MANI-Rank: Multiple Attribute and Intersectional Group Fairness for Consensus Ranking

Kathleen Cachel  
Worcester Polytechnic Institute  
kcachel@wpi.edu

Elke Rundensteiner  
Worcester Polytechnic Institute  
rundenst@wpi.edu

Lane Harrison  
Worcester Polytechnic Institute  
ltharrison@wpi.edu

Abstract—Combining the preferences of many rankers into one single consensus ranking is critical for consequential applications from hiring and admissions to lending. While group fairness has been extensively studied for classification, group fairness in rankings and in particular rank aggregation remains in its infancy. Recent work introduced the concept of fair rank aggregation for combining rankings but restricted to the case when candidates have a single binary protected attribute, i.e., they fall into two groups only. Yet it remains an open problem how to create a consensus ranking that represents the preferences of all rankers while ensuring fair treatment for candidates with multiple protected attributes such as gender, race, and nationality. In this work, we are the first to define and solve this open Multi-attribute Fair Consensus Ranking (MFCR) problem. As a foundation, we design novel group fairness criteria for rankings, called MANI-Rank, ensuring fair treatment of groups defined by individual protected attributes and their intersection. Leveraging the MANI-Rank criteria, we develop a series of algorithms that for the first time tackle the MFCR problem. Our experimental study with a rich variety of consensus scenarios demonstrates our MFCR methodology is the only approach to achieve both intersectional and protected attribute fairness while also representing the preferences expressed through many base rankings. Our real world case study on merit scholarships illustrates the effectiveness of our MFCR methods to mitigate bias across multiple protected attributes and their intersections. This is an extended version of “MANI-Rank: Multiple Attribute and Intersectional Group Fairness for Consensus Ranking”, to appear in ICDE 2022 [1].

Index Terms—Fair decision making, consensus ranking, bias.

I. INTRODUCTION

Rankings are used to determine who gets hired for a job [2], let go from a company [3], admitted to school [4], or rejected for a loan [5]. These consequential rankings are often determined through the combination of multiple preferences (rankings) provided by decision makers into a single representative consensus ranking that best reflects their collective preferences.

However, human and algorithmic decision-makers may generate biased rankings [6]–[8]. Such unfair rankings when combined may escalate into a particularly unfair consensus ranking. A prevalent type of fairness, called group fairness, is concerned with the fair treatment of candidates regardless of their values for a protected attribute. Protected attributes may be traits such as Gender, Race, or Disability, that are legally protected from discrimination. More broadly, protected attributes correspond to any categorical sensitive attributes for which bias mitigation is desired. The problem of incorporating multiple fairness objectives into the process of producing a consensus ranking, namely, multi-attribute fair consensus ranking (MFCR), remains open.

Admissions Example. Consider an admissions committee ranking applicant candidates for scholarship merit awards as seen in Figure 1. First, each of the four committee members, potentially assisted with an AI-screening tool [9], ranks the prospective candidates — illustrated with rankings $r_1$ to $r_4$. The committee seeks a consensus ranking for final decision making by consolidating the individual rankings into one single ranking. For the committee members to accept the outcome, the consensus ranking needs to reflect all the individual rankings in as much as possible.

The rankings may conflict substantially — as we can see comparing $r_4$ to $r_3$ in Fig. 1. Ranking $r_4$ exhibits significant gender and racial bias — with black candidates and women...
candidates ranked at the bottom. In contrast, r3 has relatively even gender and race distributions. In fact, r3 appears to be the only base ranking that does not have a significant preference for men candidates. For their decision making, the committee must ensure their resulting consensus ranking is unbiased with respect to the applicants’ Race and Gender.

For this, the committee must define what constitutes fair treatment in this setting. This is a challenge, as group fairness [10] in rankings with multiple protected attributes is largely unexplored. If the committee were to only consider Race and Gender independently, will this also mitigate intersectional bias? Intersectional bias [11], introduced by legal scholar Kimberle Crenshaw, refers to how identities compound to create structures of oppression, and is defined as the combination of multiple protected attributes. Attempting to treat each protected attribute individually can provide the appearance of fair treatment (e.g., Gender and Race independently), but risks neglect of mitigating intersectional group bias. For example, black women (frequently at the bottom of r1, r2 and r4) may not receive fair treatment in the consensus ranking. Complicating matters further, individual protected attributes are often independently protected from discrimination via labor laws [12] and civil rights legislation [3], [13], [14]. Thus, if the committee were to achieve intersectional fairness, it is unclear if this would also provide the necessary fair treatment of Gender and Race groups?

Most importantly, the committee is obligated to certify that their final ranking is bias free, that may require a consensus ranking that is fairer than any of the individual rankings. The above challenges demand a computational strategy capable of supporting the committee in achieving the fair treatment across multiple protected attributes while also reflecting committee preferences in the fair consensus.

State-of-the-Art and Its Limitations. Recent work addressing group fairness with multiple protected attributes focused exclusively on classification and prediction tasks [15]–[17]. Prior work on group fairness for (single) rankings has addressed only one single facet of fairness, either only one protected attribute [18]–[22] or only the intersection of attributes [23]. Further, in most cases only the restricted case was studied, where this single protected attribute is limited to be binary – i.e., only two groups [18], [20], [23], [25].

Numerous algorithmic strategies exist for combining base rankings into a good consensus ranking [26]–[30] – a task known to be NP-hard in general [27], [29]. However, all but one do not consider fairness. The exception is the recent work by Kuhlman and Rundensteiner [25] on fair rank aggregation, which is limited to providing fair consensus rankings only for the restricted case when there is one binary protected attribute. With this restriction, Race would have to be encoded as a binary value, such as {white, non-white} as opposed to any number of racial categories; while all other protected attributes such as Gender would have to be disregarded.

Our Proposed Approach. In this work, we formulate and then study the problem of multi-attribute fair consensus ranking, namely MFCR, in which we aim to create a fair consensus ranking from a set of base rankings over candidates defined by multiple and multi-valued protected attributes. The MFCR problem seeks to satisfy dual criteria – (1.) that all protected attributes and their intersection satisfy a desired level of group fairness and (2.) that the consensus ranking represents the preferences of rankers as expressed by the base rankings. We formulate the preference criteria of the MFCR problem through a new measure called Pairwise Disagreement loss, which allows us to quantify the preferences of rankers not represented in the consensus ranking. We operationalize the group fairness criteria of the MFCR problem through interpretable novel fairness criteria we propose called Multiple Attribute and Intersectional Rank group fairness (or short, MANI-RANK).

Our formulation of MANI-Rank fairness corresponds to an interpretable unified notion of fairness for both the protected attributes and their intersection. Further, this innovation empowers our family of proposed MFCR algorithms to tune the degree of fairness in the consensus ranking via a single parameter. Thus, our MFCR algorithms achieve the desired level of fairness even among base rankings that may be deeply unfair. More precisely, our optimal MFCR algorithm called Fair-Kemeny elegantly leverages our proposed formulation of MANI-Rank as constraints on the exact Kemeny consensus ranking [26]. We further design a series of polynomial-time MFCR solutions based on the efficient consensus generation methods Copeland [31], Schulze [32], and Borda [33].

We conduct a comprehensive experimental study comparing our proposed MFCR solutions against a rich variety of alternate consensus ranking strategies from the literature (both fairness-unaware and those that we make fairness-aware). We demonstrate that our solutions are consistently superior for various agreement and fairness conditions. Our experiments also demonstrate the scalability of our proposed methods for large consensus generation problems. Lastly, we showcase our MFCR solutions in removing bias in a real-world case study involving student rankings for merit scholarships [34].

Contributions. Our contributions include the following:

- We formulate open multi-attribute fair consensus ranking (MFCR), unifying the competing objectives of bias mitigation and preference representation by adopting a unified pairwise candidate disagreement model for both.
- We design the MANI-Rank group fairness criteria, that for the first time interpretably captures both protected attribute and intersectional group fairness for rankings over candidates with multiple protected attributes.
- We develop a series of algorithms from MFCR-optimal Fair-Kemeny to polynomial-time Fair-Copeland, Fair-Schulze, and Fair-Borda for efficiently solving the open problem of multi-attribute fair consensus ranking.
- Our extensive experimental study demonstrates both the efficacy and efficiency of our algorithms, along with the ability to produce real-world fair consensus rankings. We illustrate that only our proposed methodology achieves both group fairness and preference representation over a vast spectrum of consensus problems.
II. Multi-Attribute Fair Consensus Ranking Problem

A. Preliminaries

Protected Attributes. In our problem setting, we are given a database $X$ of $n$ candidates, $x_i \in X$. We assume that each candidate is described by attributes including several categorical protected attributes, such as, gender, race, nationality, or age. The set of protected attributes is denoted by $\mathcal{P} = \{p_1, \ldots, p^q\}$, with $q$ protected attributes. Each protected attribute, say $p_j$, draws values from a domain of values $\text{dom}(p_j) = \{v_{j1}, \ldots, v_{jq}\}$, with the domain size denoted by $|\text{dom}(p_j)|$, or, in short, $|p_j|$. For instance, the domain of the $k$-th protected attribute $Gender$ is composed of the three values $\text{Man, Woman, Non-binary}$.

The protected attributes $\mathcal{P}$ combined in a Cartesian product $\text{Inter} = p_1 \times \ldots \times p^q$ forms what we call an intersection [11].

The cardinality of the intersection, i.e., the number of all its possible value combinations, is $|\text{Inter}| = |p_1| \times \ldots \times |p^q|$. We denote the value of candidate $x_i$’s $k$-th protected attribute by $p_k(x_i)$ and their intersection value as $\text{Inter}(x_i)$.

For each value $v^k_j$ of a protected attribute $p^k_j$, there is a protected attribute group composed of all candidates $x_i \in X$ who all have the same value $v^k_j$ for the protected attribute $p^k_j$.

Definition 1 (Protected Attribute Group) Given a candidate database $X$ and a value $v^k_j$ for a protected attribute $p^k_j$, the protected attribute group for value $v^k_j$ is:

$$G_{(p^k_j; v^k_j)} = \{x_i \in X : p^k_j(x_i) = v^k_j\}.$$

For brevity, we refer to the protected attribute group $G_{(p^k_j; v^k_j)}$ by $G(k;j)$, when possible without ambiguity. For instance, $G(\text{Gender}; \text{Woman})$ denotes the group of all women in $X$. As candidates in $X$ are defined by their intersectional identity, an intersectional group then corresponds to all candidates $x_i \in X$ who share the same values across all protected attributes.

Definition 2 (Intersectional Group) Given a candidate database $X$ and values $v^1_j, v^2_j, \ldots, v^q_j$ for protected attributes set $\mathcal{P}$, the intersectional group for values $v^1_j, v^2_j, \ldots, v^q_j$ is:

$$\text{Inter}(G(p^1_j; v^1_j), \ldots, (p^q_j; v^q_j)) = \{x_i \in X : (p^1_j(x_i) = v^1_j) \land \ldots \land (p^q_j(x_i) = v^q_j)\}.$$

For brevity, a group of candidates in $X$ sharing an intersection value, the $j$-th intersection value, is denoted $\text{Inter}G_j$.

Notion of Group Fairness. In this work, we aim to achieve the group fairness notion of statistical parity [10]. First proposed in binary classification [10], and more recently the focus of fair learning-to-rank methods for a binary protected attribute [18], [19], [22], [24]. statistical parity is a requirement stipulating candidates must receive an equal proportion of the positive outcome regardless of their group membership in a protected attribute.

Definition 3 (Statistical Parity) Given a dataset $X$ of candidates sharing different values of protected attribute $p$, and a binary classifier $\hat{Y}$ with positive outcome $\hat{Y} = 1$. The predictions of $\hat{Y}$ are fair with respect to $p$ if candidates with different values of $p$ have a $\text{Prob}(\hat{Y} = 1)$ directly proportional to their share of $X$.

Base Rankings Model User Preferences. Our problem contains $m$ rankers — where rankers (human, algorithmic, or some combination thereof) express preferences over the database $X$ of candidates. Each ranker’s preferences over $X$ are expressed via a ranking. A ranking is a strictly ordered permutation $\pi = [x_1 \prec x_2 \prec \ldots \prec x_n]$ over all candidates in $X$. Here, $x_i \prec x_j$ implies candidate $x_i$ appears higher in the ranking $\pi$ than $x_j$, where $1$ is the top or best ordinal rank position and $n$ the least desirable. The collection of all possible rankings over the database of $X$ of $n$ candidates, denoted $S_n$, corresponds to the set of all possible permutations of $n$ candidates. The $m$ rankings produced by the $m$ rankers creates a set $R \subseteq S_n$, which we refer to as base rankings.

Consensus Ranking. From the base rankings $R$, we wish to generate a ranking that represents the preferences of the rankers, namely, a consensus ranking. A consensus ranking $\pi^C$ is the ranking closest to the set of base rankings $R$, such that a distance function $\text{dist}$ is minimized. Finding a consensus ranking corresponds to traditional rank aggregation task [33].

Definition 4 (Consensus Ranking) Thus, a consensus ranking $\pi^C$ from a set of base rankings $r_i \in R$ is defined by:

$$\pi^C = \arg\min_{\pi \in S_n} \sum_{r_i \in R} \text{dist}(\pi, r_i).$$

Multi-attribute Fair Consensus Ranking (MFCR). Our problem of creating a fair consensus ranking from many rankers’ preferences over candidates defined by multiple and multi-valued protected attributes has two components. The fairness component aims to treat all candidates equally regardless of group membership in protected groups (Definition 1) or intersectional groups (Definition 2). In order for the fairness criteria of our problem to be meaningful in practice (i.e., in the form of targets such as the “80%” rule championed by US Equal Employment Opportunity Commission (EEOC) [36]), we define “fair” as an application-specified desired level of fairness in the consensus ranking.

In Section II-B, we propose novel group fairness criteria which we integrate into the MFCR problem (Definition 11) as a concrete target (i.e, the application can select a proximity to perfect statistical parity [29]). Thus, the MFCR problem encompasses the creation of consensus rankings that may need to be fairer than the fairest base ranking.

The preference representation component of our problem ensures that all rankers see their preferences reflected in the fair consensus ranking in as much as is possible, and thus they can be expected to accept the consensus ranking. We note that even inside biased base rankings critical preference information is encoded, such as the orderings of candidates within the same group. Thus in Section II-C, we propose a new measure to quantify the preferences of the base rankings $R$ that are not represented in the fair consensus ranking, called PD loss. We integrate this measure into the MFCR problem.
(Definition 11) as criteria to be minimized in the generation of a fair consensus ranking. Doing so ensures that a fair consensus ranking does not prioritize certain rankers above others.

B. Proposed Group Fairness Criteria for Multi-attribute Fair Consensus Ranking (MFCR)

Proposed Group Positive Outcome Metric. To operationalize the group fairness component of our problem (Section II-A), we design a measure for capturing how fairly a group is being ranked. Our insight here is to define fair treatment of a group via a constant value, thus making the interpretation across group sizes comparable.

This allows us to define statistical parity (Definition 3) based on all groups having this same value, thus formulating group fairness for multiple and multi-valued protected attributes into a single easy-to-understand measure.

For many applications, the entire ranking matters — bottom ranked candidates may lose out on consequential outcomes such as funding, resource and labor divisions, or decreased scholarships, if they were placed even somewhat lower. To capture how a group is treated through a ranking we utilize pairwise comparisons of candidates. Intuitively, the more pairs a candidate is favored in, the higher up in the ranking the candidate appears. Any ranking \( \pi \) over \( n \) candidates can be decomposed into pairs of candidates \( (x_i, x_j) \) where \( x_i \prec \pi x_j \). The total number of pairs in a ranking over a database \( X \) of \( n \) candidates is:

\[
\omega(X) = (n(n - 1))/2. \tag{2}
\]

As statistical parity requires groups receive an equal proportion of the positive outcome [10], we cast positive outcome as being favored in a mixed pair — where mixed refers to candidates in a pair associated with two distinct protected groups according to protected or intersection attributes. For instance, a pair comparing two woman candidates is not a mixed pair, while a pair with a man and woman is a mixed pair.

We denote the number of mixed pairs for a group \( G \) (where \( G \) is a protected attribute group \( G_{k,j} \) or intersectional group \( \text{Inter} G_j \)) in a ranking over \( |X| \) candidates with \( |G| \leq |X| \) as:

\[
\omega_M(G, \pi) = |G|(|X| - |G|). \tag{3}
\]

The total number of mixed pairs for a protected attribute \( p^k \) or intersection \( \text{Inter} \) in a ranking over \( |X| \) candidates is:

\[
\omega_M(X, \pi) = \omega(X, \pi) - \sum_{G_{*,i} \in X} \omega(G_{*,i}, \pi), \tag{4}
\]

where \( *=k \) or \( *=\text{inter} \) for the respective attribute.

We design our measure of positive outcome allocation, called a group’s Favored Pair Representation (FPR) score.

Definition 4 (Favored Pair Representation) The \( \text{FPR} \) for a ranking \( \pi \) over candidate database \( X \) for a group of candidates \( G \subset X \), where \( G \) is either \( G_{k,j} \) or \( \text{Inter} G_j \), is:

\[
\text{FPR}_G(\pi) = \sum_{x_i \in G} \sum_{x_1 \not\in G} \frac{\text{countpairs}(x_i, x_1)}{\omega_M(G, \pi)}.
\]

One critical property of this \( \text{FPR} \) score is that it is easy to explain and interpret. It ranges from 0 to 1. When \( \text{FPR} = 0 \), the group is entirely at the bottom of the ranking. When \( \text{FPR} = 1 \), the group is entirely at the top of the ranking. By design, when \( \text{FPR} = 1/2 \), the group receives fair treatment in the ranking, i.e., a directly proportional share of favored rank positions.

Next, we assure that our \( \text{FPR} \) measure handles groups defined by multi-valued attributes. For that, we put forth that when the attribute has multiple values, we must divide by the number of mixed pairs containing that specified group (Equation (3)) as opposed to the total number of mixed pairs in the ranking (Equation (4)). Our design unlike prior work [19], [25] guarantees the critical property of the \( \text{FPR} \) measure that \( 1/2 \) is a fair positive outcome allocation even for multi-valued attributes groups.

We observe that the \( \text{FPR} \) measure allows us to compare the fair treatments of groups of different sizes purely based on their \( \text{FPR} \) scores. Better yet, when all groups receive a proportional share of the positive outcome (i.e., \( \text{FPR} = .5 \)) then all groups receive an equal proportion of the positive outcome, thus satisfying statistical parity. Thus, the \( \text{FPR} \) metric elegantly allows us to check for perfect statistical parity simply by having an \( \text{FPR} \) score of 0.5 for all groups.

Proposed Unified Multi-Attribute Group Fairness Criteria. We now propose our formal definition of the group fairness criteria of our MFCR problem. Our key design choice is to specify group fairness at the granularity of the attribute — as opposed to fairness at the group level. In this way, we provide the ability to tune the degree of fairness in every protected attribute and in the intersection. Intentionally, we do not design criteria per group. As in the multiple protected attribute setting, even a handful of attributes and their intersection can create a huge number of groups that otherwise would have to be individually parameterized and interpreted.

For the group fairness property of MFCR, we introduce a new measure to quantify statistical parity for protected attributes \( p^k \) as described below.

Definition 5 (Attribute Rank Parity) The Attribute Rank Parity (ARP) measure for the \( k \)-th protected attribute \( p^k \) in ranking \( \pi \) over candidate database \( X \) is:

\[
\text{ARP}_{p^k}(\pi) = \operatorname{argmax}_{(G_{k,i}, G_{k,j}) \in X} |\text{FPR}_{G_{k,i}}(\pi) - \text{FPR}_{G_{k,j}}(\pi)|
\]

This ARP measure simplifies the task of tuning the degree of fairness for a protected attribute. Namely, when the \( \text{ARP}_{p^k} = 1 \), then the protected attribute is maximally far from statistical parity. Meaning, one group corresponding to a value in the domain \( p^k \) is entirely at the top of the ranking, while a second group is entirely at the bottom of the ranking. When \( \text{ARP}_{p^k} = 0 \), perfect statistical parity is achieved for attribute \( p^k \).

Similarly, we now formulate intersectional fairness, which corresponds to criteria ii of our problem.

Definition 6 (Intersectional Rank Parity) The Intersection Rank Parity (IRP) measure for the intersection \( \text{Inter} \) de-
termined from the set of protected attributes $\mathcal{P}$ in ranking $\pi$ over candidate database $X$ is:

$$\text{IRP}(\pi) = \arg\max_{\forall (\text{Inter}_G_i, \text{Inter}_G_j) \in X} |\text{FPR}_{\text{Inter}_G_i}(\pi) - \text{FPR}_{\text{Inter}_G_j}(\pi)|$$

We now have two easy-to-use and interpretable measures $\text{ARP}$ and $\text{IRP}$ that directly map to the fair treatment of protected attribute groups (Definition 1) and intersectional groups (Definition 2) in the MFCR problem. We unify these objectives into one fairness notion, which we call Multiple Attribute and Intersectional Rank (MANI-Rank) group fairness. MANI-Rank is applicable to consensus and to single rankings alike over candidate databases with multiple protected attributes.

We introduce the threshold parameter $\Delta$ for fine-tuning a desired degree of fairness. $\Delta$ represents the desired (or required) proximity to statistical parity. Equation (5) below models this for every protected attribute and Equation (6) for the intersection. This is carefully designed to be easy to interpret, namely, when $\Delta$ is close to 0, the ranking approaches statistical parity for all protected attributes and intersection.

**Definition 7 (Multiple Attribute and Intersectional Rank — MANI-Rank)** $\text{MANI-Rank Group Fairness}$ for rankings of candidates with multiple protected attributes is defined as:

$$\text{ARP}_{p^k}(\pi) \leq \Delta \ (\forall p^k \in \mathcal{P})$$

$$\text{IRP}(\pi) \leq \Delta$$

Per design, the MANI-Rank fairness criteria result in perfect protected attribute and intersectional statistical parity, when $\Delta = 0$. This follows from definitions of $\text{ARP}_{p^k}$ and $\text{IRP}$.

**Customizing Group Fairness.** In most real-world settings, equal degrees of fairness for protected attributes and their intersection is desirable, which we model by choosing the same degree of fairness threshold $\Delta$ in Definition 7. However, our MANI-Rank criteria allows for applications to set up different thresholds tailored to each protected attribute ($\Delta_{p^k}$) or the intersection ($\Delta_{\text{Inter}}$). Additionally, Definition 7 can be modified to handle alternate notions of intersection by adjusting the intersectional groups (Definition 2) to be a desired subset of protected attributes (as opposed to the combination of all protected attributes). Likewise, Definition 7 can be extended to support specific subsets of protected attribute combinations by adding to Definition 7 an additional equation for every desired subset of protected attributes, such as $\text{IRP}_{\text{subset of } \mathcal{P}}(\pi) \leq \Delta$. We note, as evidenced by our empirical study in Section V-A, in order for a protected attribute or intersection to be guaranteed protected at a desired level of fairness it must be constrained explicitly.

**C. Proposed Representation Criteria for Multi-Attribute Fair Rank Aggregation**

We also propose to model the degree to which the consensus ranking captures the preferences of all the rankers corresponding to the preference representation component of MFCR. For this, we measure the distance between the base rankings $R$ and the fair consensus ranking $\pi^{C*}$. We intentionally select pairwise disagreements as a distance measure because this allows us to interpretably measure how many preferences, i.e., candidates comparisons, are not met in the fair consensus ranking. We propose to do this by summing the Kendall Tau distances between $\pi^{C*}$ and every base ranking. The Kendall Tau distance in Definition 8 is a distance measure between rankings. It counts the number of pairs in one ranking that would need to be inverted (flipped) to create the second ranking, thereby counting pairwise disagreements between two rankings.

**Definition 8 (Kendall Tau)** Given two rankings $\pi_1, \pi_2 \in S_n$, the Kendall Tau distance between them is:

$$\text{dist}_{\text{KT}}(\pi_1, \pi_2) = |\{(x_i, x_j) \in X : x_i \preceq_{\pi_1} x_j \text{ and } x_j \preceq_{\pi_2} x_i\}|.$$ 

Then we normalize the pairwise disagreement count by the total number of candidate pairs represented in the base rankings $R$. This leads us to Definition 9 of our proposed preference representation metric Pairwise Disagreement loss (PD loss).

**Definition 9 (Pairwise Disagreement Loss)** Given a set of base rankings $R$ and a consensus ranking $\pi^C$, the pairwise disagreement loss (PD loss) between $\pi^*$ and $R$ is:

$$\text{PD Loss}(R, \pi^C) = \frac{\sum_{r_i \in R} \text{dist}_{\text{KT}}(\pi^C, r_i)}{n! \omega(X, \pi^C) * |R|}$$

We now have a measure that expresses how many preferences of rankers are not captured in a fair consensus ranking. By design, PD loss is between the values of 0 and 1, with 0 offering the interpretation that every pairwise preference in the base rankings is represented in the fair consensus ranking (i.e., all the base rankings are the same and match exactly the fair consensus ranking) and PD losses 1 when no single pairwise preference in the base rankings is represented in the fair consensus ranking.

**D. MFCR Problem: Multi-Attribute Fair Consensus Ranking**

Pulling together our proposed group fairness and preference representation models, we now are ready to formally characterize our fair consensus ranking problem.

**Definition 10 (Multi-Attribute Fair Consensus Ranking - MFCR)** Given a database of candidates $X$ defined by multiple and multi-valued protected attributes $\mathcal{P}$, a set of base rankings $R$, and proximity to statistical parity parameter $\Delta$, the multi-attribute fair consensus ranking (MFCR) problem is to create a consensus ranking $\pi^{C*}$, that meets two criteria:

- **MFCR group-fair criteria:**
  - satisfies MANI-Rank group fairness (Definition 7) subject to parameter $\Delta$, and,

- **MFCR pref criteria:**
  - minimizes PD loss (Definition 9) between $R$ and $\pi^{C*}$. Our problem formulation emphasizes the dual ability to specify a desired level of group fairness in the consensus ranking while minimizing unrepresented ranker preferences in the consensus ranking. We note that, by design, this allows for
applications to create a consensus ranking fairer than all the base rankings by setting a low $\Delta$ parameter.

**MFCR Applied to Motivation Admissions Example.** Returning to the task faced by the Admissions Committee from Section 1, we demonstrate how the committee can apply the MFCR problem to create a fair consensus ranking of applicant candidates. Figure 2 illustrates the consensus ranking generated for admissions decisions with and without MANI-Rank group fairness. The ranking 2(a) is determined from premininent consensus ranking method, Kemeny [26]. We see that it exhibits significant bias with respect to Gender as men are clustered at the top. Likewise, the intersectional bias is substantial due to white men being significantly advantaged in the ranking. In contrast, ranking 2(b), generated with MANI-Rank $\Delta = .1$, has ARP and IRP scores nearly at perfect statistical parity - indicating the promise of our proposed formulation to remove Gender, Race and intersectional bias. Apply the MFCR framework helps the admissions committee to create a fair consensus ranking; as otherwise biases present in base rankings would be reflected in the final applicant ranking.

**III. ALGORITHMS FOR SOLVING MULTI-ATTRIBUTE FAIR CONSENSUS RANKING**

We propose a family of algorithms for solving the MFCR problem. For specific algorithms, we utilize an a precedence matrix representation $W$ of the base rankings $R$ that captures all pairwise relationships in $R$. Put differently, $W$’s entries represent pairwise disagreements.

**Definition 11 (Precedence Matrix)** Given a database $X = \{x_1, x_2, \ldots, x_n\}$ of candidates, and set of base rankings $R$, the precedence matrix $W = [W_{x_a,x_b}]_{a,b=1,\ldots,n}$ is defined by:

$$W_{x_a,x_b} = \sum_{r \in R} 1(x_b \prec_r x_a)$$

where $1$ is the indicator function, namely, equal to 1 when $x_b \prec_r x_a$, 0 otherwise.

\[\text{Algorithm 1: Fair-Kemeny}\]

$$\text{Minimize } \sum_{(x_a,x_b) \in X} Y_{(x_a,x_b)} W_{(x_a,x_b)}$$

subject to: $Y_{(x_a,x_b)} \in \{0,1\}$

$$Y_{x_a,x_b} + Y_{x_b,x_a} = 1, \forall x_a, x_b$$

$$Y_{x_a,x_b} + Y_{x_b,x_c} + Y_{x_c,x_a} \leq 2, \forall x_a, x_b, x_c$$

$$\sum_{x_a \in G_{ki}, x_b \notin G_{ki}} (\frac{1}{\omega_m(G_{ki})} Y_{x_a,x_b}) - \sum_{x_c \in G_{kj}, x_d \notin G_{kj}} (\frac{1}{\omega_m(G_{kj})} Y_{x_c,x_d}) \leq \Delta,$$

$$\forall (G_{ki}, G_{kj}) \in X, \forall p^k \in \mathcal{P}$$

$$\sum_{x_c \in \text{Inter}G_i, x_d \notin \text{Inter}G_i} (\frac{1}{\omega_m(\text{Inter}G_i)} Y_{x_c,x_d}) - \sum_{x_c \in \text{Inter}G_j, x_d \notin \text{Inter}G_j} (\frac{1}{\omega_m(\text{Inter}G_j)} Y_{x_c,x_d}) \leq \Delta,$$

$$\forall (\text{Inter}G_i, \text{Inter}G_j) \in X$$

In the formulation of Fair-Kemeny above, matrix $Y$ specifies the pairwise order of candidates in the consensus ranking $\pi^C$ and matrix $W$ represents the precedence matrix from Definition 11. The objective function in Equation (7) formulates the Kemeny criteria, minimizing the number of pairwise disagreements between base rankings $R$ and $\pi^C$.

As shown in Conitzer et al. [29], the first three constraints, Equations (8), (9), (10) enforce a valid ranking (no cycles, not multiple candidates in the same position, or no invalid pairwise orderings).

Next, our formulation of MANI-Rank group fairness is modeled by the constraints in Equation (11) and Equation (12) which enforces group fairness for every protected attribute.
as well as intersectional group fairness, respectively. These constraints leverage our formulation of ARP (Definition 5) and IRP (Definition 6) constraining the maximum absolute difference in FPR scores for all groups representing values in the $\text{dom}(p^k)$ for each protected attribute and in $\text{dom}(\text{Inter})$.

**Complexity of Fair-Kemeny Solution.** General (fairness unaware) Kemeny is an NP-hard problem [26], [29], though tractable in smaller candidate databases. Our Fair-Kemeny method inherits this complexity. Our MANI-Rank criteria formulated via Equation (11) adds $\left(\text{dom}(p^k)\right)^2$ and Equation (12) adds $\left(\text{dom}(\text{Inter})\right)^2$ constraints. Yet, our empirical study in Section IV-D confirms that in practice the runtime of Fair-Kemeny is not substantially greater than that of the traditional Kemeny, and both can handle thousands of base rankings.

**B. Fair-Copeland, Fair-Schulze, and Fair-Borda**

Aiming to handle larger candidates databases than Fair-Kemeny or Kemeny, we now design a series of algorithms utilizing polynomial time consensus generation methods. They all utilize a novel pairwise bias mitigation algorithm we propose, called Make-MR-Fair, specifically designed to efficiently achieve MFCR group-fair criteria while minimizing increases PD loss caused by introducing fairness.

**Make-MR-Fair** takes as input a consensus ranking $\pi^C$ to be corrected. Initially, it determines the FPR and IRP scores of the protected attributes and intersection, checking if MFCR group-fair criteria is satisfied with respect to the $\Delta$ semantics. When this condition is not true, the algorithm determines the attribute (either a protected attribute or the intersection) that has the highest ARP or IRP score. This attribute is now “the attribute to correct”. By honing in on the least fair attribute we aim to minimize pairwise swaps - thus minimizing PD loss. Within the attribute to correct, the group with the highest and lowest FPR score is said to be $G_{\text{highest}}$ and $G_{\text{lowest}}$, respectively.

Within the group $G_{\text{highest}}$, the candidate lowest in the ranking is selected as $x_{G_h}$. Then the ordered mixed pairs of $x_{G_h}$ are searched to determine the first unfavored candidate, $x_{G_l}$, in the list of mixed pairs which belong to $G_{\text{lowest}}$. If there is no $x_{G_l}$ candidate then the $x_{G_h}$ candidate is altered to be the first candidate in $G_{\text{highest}}$ higher than the original $x_{G_h}$ candidate. We intentionally choose to redefine the $x_{G_h}$ candidate as opposed to $x_{G_l}$. The reason is to enforce repositioning candidates into positions at the top of the ranking, making fewer, but more impactful swaps. This helps to minimize the increase in PD loss caused by this fairness mitigation process. When the pair $x_{G_h} \prec x_{G_l}$ is found, the two candidates are swapped - resulting in $x_{G_l} \prec x_{G_h}$. Each pair swap is guaranteed to lower the FPR score of $G_{\text{highest}}$ and increase the FPR score of $G_{\text{lowest}}$, thus increasing proximity to statistical parity for the attribute to correct. The algorithm terminates once all protected attributes and the intersection are below $\Delta$, and returns the corrected ranking $\pi^C$ as fair consensus ranking $\pi^{C*}$.

**Algorithm 2: Make-MR-Fair**

**Input:** consensus ranking $\pi^C$, candidate database $X$, thresholds $\Delta$

**Output:** fair consensus ranking $\pi^{C*}$

```python
for all $p^k \in P$ and inter do
    atr = $p^k \in P$ or Inter with max IRP/ARP
    // entity to correct
    $G_{\text{highest}}$ = group of atr with max FPR score
    $G_{\text{lowest}}$ = group of atr with min FPR score
    $x_{G_h}$ = lowest $x_i \in G_{\text{highest}}$
    if $(x_{G_h} \prec x') \in \omega_m(x_{G_h})$ s.t.
        $x'$ is the highest $x' \in G_{\text{lowest}}$
        /* find next highest $x_{G_h}$ */
        else
            $x_{G_h}$ = next lowest $x_i \in G_{\text{highest}}$ s.t.
            $(x_i \prec x') \in \omega_m(x_i)$ s.t. $x'$ is the highest
            $x' \in G_{\text{lowest}}$
            $x_{G_l}$ = is the highest $x' \in G_{\text{lowest}}$
            swap $x_{G_l}$ and $x_{G_h}$
    $\pi^{C*} = \pi^C$
    return $\pi^{C*}$
```

IRP score. Each ARP and IRP score is calculated with one traversal of $\pi^C$ to determine the FPR scores that ARP and IRP are calculated from. Thus each score computation costs $O(n)$ work. This work is done before each swap. In the worst case, the algorithm would flip $\omega(X) = n * (n - 1)/2$ pairs. Thus, assuming $|P|$ protected attributes and one intersection, the resulting worst case runtime is $O(n^2 * (|P| + 1) * n)$. Additionally, the runtime can be reduced by adjusting the $\Delta$ parameter as will be empirically illustrated in Section IV-D.

**Fair-Copeland Solution for MFRA.** We create Fair-Copeland based on Copeland [31], because the later is the fastest (polynomial) pairwise consensus rank generation method [39]. Copeland [31] creates a consensus ranking that ranks candidates in descending order by the number of pairwise contests a candidate has won (a tie is considered a "win" for each candidate). Intuitively, this can be understood through our precedence matrix $W$, where the number of pairwise contests candidate $x_b$ wins is $\sum_{x_a \in X} \mathbb{1}(W_{x_a,x_b} \geq W_{x_b,x_a})$. Our Fair-Copeland method satisfies MFCR pref criteria by producing the Copeland consensus, ordering candidates by descending $\Delta$ parameter as will be empirically illustrated in Section IV-D.

**Complexity of Fair-Copeland.** Fair-Copeland takes $O(n^2 * |R|)$ to create the precedence matrix $W$, $O(n^2)$ to check $W$ to determine the winners of pairwise contests, $O(n\log n)$ time to sort the candidates by the number of contests won, and in the worst case Make-MR-Fair takes $O(n^2 * (|P| + 1) * n)$ to fairly reorder candidates.
Fair-Schulze Solution for MFCR. We introduce Fair-Schulze as it is polynomial-time, Condorcet, and pairwise like Fair-Copeland, and in addition its fairness-unaware method is extremely popular in practice. Schulze is used to determine multi-winner elections in over 100 organizations worldwide (such as Wikimedia Foundation to elect a board of Trustees, Political Parties, Gentoo, Ubuntu, and the Debian Foundation, see [32] for a full list). The Schulze [32] method generates a consensus ranking via pairwise comparisons which naturally helps address MFCR pref criteria.

The Shulze rank aggregation method first determines the precedence matrix \( W \). Then \( W \) is treated as a directed graph with weights representing the pairwise agreement counts between every pair of candidates. Next, the algorithm computes the strongest paths between pairs of candidates by means of a variant of the Floyd Warshall algorithm [32], [31]. Here the strength of a path between candidates is the strength of its lowest weight edge. Schulze then orders candidates by decreasing strength of their strongest paths. Our Fair-Schulzemethod satisfies MFCR pref criteria by producing the Schulze consensus ranking, ordering candidates by the strength of path - thereby optimizing for pairwise agreement, then correcting it to satisfy MFCR group-fair criteria with \( \Delta \) desired fairness using Make-MR-Fair. The algorithm’s pseudocode is provided in the supplement [40].

Fair-Schulze Complexity. Fair-Schulze takes \( O(n^2 \log |R|) \) to create the precedence matrix \( W \), \( O(n^2) \) to compute strongest paths to create a Schulze ranking [32], [41], \( O(n\log n) \) time to sort the candidates into correct order, and in the worst case \( O(n^2 \times (|P| + 1) \times n) \) to fairly reorder candidates.

Fair-Borda Solution for MFCR. We create Fair-Borda specifically as an MFCR solution for large consensus ranking problems. In a comparative study of Kemeny consensus ranking generation, Borda [33] was shown to be the fastest Kemeny approximation method [27]. Thus it is an ideal strategy for tackling our MFCR problem. More precisely, Borda [33] is a positional rank aggregation function that ranks candidates in descending order based on a total number of points allotted to each candidate. The total points allotted to candidate \( x_i \) correspond to the total number of candidates ranked lower than \( x_i \) in all the base rankings. Our Fair-Borda method utilizes Borda to efficiently aggregate the base rankings with minimal error by minimizing pairwise disagreement with the base rankings. Thus, it satisfies the MFCR pref criteria criteria.

Next, the Make-MR-Fair subroutine is applied to the resulting Borda ranking so that it satisfies MFCR group-fair criteria with parameter \( \Delta \). The algorithm’s pseudocode is found in the supplement [40].

Complexity of Fair-Borda. Fair-Borda takes \( O(n \times |R|) \) time to determine the points per candidates, \( O(n\log n) \) time to order the candidates by total points, and \( O(n^2 \times (|P| + 1) + n) \) to fairly reorder candidates.

C. Price of Fairness Measure for Fair Rank Aggregation

Satisfying MFCR group-fair criteria incurs a toll in terms of MFCR pref criteria. Intuitively, this toll is greatest when the base rankings \( R \) have a low degree of fairness and the \( \Delta \) parameter requires a high degree of fairness. Thus, we now design the concept of Price of Fairness (PoF) as a metric to quantify the cost of MFCR group-fair criteria as an increase in PD loss from the consensus ranking satisfying only MFCR pref criteria. We compute PoF as the difference in the PD loss of the fair consensus ranking \( \pi^C \) and PD loss of the fairness unaware consensus ranking \( \pi^C \):

\[
PoF = PD\ Loss(R, \pi^C) - PD\ Loss(R, \pi^C) \quad (13)
\]

We note that \( PoF \) is always \( \geq 0 \) as the fair consensus ranking at \( best \) represents the preferences of the base rankings equally as well as the fairness-unaware consensus ranking. As is true of the fairness and utility tradeoff in general [18], [24], we note here and observe in our experiments that \( PoF \) for the MFCR problem can be significant.

TABLE I: Mallows datasets: \(|R| = 150\) base rankings over 90 candidates, with 6 candidates in each of 15 intersectional groups from \( \text{dom}(\text{Race}) = 5 \) and \( \text{dom}(\text{Gender}) = 3 \).

| Mallows Dataset | Fairness metrics on modal ranking | ARP_Male | ARP_Female | ITP |
|----------------|----------------------------------|---------|------------|-----|
| Low-Fair       | 0.70                             | 0.70    | 1.00       |
| Medium-Fair    | 0.50                             | 0.50    | 0.75       |
| High-Fair      | 0.30                             | 0.30    | 0.54       |

IV. EXPERIMENTAL STUDY

Experimental Methodology. We conduct a systematic study of alternate group fairness approaches evaluating our MFCR methods against baselines under a rich variety of conditions in the base rankings modeled using the Mallows model [27], [42]. In particular, we analyze how the degree of consensus and fairness present within the base rankings along with the \( \Delta \) parameter affect the PoF of the consensus ranking. We also study the scalability of our MFCR solutions. We conclude with a case study on student rankings and merit scholarship [34].

A. Studying Alternate Group Fairness Constraints

Datasets. We leverage the Mallows Model [27], [42], a probability distribution over rankings, as a data generation procedure. Widely used to evaluate consensus ranking tasks [19], [27], [43], the Mallows Model is an exponential location-spread model, in which the location is a modal ranking among a set of base rankings, denoted by \( \pi' \), and the spread parameter, denoted by \( \theta \), is a non-negative number. For a ranking \( \pi' \in S_n \), the Mallows model is the following probability distribution:

\[
P(\pi') = \frac{e^{\exp(-\theta \cdot d(\pi', \pi))}}{\psi(\theta)} \quad (14)
\]

where \( \psi(\theta) \) is the normalization factor that depends on the spread parameter \( \theta \), and has a closed form. Utilizing the Kendall-Tau distance as the distance metric \( d(\pi', \pi) \), the Kemeny consensus ranking corresponds to the maximum likelihood estimator of the Mallows model [44]. We control the fairness of base rankings by setting the fairness in the modal ranking, consensus is adjusted via the \( \Delta \) parameter.
When $\theta = 0$, there is no consensus among the base rankings around the modal ranking $\pi'$. As $\theta$ increases, the base rankings gain consensus around the modal ranking $\pi'$. 

**Experimentally Comparing MANI-Rank Criteria with Alternate Group Fairness Criteria.** To evaluate group fairness notions, we compare three alternate group fairness approaches in rank aggregation with traditional Kenny rank aggregation (Figure 3). As depicted in Figure 3 we vary the spread parameter $\theta$ to create four sets of base rankings with different degrees of consensus for each modal ranking.

Group fairness strategies we compare include: only constraining the protected attributes – our Fair-Kemeny with Equation (12) removed, only constraining the intersection – our Fair-Kemeny with Equation (11) removed, and our proposed Fair-Kemeny. In all cases, we set our desired fairness threshold $\Delta = .1$ to specify close proximity to statistical parity as per Definition 7.

In Figure 3, we observe that, under all fairness conditions and consensus scenarios ($\theta$), the Kemeny fairness unaware aggregation method creates a consensus ranking with ARP/IRP scores significantly above the desired threshold. The protected attribute-only approach consistently results in consensus rankings with Gender and Race at or below the threshold. But it still leaves IRP significantly higher than desired. The intersection only approach successfully constrains the intersection to the desired fairness level. But it leaves the ARP of Gender higher than $\Delta$. Our proposed MANI-Rank criteria is the only group fairness approach formulation which ensures that both the individual and intersection of protected attributes are at or below the desired level of fairness. Thus, we conclude for a protected attribute or intersection to be guaranteed protected at a desired level of fairness it must be constrained explicitly.

**B. Comparison of MFCR Solutions: Fairness and Preference Representation of Generated Consensus Ranking**

To analyze the efficacy of our proposed MFCR solutions in achieving MFCR group-fair criteria and MFCR preference criteria, we compare our four methods against four baselines in Figure 4. Baselines are (1) traditional Kemeny, (2) Kemeny-Weighted, which orders the base rankings from least to most fair and weights the fairest ranking by $|R|$ and the least fair by 1, (3) Pick-Fairest-Perm a variation of Pick-A-Perm [39] which returns the fairest base ranking, and (4) Correct-Pick-A-Perm which utilizes Make-MR-Fair to correct the finest base ranking to satisfy $\Delta$.

Examining Figure 4 we see that Kemeny and Kemeny-Weighted perform best at representing the base rankings. Pick-Fairest-Perm in the case the base rankings have a high degree of consensus ($\theta$) represents the base rankings as well. However, these methods do not achieve our fairness criteria, i.e., they perform the worst at satisfying MFCR group-fair criteria. While Kemeny does not attempt to incorporate group fairness, Pick-Fairest-Perm and Kemeny-Weighted aim to incorporate fairness. They do not succeed as solutions to the MFCR problem - because they do not achieve a desired level of fairness ($\Delta$ parameter). This results in a consensus ranking that at best represents the fairest ranking in the base set, i.e., Pick-Fairest-Perm indeed returns the fairest among the base rankings. Also, not surprisingly, Kemeny-Weighted is not fairer than Pick-Fairest-Perm ("fairer" defined as lower IRP/ARP scores). The last baseline Correct-Fairest-Perm satisfies MFCR group-fair criteria (due to the utilization of Make Mani-Rank), but with a significantly higher PD loss. This indicates it does not represent the base rankings well, making it a poor MFCR solution.
Next, we examine our proposed methods. It is clear from the ARP and IRP graphs that all our methods achieve the desired level of fairness ($\Delta = 0.1$), thus satisfying MFCR group-fair criteria. When examining PD loss, we see Fair-Kemeny performs best, which intuitively makes sense as it optimally minimizes pairwise disagreements subject to fairness constraints. Next, in order of decreasing PD loss is Fair-Copeland, then Fair-Schulze, followed by Fair-Borda. This is also as expected, as the first two methods are Condorcet methods, and Fair-Borda is not. However, these polynomial-time algorithms compared to Fair-Kemeny perform comparably well in representing the base rankings – this is particularly true when there is less consensus in the base rankings.

C. Studying the Price of Fairness

In Figure 5, we evaluate the Price of Fairness ($PoF$) using the metric from Equation (13). We analyze how the amount of consensus in the base rankings and the $\Delta$ parameter affect $PoF$. Utilizing the Fair-Kemeny method, we observe that the fairness of the modal ranking has the biggest impact on the price of fairness. When the modal ranking has a higher level of fairness, the level of consensus around that ranking does not significantly impact the price of fairness. But when the modal ranking has a very low level of fairness, the degree of consensus ($\theta$) has a larger impact. A low degree of consensus has the effect of “cancelling out” the fairness in the modal ranking. Intuitively, a high degree of consensus around a low fairness modal ranking results in a higher price of fairness.

Next, we examine the effect of the $\Delta$ parameter on $PoF$. We uncover a steep inverse linear trend between $\Delta$ and the $PoF$ for our four methods and Correct-Fairest-Perm on the Low-Fair dataset with $\theta = 0.6$. Across all methods that utilize the $\Delta$ parameter, when $\Delta$ is high, $PoF$ is lower. This is intuitive as when $\Delta$ is small, the consensus ranking is likely required to be fairer than the base rankings. This in turn increases the amount of disagreement between the consensus ranking and the base rankings.

D. MFCR Solutions: Study of Scalability

We evaluate the scalability of all methods presented above. We have implemented our methods in python and used IBM CPLEX optimization software for Kemeny, Fair-Kemeny, and Kemeny-Weighted. All experiments were performed on a Windows 10 machine with 32GB of RAM.

Scalability in Number of Rankers. In Figure 6, we analyze the efficiency of our proposed methods in handling increasingly large numbers of base rankings. We create a Mallows model dataset with a modal ranking with $ARP(Race) = 0.15$, $ARP(Gender) = 0.7$, $IRP = .55$, $dom(Race) = 2$ and $dom(Gender) = 2$, we set ($\theta = .6$), $n = 100$, and specify a tight fairness requirement with $\Delta = .1$.

In Figure 7, we see that three tiers of methods emerge. The fastest tier includes Fair-Borda, Pick-Fairest-Perm, and Correct-Pick-A-Perm, while the second is Fair-Schulze, Fair-Copeland, Fair-Kemeny, and Kemeny. We note that our proposed methods perform no slower than regular (not-fair) Kemeny. Lastly, Kemeny-Weighted performs the slowest due to having to order and weight large numbers of base rankings.

Scalability in Number of Candidates. In contrast to base rankings, large numbers of candidates is a greater challenge for consensus generation. In Figure 7, we analyze the efficiency of our methods for increasingly large numbers of candidates and the effect of the $\Delta$ parameter on the execution time. We create a Mallows model dataset with a modal ranking with $ARP(Race) = 0.31$, $ARP(Gender) = 0.44$, $IRP =$
E. Empirical Takeaways

When utilizing our MFCR methods, we recommend Fair-Kemeny for smaller candidate databases and note that the number of rankings can be reasonably large (thousands). Next, Fair-Copeland and Fair-Schulze provide nearly comparable performance on larger candidate databases and number of rankings when fairness requirement is strict. However, if fairness requirement is looser, we recommend Fair-Copeland for more efficiency with decreased PoF. If the consensus ranking problem is very large, Fair-Borda is the best choice with significant speed-up and minimal increase in PoF.

F. Case Study of Student Merit Scholarships

We demonstrate that our MFCR solutions create real-world fair consensus rankings over students with multiple protected attributes. Entrance exam and test scores are commonly used as part of admissions decisions or merit scholarship allocations for educational institutions ranging from magnet schools \(^45\) to graduate admission \(^46\). As sociodemographic factors such as student socioeconomic status, race, and gender can have a large effect on exam outcomes \(^47\)–\(^49\), schools and testing organizations are exploring ways to level the playing field when using exam scores for admission decisions \(^50\).

We utilized a publicly available dataset \(^34\) modeling student exam scores for math, reading and writing exams. The data contained three protected attributes of Gender (man or woman), Race (5 racial groups) and Lunch (if student received reduced cost lunches). We utilized the exam scores provided in each subject to create \(|R| = 3\) base rankings (ordered by score) over 200 students.

In Table IV all protected attributes have an ARP \(\geq 0.2\) (with higher TRP scores) across all base rankings – indicating statistical parity is far from being achieved. This contrast is particularly stark as we can see students with subsidized lunches are ranked low, along with NatHawaiin students. Also, there appears to be a substantial gender imbalance. We create a (fairness-unaware) Kemeny consensus ranking, and observe that the biases in the base rankings are unfortunately also reflected in the consensus ranking. Thus, if the consensus ranking was used to determine merit scholarships then the students who receive subsidized lunch would receive almost three times less aid as the group of students who do not require subsidized lunches. The group of men would also receive more merit aid then women. NatHawaiian students would receive only one percent of the amount of aid as Asian and Black students.

We then compare this to utilizing our four proposed MFCR solutions, employing the MANI-Rank criteria and setting \(\Delta = 0.05\) to ensure almost perfect statistical parity across all protected attributes and their intersection. All methods generate a de-biased consensus ranking, with ARP \(\leq 0.05\) and IRP \(\leq 0.05\). This translates to all groups receiving an extremely close to equal proportion of merit scholarships.
The disparities between merit aid received by the men and women are nearly nonexistent, and the difference between racial groups is leveled. The severe advantage of students who do not require subsidized lunch is also removed. As conclusion, utilizing the fair consensus rankings created by our MFCR solutions ensures certain student groups are not disadvantaged in the merit aid process.

V. RELATED WORK

**Fair Ranking Measures.** While the majority of algorithmic fairness work has concentrated on the task of binary classification, several notions of group fairness have been defined for (single) rankings. The most widely adopted notion of group fairness in rankings is statistical parity [18]–[22], [25], [51], [52]. Most works focus on measuring and enforcing statistical parity between two groups defined by a single binary protected attribute [18]–[20], [22], [24]. The pairwise fairness metrics for a single protected attribute of binary groups of [19] inspire our metrics, we propose metrics extending the pairwise approach to multi-valued attributes and for multiple attributes.

The recent works of Narasimhan et al. [51] and Geyik et al. [21] propose group fairness metrics for a single multi-valued protected attribute of multiple values. Narasimhan et al. [51] introduce a pairwise accuracy statistical parity notion for ranking and regression models. Our pairwise statistical parity notion differs by counting pairs directly as consensus ranking is not performed from a model whose accuracy we can constrain. Geyik et al. [21] formulate group fairness by casting the group positive outcome as inclusion in the top-\( k \). As consensus generation combines multiple whole rankings, simply setting \( k = n \) would not capture group fairness for a consensus ranking.

**Multiple Protected Attributes.** Recent work addressing group fairness in multiple protected attribute settings is entirely focused on the binary classification task. Kearns et al. [15] introduced procedures for auditing and constructing classifiers subject to group fairness notions where groups could be defined as combinations of protected attributes. Hebert-Johnson at al. [17] proposed an approach which ensures accurate predictions on all combinations of groups formed from protected attributes. Foulds et al. [16] proposed differential fairness; an intersectional fairness notion, for ensuring group fairness with respect to classification outcome probabilities for all possible combinations of sub-groups while ensuring differential privacy.

Within the domain of rankings, Yang et al. [53] design algorithms to mitigate within-group unfairness in the presence of diversity requirements for groups encoded by multiple protected attributes – we consider fairness between groups. Yang et al. [23] present a counterfactual causal modelling approach to create intersectionally fair rankings in score-based and learning-to-rank models. While we assume access to the base rankings, we do not know why they reflect a specific order or how the rankings would differ based on changes in the protected attributes of candidates. Without these counterfactual outcomes, causal fairness notions are difficult to deploy.

**Rank Aggregation.** Rank aggregation originates from the study of ranked ballot voting methods in Social Choice Theory [26], [35], [54]. Rank aggregation aims to find the consensus ranking which is the closest to the set of base rankings. This has been studied in information retrieval [29], machine learning [55], databases [43] and voting theory [56]. Kemeny rank aggregation, a preeminent rank aggregation method satisfying several social choice axioms [35], has been applied to a wide set of applications: MRNA gene rankings [57], teacher evaluations [58], and conference paper selections [59]. Recent work [25], which introduced the fair rank aggregation problem, also leverages Kemeny rank aggregation. However, they assume a single binary protected attribute. Our work instead now handles multi-valued as well as multiple protected attributes.

VI. CONCLUSION

This work introduces the first solution to the multi-attribute fair consensus ranking (MFCR) problem. First, we design novel MANI-Rank fairness criteria to support interpretable tuning of fair outcomes for groups defined by multiple protected attributes in consensus rankings. We then design four alternate MFCR algorithms using our proposed MANI-Rank model. We demonstrate the efficacy, scalability, and quantify the price of fairness achieved by our MFCR solutions in selecting a fair consensus ranking over a vast array of rank aggregation scenarios. Code, metrics, supplemental material, and experiments made publicly available. https://github.com/KCachel/MANI-Rank

| Ranking   | Men  | Women | Gender | NoSub | SubLunch | Lunch | Asian | White | Black | AlaskaNat. | NatHaw. | Race | IRP |
|-----------|------|-------|--------|-------|----------|-------|-------|-------|-------|------------|---------|------|-----|
| Math      | 0.39 | 0.61  | 0.37   | 0.72  | 0.28     | 0.44  | 0.59  | 0.49  | 0.56  | 0.55       | 0.22    | 0.37 | 0.65|
| Reading   | 0.60 | 0.4   | 0.20   | 0.63  | 0.37     | 0.26  | 0.35  | 0.45  | 0.36  | 0.56       | 0.29    | 0.27 | 0.47|
| Writing   | 0.63 | 0.37  | 0.24   | 0.68  | 0.32     | 0.36  | 0.56  | 0.47  | 0.56  | 0.52       | 0.32    | 0.24 | 0.51|
| Kemeny    | 0.57 | 0.43  | 0.14   | 0.67  | 0.33     | 0.34  | 0.57  | 0.46  | 0.57  | 0.54       | 0.27    | 0.30 | 0.52|
| Fair-Kemeny | 0.51 | 0.49  | 0.02   | 0.52  | 0.48     | 0.04  | 0.52  | 0.50  | 0.51  | 0.50       | 0.47    | 0.04 | 0.05|
| Fair-Schulze | 0.50 | 0.50  | 0.00   | 0.52  | 0.48     | 0.04  | 0.52  | 0.50  | 0.50  | 0.51       | 0.47    | 0.05 | 0.05|
| Fair-Borda | 0.50 | 0.50  | 0.00   | 0.52  | 0.48     | 0.04  | 0.52  | 0.50  | 0.51  | 0.50       | 0.48    | 0.04 | 0.05|
| Fair-Copeland | 0.51 | 0.49  | 0.02   | 0.52  | 0.48     | 0.04  | 0.52  | 0.49  | 0.51  | 0.50       | 0.48    | 0.04 | 0.05|

TABLE IV: Exam Case Study: Attribute values columns (e.g, Men, SubLunch) indicate FPR scores, and Gender, Race, and Lunch indicate ARP scores.
While group fairness concerns typically emerge when people are being ranked, they also apply for other ranked entities. We illustrate the effectiveness our framework on removing bias in Computer Science Department rankings using the publicly available CSRankings from csrankings.org. We collected rankings of 65 departments in the US over the period of 2000–2020 (utilizing the relative order presented on csrankings.org as the ranking for each year) to generate a 20-year consensus ranking. To examine geographic and private vs. public institutional bias, we used the location (Northeast, South, West, MidWest) and type (Public or Private) of each institution as protected attributes.

In Table V, we observe that the rankings over the years indeed exhibit bias, namely, a strong bias in the location attribute (high ARP) - which stems from institutions in the Northeast region commonly appearing at the top of the rankings, along with those in South region exhibiting the opposite trend. Additionally, there is a significant amount of intersectional bias — resulting from the base rankings having Private and Northeast colleges highly ranked.

When using fairness-unaware Kemeny to create the 20-year consensus ranking, we observe that this bias is amplified, with the Location ARP score resembling that of base rankings with higher such scores, and the IRP score close to 0.6. By utilizing our MANI-Rank criteria and setting the group fairness threshold $\Delta = .05$, it can be seen that we were able to remove the bias in the consensus ranking. All our proposed methods remove the bias toward Northeast and Private universities.
TABLE V: CSRankings Study: Attribute values columns (e.g, Northeast, Prive) indicate FPR scores, and Location, and Type indicate ARP scores.

| Ranking | Northeast | Midwest | West  | South | Location | Private | Public | Type | IRP |
|---------|-----------|---------|-------|-------|----------|---------|--------|------|-----|
| 2000    | 0.692     | 0.373   | 0.574 | 0.315 | 0.377    | 0.642   | 0.358  | 0.284| 0.536|
| 2001    | 0.712     | 0.370   | 0.598 | 0.270 | 0.442    | 0.608   | 0.392  | 0.217| 0.582|
| 2002    | 0.700     | 0.451   | 0.610 | 0.203 | 0.497    | 0.569   | 0.431  | 0.138| 0.574|
| 2003    | 0.689     | 0.392   | 0.589 | 0.287 | 0.402    | 0.562   | 0.438  | 0.124| 0.557|
| 2004    | 0.658     | 0.481   | 0.570 | 0.265 | 0.392    | 0.610   | 0.390  | 0.221| 0.520|
| 2005    | 0.680     | 0.457   | 0.552 | 0.278 | 0.402    | 0.607   | 0.393  | 0.213| 0.480|
| 2006    | 0.707     | 0.451   | 0.550 | 0.255 | 0.452    | 0.582   | 0.418  | 0.164| 0.476|
| 2007    | 0.691     | 0.435   | 0.601 | 0.236 | 0.455    | 0.607   | 0.393  | 0.213| 0.553|
| 2008    | 0.724     | 0.416   | 0.550 | 0.265 | 0.459    | 0.610   | 0.390  | 0.221| 0.515|
| 2009    | 0.693     | 0.469   | 0.599 | 0.205 | 0.488    | 0.591   | 0.409  | 0.181| 0.570|
| 2010    | 0.687     | 0.426   | 0.582 | 0.268 | 0.419    | 0.608   | 0.392  | 0.217| 0.521|
| 2011    | 0.714     | 0.426   | 0.573 | 0.245 | 0.470    | 0.607   | 0.393  | 0.215| 0.559|
| 2012    | 0.694     | 0.460   | 0.574 | 0.237 | 0.457    | 0.582   | 0.418  | 0.164| 0.542|
| 2013    | 0.738     | 0.436   | 0.533 | 0.249 | 0.489    | 0.606   | 0.394  | 0.211| 0.545|
| 2014    | 0.674     | 0.476   | 0.561 | 0.259 | 0.416    | 0.582   | 0.418  | 0.164| 0.495|
| 2015    | 0.726     | 0.469   | 0.532 | 0.236 | 0.490    | 0.599   | 0.401  | 0.197| 0.561|
| 2016    | 0.711     | 0.470   | 0.517 | 0.267 | 0.445    | 0.606   | 0.394  | 0.211| 0.512|
| 2017    | 0.688     | 0.478   | 0.540 | 0.264 | 0.424    | 0.547   | 0.453  | 0.095| 0.448|
| 2018    | 0.661     | 0.504   | 0.559 | 0.253 | 0.409    | 0.551   | 0.449  | 0.103| 0.457|
| 2019    | 0.680     | 0.499   | 0.531 | 0.264 | 0.416    | 0.566   | 0.434  | 0.132| 0.467|
| 2020    | 0.666     | 0.445   | 0.540 | 0.318 | 0.348    | 0.579   | 0.421  | 0.158| 0.415|
| Kemenen | 0.712     | 0.445   | 0.570 | 0.233 | 0.479    | 0.602   | 0.398  | 0.203| 0.570|
| Fair-Kemenen | 0.532 | 0.499 | 0.532 | 0.432 | 0.100  | 0.498  | 0.502 | 0.004| 0.099|
| Fair-Schulze | 0.529 | 0.501 | 0.529 | 0.436 | 0.093  | 0.494  | 0.501 | 0.012| 0.089|
| Fair-Borda | 0.519 | 0.499 | 0.534 | 0.445 | 0.089  | 0.496  | 0.504 | 0.008| 0.094|
| Fair-Copeland | 0.531 | 0.501 | 0.531 | 0.432 | 0.099  | 0.494  | 0.506 | 0.012| 0.099|