispersion parameter α, which, as it increases, the profile contains more consensus (agreement) with the provided central ranking. For our central ranking, we use a completely biased ordering of candidates where one group is stacked on top of another. Specifically, we place order candidates by group 1, group 2, group 3, and group 4.

Using the Mallows model we create 20 different preference profiles. First, we start by creating profiles of 50 voter preference rankings with six different α dispersion values, e.g. $\alpha = 0.0, 0.2, 0.4, 0.6, 0.8, 1$. Next, in order to model different degrees of incompleteness in the preference profile we introduce a *preference completeness* $= 0.2, 0.4, 0.6, 0.8, 1$ parameter. The *preference completeness* parameter "truncates" the profile at *preference completeness* $*100$ candidates. For instance, for the profile $\alpha = 0$ and *preference completeness* $= 0.2$ we take the preference profile from the Mallows model and use only the top 20 candidates. In this way we are able to study how the conditions of voter agreement (modeled by α) and preference incompleteness (modeled by *preference completeness*) interact. To provide a sense of the fairness of each of the profiles, Figure 7 displays the KL-divergence between group proportions in the resulting profile and the overall candidate pool. As we can see, the KL-divergence decreases as *preference completeness* increases because more candidates are in the profile as a result of the length of voter preference. The KL-divergence also increases as α increases since more voter agreement implies fewer candidates are ranked in the profile. The code to generate these datasets, run the experiment, and plot the visuals can be found in https://github.com/KCachel/prefair.

D ADDITIONAL EXPERIMENTAL DETAILS

This section provides additional details for the methods used in Section 5.

D.1 Characterizing Methods

See table 4. We implemented the Borda, STV, and MC4 methods ourselves and used the code provide in Cachel et al. [12], Wei et al. [60], and Celis et al. [13], for Epira, Rapf, and Fmwv respectively.

D.2 Implementation Details for Section 5.3.3

In Sections 5.3.1 and 5.3.2 the Epira and Rapf methods are directly passed partial preference input (and some full preference in scenarios in Section 5.3.2). Yet, in Section 5.3.3 we perform an ablation of PreFair, and utilize it's first step (Configuration as in Section 4.2.1) and then entirely replace PreFair's Aggregationstep with Epira and Rapf. This is shown in tables 3 and 5.

E ADDITIONAL EXPERIMENTAL RESULTS

This section presents additional experiments complementing those in Section 5.

E.1 Additional Performance Metrics for Controlled Study

We provide the results of additional metrics from the performance study on the effects of voter agreement conducted in Section 5.3.2. First, examining SFD the selection fairness metrics in Figure 8a we see that only PreFair is fair under all voter agreement and

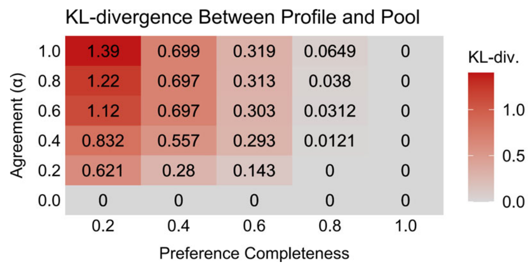


Figure 7: Representation of the candidate pool C in the preference profiles included in the *ACSEmp-Mallows* dataset.

Method	Native Partial Preference Support	Ordering Fairness Support	Selection Fairness Support
Borda [9]	✓	✕	✕
STV [57]	✓	✕	✕
MC4 [25]	✓	✕	✕
Epira [11]	✕ (adapt by using Borda)	✓ (fairness of exposure)	✕
Rapf [60]	✕*	✓ (p fairness)	✕
Fmwv [13]	✓**	✕	✓
PreFair (Ours)	✓	✓	✓

Table 4: Overview of Methods. *Wei et al. [60] note Rapf is not conceptually geared toward handling partial preferences, but we find it does not break on partial preference input. **Celis et al. [13] note Fmwv can handle individual voters ranking disjoint subsets, however, they assume that all candidates in candidate pool C are ranked at least once by a voter, meaning $S = C$.

preference profile completeness conditions. Figure 9 augments Figure 8a, by providing the count of the distinct number of groups included in each consensus decision. Looking specifically at Figure 8a we see all methods are fairer when there is no voter agreement with the central biased ranking (e.g., $\alpha = 0$). Then Fmwv, Epira, and Rapf are strongly affected by the partial-ness of voter preferences; but, Fmwv does the best of these three methods because it is designed for selection fairness, provided the necessary count of candidates per group are in the profile. Lastly, Fmwv, Epira, Borda, MC4, and STV are all affected by voter agreement conditions. Only PreFair has the lowest SFD values (Figure 8a) and included all four groups in every consensus decision (Figure 9).

Turning to the NDKL values expressed in Figure 8b, on first sight it appears that most methods produce fair orderings. However, the distinct group counts in Figure 9 explain that this is because in the fairness-unaware methods, e.g., Borda or MC4, these consensus rankings only have one group in them, thus by definition of NDKL they have high NDKL values. As expected since neither Fmwv or STV optimize for fair orderings whatsoever these methods, on the whole, have higher NDKL values. Then examining, Epira and Rapf they have low NDKL values since these methods optimize for fairly ordering candidates in their consensus rankings. However, recalling the selection fairness results in Figure 8a, only PreFair always has both selection fairness (low SFD values) *and* ordering fairness (low NDKL values).

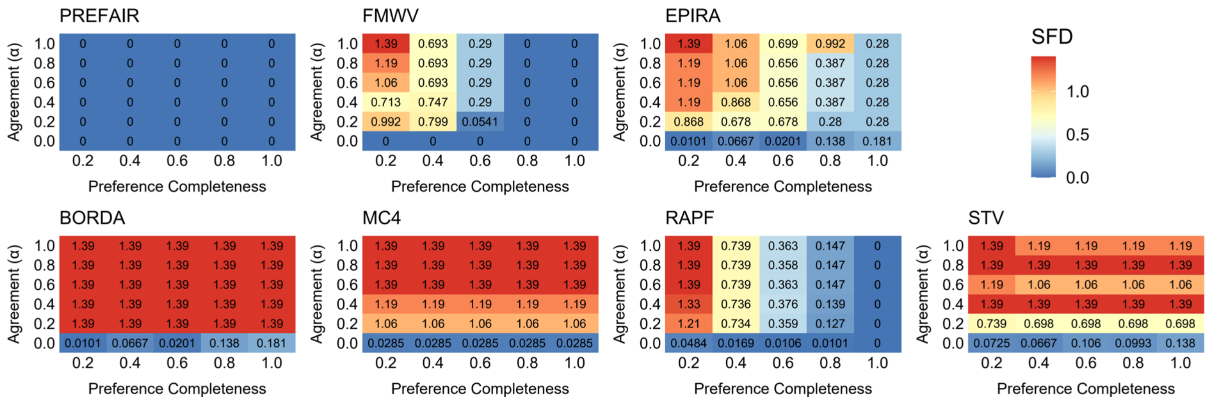
E.2 Section 5.3.3 Ablation Study Results for Proportional Representation

In comparing PreFair to ablated versions of PreFair using Epira and Rapf, we see for proportional representation in Table 5 the same conclusions as in equal representation shown in Table 3. That is, first, by using PreFair in its entirety, we do much better in all fairness objectives - SFD, NDKL, and pNDKL. And second PreFair has the highest voter satisfaction in half of all Average Sat comparisons. And lastly, if PreFair's Configuration step is used as a pre-processing step, Epira and Rapf perform much better compared to prior studies in Section 5.3.1 and 5.3.2. Nonetheless, using PreFair's Aggregation step with either Epira and Rapf underperforms the full PreFair approach in terms of both fairness objectives of fair partial preference aggregation, and risks decreased voter satisfaction.

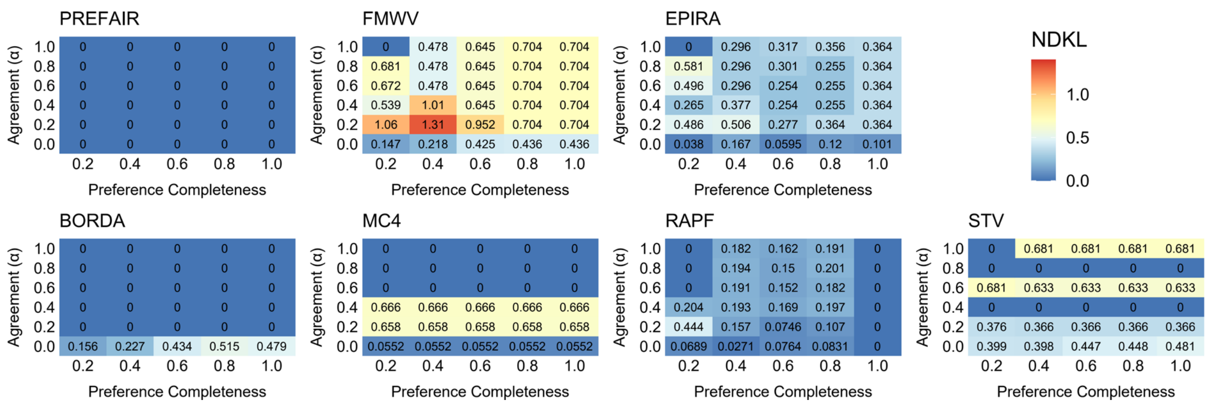
E.3 Examining PreFair's Configuration Pulling Up Approach

To ensure the PreFair Configuration strategy of pulling up additional candidates to mitigate voter selection bias is effective, we compare it to a simple alternative approach. Specifically, we contrast the strategy presented in Algorithm 1, with a naive approach of appending to each partial voter preference ranking a random ordering of the remaining candidates that the voter did not rank.

To test this, we do the following. As used and described in Section 5.3, for each dataset, each voter preference ranking contains the top x candidates of each voter. We intentionally, take the top x candidates in creating a partial preference profile from the dataset, so that we may have existing longer rankings (more candidates) to compare against. We call these *existing voter preferences.* Note these existing preference rankings are still considered partial preferences since they order disjoint subsets due to how the data was collected. For example, in the *World Happiness* the same countries are not ranked every year. Nonetheless, these existing preferences give us



(a) SFD results.



(b) NDKL results.

Figure 8: Only PreFair has low SFD and NDKL values under all voter agreement and preference profile completeness conditions. We see all methods are fairer when there is no voter agreement, both fairness objectives of Fmwv, Epira, and Rapf are strongly affected by the partial-ness of voter preferences, and the SFD and NDKL values of Fmwv, Epira, Rapf, MC4 and STV are affected by voter agreement conditions.

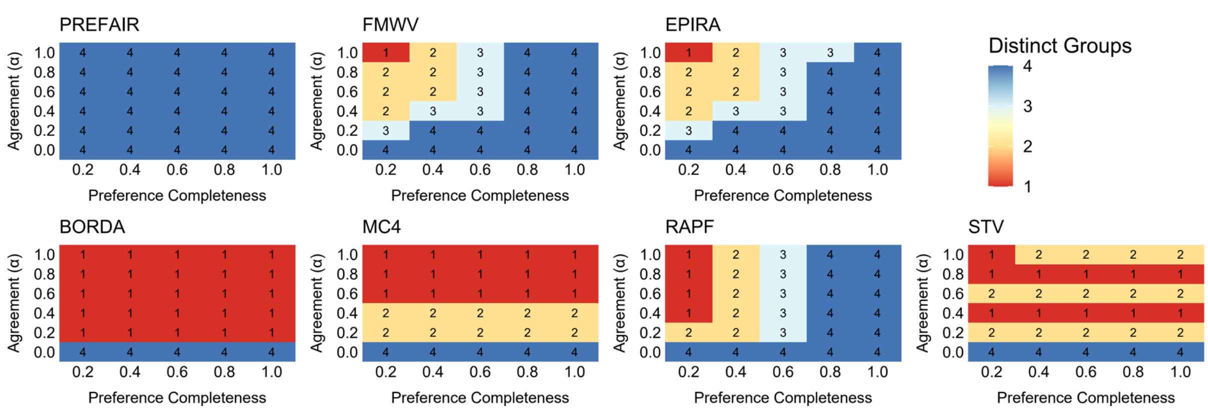


Figure 9: Only PreFair includes all four groups in the consensus decision under all voter agreement and preference profile completeness conditions. We see all methods include all groups when there is no voter agreement and all methods, besides PreFair, are affected by voter agreement. Moreover, Fmwv, Epira, and Rapf include less groups when the profile is less complete.

Dataset	Method		SFD ↓	NDKL ↓	pNDKL ↓	Average Sat ↑				
	Config.	Agg.				0.10	0.20	0.30	0.40	0.50
Econ Freedom	PreFair	PreFair	**0.004**	**0.024**	**0.024**	**0.325**	**0.325**	**0.500**	**0.725**	0.725
	PreFair	Epira	0.175	0.252	0.252	0.000	0.250	0.650	0.725	**0.775**
	PreFair	Rapf	0.011	0.156	0.156	0.133	0.320	0.393	0.567	0.595
GSCI	PreFair	PreFair	**0.001**	**0.064**	**0.056**	**0.143**	**0.286**	0.286	0.405	0.405
	PreFair	Epira	0.196	0.474	0.474	0.000	0.000	0.000	0.143	0.429
	PreFair	Rapf	0.002	0.180	0.180	0.143	0.257	**0.286**	**0.429**	**0.567**
IBM HR	PreFair	PreFair	**0.001**	**0.220**	**0.220**	0.050	0.110	0.150	0.195	0.255
	PreFair	Epira	0.203	0.341	0.341	**0.085**	**0.170**	**0.260**	**0.330**	**0.385**
	PreFair	Rapf	**0.001**	**0.247**	**0.247**	0.065	0.122	0.173	0.230	0.286
World Happiness	PreFair	PreFair	**0.004**	**0.025**	**0.025**	**0.324**	**0.352**	**0.647**	**0.6**	**0.618**
	PreFair	Epira	0.148	0.435	0.435	0.029	0.029	0.029	0.324	0.382
	PreFair	Rapf	0.011	0.091	0.091	0.088	0.321	0.326	0.468	0.500

Table 5: Ablation results for proportional representation selection fairness. Best performance indicated in bold.

	Method	*Econ Freedom*	*GSCI*	*IBM HR*	*World Happiness*
Average Kendall tau ↓	PreFair	4005.50	**2488.33**	**224395.25**	**1657.35**
	Random	**3455.52**	2939.73	231373.65	2162.21

Table 6: Results for pulling up additional candidates via PreFair's Configuration (Section 4.2.1) approach compared to randomly adding unranked candidates. Best performance indicated in bold.

a baseline to compare our PreFair inferred rankings against. Thus, for PreFair, denoted as PreFair in Table 6, and for the simple inference method described above, denoted as Random in Table 6, we compute the average Kendall tau distance between each *inferred* voter preference ranking and the existing voter preference ranking. Since existing preferences are partial, we omit candidates that do not appear in the existing candidate when measuring Kendall tau. Additionally, since the compared approach involves randomness, Table 6 reports Random's average Kendall tau over 10 trails.

We observe that PreFair has lower average Kendall tau distances in three out of four datasets, indicating that the inferred rankings are significantly closer to the existing rankings than when using a random inference approach. However, in the *Econ Freedom* dataset, we see that Random has a better performance. We suspect this is because of the relatively large number of features used in PreFair's inference; specifically, the *Econ Freedom* dataset has 52 features compared to 8, 25, and 6 in the respective *World Happiness*, *IBM HR*, and *GSCI* datasets. PreFair may have overfit to a pattern, uncovered over so many features, that was not present in the organic existing rankings. This echoes our call in Section 4 for future work developing more custom and nuanced methods to pulling up additional candidates. Nonetheless, averaging over the *World Happiness*, *IBM HR*, and *GSCI* datasets, the average Kendall tau distance between inferred and existing rankings for method was 86% of the distance of the Random approach, indicating PreFair provides gains in modeling unknown voter preferences.