Wise Fusion: Group Fairness Enhanced Rank Fusion

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

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

Abstract

Rank fusion is a technique for combining multiple rankings into a single aggregated ranking, commonly used in high-stakes applications. For hiring decisions, a fused ranking might combine evaluations of different candidates from various job boards into one list. Ideally, such fused rankings are fair. Meaning they do not withhold opportunities or resources from marginalized groups of candidates, even if such biases may be present in the to-be-fused rankings. Prior work fairly aggregating rankings is limited to ensuring proportional (not addressing equality) fairness when combining ranked lists containing the same candidate items. Yet, real-world fusion tasks often combine rankings of varying candidate sets, may also contain relevance scores, or are better suited to equal representation. To address fairness in these settings, we present a new plug-and-play fairness-aware fusion strategy: Wise fusion. Wise works in fusion settings where we have closed-box access to a score-powered rank fusion (SRF) method, making it possible to fairness-enhance existing fusion pipelines with little added cost. Wise uses existing evaluations of candidates from an as-is SRF method to achieve proportional or equal rank fairness in the final fused ranking. Our experimental study demonstrates that Wise beats the fairness and utility performance of state-of-the-art methods applied to these new fair rank fusion settings.

CCS Concepts

• **Information systems** → **Information retrieval**; • **Social and professional topics** → **User characteristics**.

Keywords

Group Fairness; Fair Rank Fusion; Algorithmic Fairness.

ACM Reference Format:
Kathleen Cachel and Elke Rundensteiner. 2024. Wise Fusion: Group Fairness Enhanced Rank Fusion. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3627673.3679649

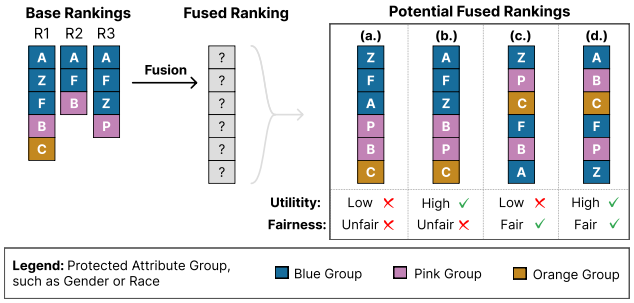

Figure 1: Fair Rank Fusion. Given to-be-fused rankings with varying degrees of candidate overlap and possibly relevance scores, the goal, illustrated with the fused ranking (d.), is to create a ranking that is fair with respective to multiple groups and provides a high utility representation of base rankings. The balance of these two objectives is also tunable.

1 Introduction

Background. Rank fusion methods combine multiple, potentially conflicting, base rankings of candidate items into a single fused ranking [7, 56]. Such systems play a pivotal role in fusing rankings in impactful domains, including employment [49, 54], medical decision-making [25], and modern metasearch engines [10, 56]. The leading approach to fusing multiple rankings is the class of *score-powered rank fusion (SRF)* algorithms, e.g., Borda Fuse [6], Coomb Fuse [39], and many others [5–7, 26, 32, 58, 60, 61, 63, 66–68]. These SRF methods take in multiple rankings, each of which may contain relevance scores. The final fused ranking is generated by ordering candidates based on each method's specific scoring function, which incorporates factors such as candidate positioning, accompanying relevance scores, and popularity among base rankings.

Existing SRF methods are ideal for providing a fused ranking that represents the base rankings, i.e., has high utility. However, in high-stakes domains, we must also ensure that a high utility fused ranking does not unfairly disadvantage marginalized groups, such as race or gender. When fusing rankings, the detrimental effects of societal biases in individual rankings can be exacerbated, as the very act of combining potentially biased rankings can strengthen existing biases or even introduce new ones [17, 54]. This realization has led to the recent proposal of methods that incorporate group proportional fairness into Kemeny rank aggregation [50], e.g., *fair rank aggregation methods (FRA)* [17, 18, 54, 89].

Limitation of the State-of-the-Art. Despite significant advances, these solutions [17, 18, 54, 89] do not support many prevalent types of to-be-fused ranked data. In particular, existing FRA methods make two restrictive assumptions. First, they assume that base rankings are conjoint ranked lists [1]. Then, they ignore any relevance scores associated with candidates which risks unnecessarily

[1]Throughout this work we use the term conjoint to describe rankings that order the exact same items.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679649

lowering the final fused ranking's utility. For instance, some hiring recruitment systems combine the results of multiple ranking models to surface a final candidate ranking for the recruiter to act on [4]. Each ranking model may produce an ordering of applicable candidates from a platform-specific API, e.g., Monster, Indeed, Linkedin, etc. Such models might also provide a relevance score for a desired purpose, e.g., experience level or "company fit" [73]. However, when combining rankings from different platforms, it is unlikely that the same people are all on the same platform. Ignoring the learned relevance score could also overlook valuable candidates. The fairness community has yet to address such fusion contexts.

Furthermore, adopting existing FRA methods [17, 18, 54, 89] requires replacing a fusion system with an entirely new one. Because FRA methods adopt the Kemeny approach to rank aggregation [17, 18, 54, 89], the internal mechanics of the fusion approach must also use the Kemeny criteria [50]. This is undesirable for domains that have rigorously validated and designed custom SRF methods.

Problem. Addressing previously overlooked forms of to-be-fused rankings, we conceptualize a *fair rank fusion problem.* Consider a set of base rankings that may be positional, score-enhanced, and/or non-conjoint. The goal is to combine these rankings into a fused ranking that ensures *all groups* receive *favorable shares of higher positions* in the fused ranking while maintaining as much overall utility as possible. This is a multi-objective problem – with the goals of *fairness* and *utility* often conflicting. Section 3 presents our detailed problem, while Figure 1 depicts our ideal solution.

Challenges. We identify three core challenges in addressing this fair rank fusion problem. First, *multiple popular notions of fairness* apply to rankings [71, 74], and the decision of which notion to utilize when debiasing depends on the task at hand. Stakeholders may wish to represent groups either equally or proportionally at favorable positions in the fused ranking. Yet, these two distinct notions of fairness require different debiasing approaches [40, 71]. Addressing this challenge is of practical value as FRA methods to date only support proportional fairness [17, 18, 54, 89].

Second, prior to fusion, the dynamics of the *tradeoff* between utility and fairness is unknown. As illustrated in Figure 1, multiple combinations of utility and fairness exist in a fused ranking. Practical solutions must facilitate navigating the tradeoff between the two objectives. Third, while we aim to support fairly fusing a variety of input base rankings, e.g., *disjoint, conjoint, positional, and relevance-scored (and combinations thereof),* proposing many distinct custom solutions coupled with multiple fairness notions is impractical. To ensure the ease of bias-mitigation adoption, a *single easy-to-use solution agnostic* to specific input forms is essential.

Proposed Approach. Addressing the aforementioned challenges, we present *Within-a-group Similarity Enhanced fusion*, in short, Wise. As a conceptual foundation for Wise's machinery, we introduce the *fair fusion principle*, stating that candidates who perform similarly within their respective groups should receive similar favorable positions in the fused ranking. Wise operationalizes this principle through novel group similarity-determination strategies for proportional and equal fairness.

Advancing in capabilities beyond prior methods [17, 18, 54, 89] (see Table 1), Wise integrates with an off-the-shelf fairness-unaware SRF method (e.g., [6, 7, 39]), making it applicable to fusing a wide variety of input base rankings. We design Wise with a fairness-control mechanism, which allows users to trade-off between the maximal utility of the existing SRF method and fairness by adjusting a single parameter λ. Wise is an easy-to-use approach with a relatively low engineering footprint, and, unlike past methods [17, 18, 54, 89], can be easily added on top of existing rank fusion pipelines.

In summary, we make the following contributions:

- We define the *fair rank fusion problem* bridging a gap between the capabilities of fair rank aggregation methods and modern forms of to-be-fused rankings.
- We design Wise, the first solution to this problem. By enhancing a given SRF method with our proposed within-a-group similarity knowledge, Wise operationalizes the fair rank fusion principle introduced for debiasing rank fusion.
- We conduct extensive experiments comparing Wise to state-of-the-art bias mitigation methods [17, 37, 42, 89] adapted to our problem. On real-world and controlled datasets representing previously unstudied forms of rank fusion, we observe Wise outperforms these methods in terms of fairness and utility while providing similar computational efficiency.

2 Related Work

Fairness-Unaware Rank Fusion. The task of combining multiple ranked lists has origins in the field of social choice theory [16], where such methods, generally referred to as voting rules [95], are studied for democratic elections [24, 34, 35]. Likewise, rank aggregation mechanisms underpin many modern information access systems [9, 70]. This branch of work designs algorithms that can be categorized into three algorithmic approaches: *distance-based optimization methods* [11, 23, 28, 80] (including deterministic heuristics [21, 87]), *randomized methods* [1, 2], or, as described at the outset, *score-powered rank fusion* [5–7, 26, 32, 39, 58, 60, 61, 63, 66–68].

For distance-based aggregation, popular optimization functions include the Kemeny criteria which minimizes the Kendall Tau distance between a final ranking and each of the base rankings [3, 94]. Pick-A-Perm [2] is an example of a randomized aggregation approach. This method randomly selects a single base ranking as the final ranking. Score-powered rank fusion (SRF), has gained popularity in information access systems, because unlike distance-based aggregation methods [11, 21, 23, 28, 80, 87], these methods do not require that base rankings be conjoint ranked lists. Moreover, SRF methods often incorporate relevance scores that may be associated with candidates in each of the to-be-fused rankings [6, 39, 57]. Prominent SRF algorithms include Borda Fuse [6], Coomb Fuse family [39], ProbFuse [63], Reciprocal Rank Fuse [26], and Rank-Biased Centroid Fusion [7]. Unlike our work, the above methods do not address group fairness during the rank fusion process.

Fairness in Rankings. The field of fair ranking is a broad research area, including many different fairness metrics [12, 30, 40, 69, 79, 83, 93], bias-mitigation methods, empirical studies [8, 22, 38, 41, 72, 86, 90], and relevant surveys [33, 59, 71, 98, 99]. Fair ranking metrics typically quantify either individual [13, 15, 83] or group fairness [12, 31, 69, 79, 82, 93]. Bias-mitigation methods are described as either pre-processing [19, 36], in-processing [12, 75, 84, 97], or

Table 1: Comparison of fair rank aggregation (FRA) methods.

Approaches:	Distance-based FRA [17, 18, 54]	Randomized FRA [89]	This work (Wise)
Supports Equal & Prop. Fairness	X	X	✓
Incorporates Relevance Scores	X	X	✓
Fuses Non-conjoint rankings	X	✓	✓
Has Fairness Control Parameter	✓	X	✓

post-processing [20, 30, 37, 42, 83, 96] with respect to the task of learning to rank. Empirical studies have analyzed gender bias in resume rankings [22], and performed fairness analyses of specific platforms such as TaskRabbit ranking algorithms [86], Pymetrics candidate rankers [91], and music streaming services [38]. The aforementioned works in this category do not address fairness concerns when combining multiple rankings. Our work complements this line of research, by employing existing fair ranking metrics [40] and adding to the toolkit of bias-mitigation methods.

Fair Rank Aggregation. As stated in Section 1, the prior work most closely related to our target task is *fair rank aggregation* (*FRA,* for short) [17, 18, 54, 89]. We group these methods into two categories summarized in Table 1. The *distance-based FRA* category encompasses strategies that enforce notions of proportional group fairness as constraints in the Kemeny rank aggregation optimization program [17, 18, 54]. Departing from this literature, our work does not assume to-be-fused rankings are conjoint lists, and we handle incorporating candidate relevance scores in the fusion process.

Rank Aggregation Proportional Fairness, or Rapf [89] is a FRA approach we refer to as *randomized FRA* because it, like Pick-a-Perm [2], picks a random base ranking to re-rank until proportional fairness is satisfied. Rapf does not contain a mechanism to control the level of bias-mitigation. It also ignores any relevance scores associated with ranked candidate items. While Wei et al. [89] explicitly state that Rapf assumes all base rankings rank an identical set of candidates, we observe in our experiments (Section 5) that its Pick-A-Perm-based approach does not break on non-conjoint input. We thus categorize this method as "handling" non-conjoint base rankings (Table 1). Nonetheless, only the candidates that are in the selected individual ranking will be in the final Rapf ranking.

As neither the *distance-based FRA* [17, 18, 54] nor *randomized FRA* [89] approaches explicitly address combining non-conjoint lists, equal rank fairness, or make use of candidate relevance scores we experimentally find that they underperform our proposed solution in such fusion settings.

3 Problem Formulation

Fusing Ranked Lists. Our setting involves a set $X = \{x_1, x_i, ..., x_m\}$ of m candidate items, and a set $\mathcal{R} = \{R_1, R_2, ..., R_n\}$ of n ranked lists over set X [7, 39, 61]. For instance, ranked list R_i could represent a base ranking produced by a ranking model i. Every base ranking R_i ranks $\leq |X|$ candidates. We do not assume that each base ranking necessarily orders exactly the *same* set X. We assume that each candidate $x_j \in X$ appears in at least one base ranking.

We consider two main types of base rankings. First, those that only provide an ordering of candidates, e.g., $R_i = x_1 \prec x_4 \prec x_7$, where $a \prec b$ means a is ranked higher than b. We refer to these as *positional-only rankings* or in short $\mathcal{R}_p$. The second form is base rankings which provide an ordering and a numeric score associated with each candidate item [39]. Each base ranking is a ranked list of items of the format $(x_j, s_i(x_j))$, with x_j indicating the candidate item and $s_i(x_j)$ denoting a relevance value associated with candidate x_j for base ranking R_i. We refer to these base rankings as *score-enhanced base rankings*, or in short $\mathcal{R}_s$. Base rankings (either $\mathcal{R}_p$ or $\mathcal{R}_s$) are combined into a single fused ranking R^*.

Fairness. Specific to our rank fusion problem, candidate items have an associated categorical protected attribute p. Examples of protected attributes include nationality, race, gender, or their combination. The set of candidates in X that share the same value in the protected attribute is referred to as a *group*. We use $G = \{g_1, ..., g_v\}$ to denote the set of v groups in candidate set X. For fairness, we mitigate bias with respect to the protected attribute p.

Our Problem: Fair Rank Fusion. The notion of what constitutes a *fair* representation of candidate groups at favorable fused ranking R^* positions is a contextual decision dictated by the domain [85]. Thus, we handle both popular forms of group rank fairness, equal and proportional, so that practitioners can customize the approach to their needs. Table 2 defines the two fairness metrics we employ and our measure of utility (fusion quality). For *fairness metrics,* we use Normalized Discounted KL-Divergence [40] since it has equal, i.e., NDKL *eq.*, and proportional, i.e., NDKL *prop.* formulations. In each formula, d_{KL} is the KL-divergence score [55], $\mathbf{D_{r:i}}$ the proportions of each group in the first i positions of ranking R, and $Z = \sum_{i=1}^{|R|} \frac{1}{log_2(i+1)}$ is a normalizing factor [40].

In Table 2 we define Average Rank Biased Overlap, or ARBO as our utility metric. In ARBO, we utilize Rank Bias Overlap *rbo*, with the persistence parameter $p = 1$ to evaluate all positions [88]. We choose Rank Bias Overlap as opposed to other similarity measures (e.g., Kendall Tau [50]), as it does *not require conjoint ranked lists.*

Having described its elements, our *fair rank fusion* problem is to combine a given set of base rankings (e.g., positional-only ranked lists $\mathcal{R}_p$ or score-enhanced ranked lists $\mathcal{R}_s$), representing candidate item set X, into a final fused ranking R^*. R^* ensures all groups receive comparable shares of favorable rank positions (equal or proportional) and simultaneously provides a high-quality fusion of the potentially conflicting ranked lists. Ranking R^* has a low NDKL (*eq.* or *prop*) value and a high ARBO value. *Fair rank fusion* prioritizes fairness while also seeking to maximize utility. However, how to balance the tradeoff between the two objectives is a question best answered by practitioners and stakeholders [27, 85]. Thus, we seek a tunable solution design that allows for adjusting the level of bias-mitigation applied to the task at hand.

4 Proposed Methodology: Wise

We provide the key conceptual idea underlying our proposed solution to the fair rank fusion task in Section 4.1. Then, we present the proposed Wise fusion solution. in Section 4.2.

4.1 Wise: Key Idea of Fair Fusion Principle

Conventional SRF methods [6, 7, 39] take as input a set of base rankings and calculate for every $x_i \in X$ a single fusion score $f(x_i)$.

Table 2: Fairness and utility metrics employed in evaluating Fair Rank Fusion. *rbo* refers to rank biased overlap [88].

Metric	Measures	Formula	Interpretation
NDKL *eq.* [40]	Equal group representation at favorable positions in R^*	$\frac{1}{Z}\sum_{i=1}^{\lvert R^*\rvert}\frac{1}{log_2(i+1)}d_{KL}(D_{R^*:i}\lvert\lvert[\lvert X\rvert\backslash\lvert g_1\rvert,...,\lvert X\rvert\backslash\lvert g_v\rvert])$	↓ more fair
NDKL *prop.* [40]	Proportional group representation at favorable positions in R^*	$\frac{1}{Z}\sum_{i=1}^{\lvert R^*\rvert}\frac{1}{log_2(i+1)}d_{KL}(D_{R^*:i}\lvert\lvert[\lvert G\rvert^{-1},...,\lvert G\rvert^{-1}]_{1x\lvert G\rvert})$	↓ more fair
ARBO (*rbo* [88])	Utility of fused R^* in representing $\mathcal{R}$	$\frac{1}{n}\sum_{j=1}^{n} rbo(R^*, R_j)$	↑ more utility

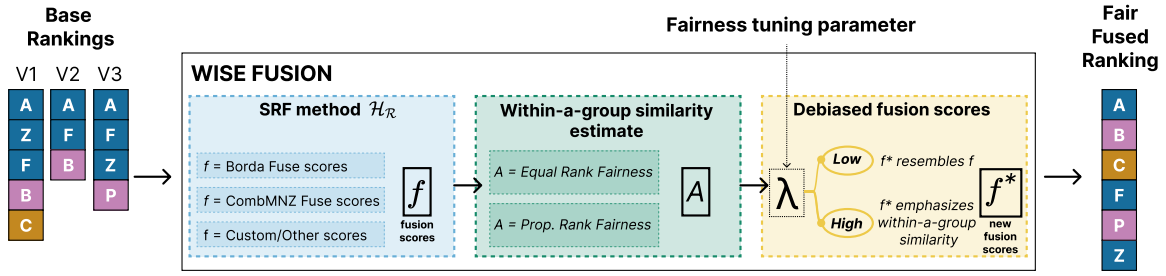

Figure 2: Wise uses evaluations of candidates from an existing SRF method to achieve proportional or equal rank fairness in the final fused ranking. As parameter $\lambda \in (0, 1]$ increases, the final fused ranking incorporates more within-a-group similarity information. Candidates that perform similarly within their respective groups receive similar positions in the fused ranking.

Thereafter, the fused ranking orders candidate items in X by decreasing fusion scores. These methods only focus on maximizing utility; thus, when certain groups occupy large shares of favorable positions, their final fused ranking can be biased or unfair.

Remark 1 formally establishes our core conceptual principle for mitigating unfair group treatment during rank fusion.

Remark 1 (Fair Fusion Principle). *If two candidate items belonging to disjoint groups perform similarly within their respective groups, they should be placed in similar positions in the fused ranking.*

Our motivation for proposing the fair fusion principle is the *group model of bias* [14, 52, 78], which was introduced by Kleinberg and Raghavan [52]. The group model of bias is a theoretical framework for conceptualizing how discriminatory bias affects empirical evaluations of demographic groups. It states that bias is applied inconsistently to different groups but consistently within each group [52, 77]. In other words, given empirical evaluations of candidates, those evaluations are often accurate within each group but are not necessarily accurate when performing comparisons across groups.

The fair fusion principle addresses the concerns of the group model of bias in the context of fusing multiple rankings. The core implication of the fair fusion principle is that fusion scores $f(x_i)$ within each group reveal performance within a group. Then, similar performance *across groups* can be derived from within-a-group performance. For example, the best candidates in each group should be somewhat closely positioned to one another. Likewise, the best candidate for one group should not be in a position similar to that of the 10th, 20th, 50th, etc. candidate of another group. Mechanistically, we can use the introduced notion of similar within-a-group performance to "lift up" candidate items in disadvantaged groups. For now, we intentionally under-specify the definition of similar within-a-group performance to provide applicability to both equal and proportional fairness notions. We formalize within-a-group similarity in Section 4.2.2, illustrating that the fair fusion principle can be applied to both forms of fairness.

To the best of our knowledge, this is the first conceptual framework for integrating contemporary group fairness with SRF approaches. Past work has introduced conceptual frameworks focused on solely increasing the utility of the final fused ranking. Most prominently, for contexts in which ranked candidate items are textual documents, the *cluster hypothesis* [48] underpins a number of methods for combining rankings [29, 51, 61]. The cluster hypothesis states that documents with similar contents have a similar degree of relevance for the given information need [48].

Cluster hypothesis-powered SRF methods [29, 51, 61] operationalize the cluster hypothesis by regularizing fusion scores with inter-document similarity information using the regularization paradigm proposed in Zhou et al. [100] and Liu et al. [64]. This paradigm is an efficient closed-form solution to regularizing item scores with additional pairwise item information. Like prior state-of-the-art rank fusion methods [29, 51, 61], we also adopt this regularization paradigm, but for a new objective – namely, bias mitigation.

4.2 The Wise Fusion Solution

We introduce Wise. Wise is SRF method agnostic in the sense that it relies only on the fusion scores provided by a score-powered rank fusion method $\mathcal{H}_\mathcal{R}$, making it applicable to fairness-enhance a wide array of rank fusion systems. The high-level Wise approach is illustrated in Figure 2. Once fairness-unaware fusion scores $f(x_i)$ are determined from $\mathcal{H}_\mathcal{R}$, Wise extracts how candidates perform within their respective groups. Then, addressing the need to support multiple notions of rank fairness, Wise introduces two within-a-group similarity-determination procedures designed to "lift up" disadvantaged candidate groups. This allows practitioners to control whether the fused ranking satisfies equal or proportional rank fairness. Wise then regularizes fusion scores $f(x_i)$ with this new

Algorithm 1 Wise Fusion

Input: Base rankings $\mathcal{R}$ (e.g., $\mathcal{R}_p$, scored) of a candidate item set X, including each candidate item's group membership in the set of groups G, a desired rank fairness $\phi \in \{\text{equal, proportional}\}$, and fairness tuning parameter $\lambda \in (0, 1]$.
Require: Use of a fairness-unaware SRF method $\mathcal{H}_{\mathcal{R}}$.
Output: Fused ranking R^* ordering a candidate item set X.

1: Obtain fusion scores $f \leftarrow$ from $\mathcal{H}_{\mathcal{R}}$
2: **for** every $x_i \in X$ **do**
3: **for** every $x_j \in X$ **do**
4: **if** x_j and x_j belong to different groups in G **then**
5: **if** ϕ == equal **then**
6: $\mathbf{A}_{ij} \leftarrow \delta_{eq}(x_i, x_j)$ // Eq. 2
7: **if** ϕ == proportional **then**
8: $\mathbf{A}_{ij} \leftarrow \delta_{prop}(x_i, x_j)$ // Eq. 3
9: $\mathbf{D} \leftarrow$ Diagonal normalizing matrix of $\mathbf{A}$
10: $f^* \leftarrow \left(\mathbf{I} - \lambda \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}\right)^{-1} f$
11: Obtain the fused ranking R^* by sorting candidate items $x_i \in X$ by decreasing debiased fusion scores f^*
12: **return** R^*

within-a-group similarity knowledge. The level of debiasing is controlled via the fairness-tuning parameter λ. Increasing λ integrates more within-a-group similarity knowledge into the fusion process.

4.2.1 Wise: Enhancing existing SRF methods. As illustrated in Figure 2, and shown in line 1 of Algorithm 1, Wise uses an existing fairness-unaware SRF method $\mathcal{H}_{\mathcal{R}}$ as-is. $\mathcal{H}_{\mathcal{R}}$ determines fusion scores for the entire candidate item set X based on the provided base rankings. Wise produces fair fused rankings via as free-standing "plug and play" component that integrates with $\mathcal{H}_{\mathcal{R}}$. This design choice frees Wise from restrictive assumptions on the type of base rankings that it can fairly fuse. For example, prior FRA ranking methods (e.g., [17, 54]) are restricted to combining only conjoint ranked lists into a final fair ranking. Wise can combine any form of to-be-fused base rankings that $\mathcal{H}_{\mathcal{R}}$ can. This opens up fairness support for *disjoint, conjoint, positional, and relevance-scored (and combinations thereof)* base rankings.

In our experiments, we utilize two SRF methods: Borda [6] for $\mathcal{R}_p$ base rankings and Combmnz [39] for $\mathcal{R}_s$ base rankings, as each is designed for that respective setting. The standard fusion scores f produced by this first step, namely, $\mathcal{H}_{\mathcal{R}}$, are then used as part of the subsequent debiasing process.

4.2.2 Wise: Debiasing Fusion Scores. To generate fair fused rankings, Wise enhances fusion scores to incorporate pairwise information on how two candidates compare in terms of their performance *within* their groups. Specifically, Wise determines debiased fusion scores f^* as shown in line 10 of Algorithm 1. The key ingredient in determining f^* is a matrix, A, capturing the within-a-group similarities of candidate items. We define $\mathbf{A} \in \mathbb{R}^{m \times m}$ to be composed of candidate similarity information. $\mathbf{A}$ is generated from the overarching fairness objective stipulated by the user, i.e., if input ϕ in Algorithm 1 is equal or proportional rank fairness. Matrix $\mathbf{A}$ values are set to zero when the compared candidate items x_i and x_j belong to the same group, since our objective is to establish

Figure 3: Example of within-a-group similarity for both forms of fairness in the sample data shown in Figure 3a.

Item x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
Group g_v	g_1	g_1	g_1	g_1	g_1	g_1	g_2	g_2	g_3	g_3	g_3
Fusion score $f(x_i)$	30	27	24	21	18	15	12	9	6	3	0

(a) X with group membership and standard fusion scores.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	0	0	0	0	ϵ	0	1	0	1	ϵ	ϵ
x_2	0	0	0	0	0	0	ϵ	1	ϵ	1	ϵ
x_3	0	0	0	0	0	0	ϵ	ϵ	ϵ	ϵ	1
x_4	0	0	0	0	0	0	ϵ	ϵ	ϵ	ϵ	ϵ
x_5	0	0	0	0	0	0	ϵ	ϵ	ϵ	ϵ	ϵ
x_6	0	0	0	0	0	0	ϵ	ϵ	ϵ	ϵ	ϵ
x_7	1	ϵ	ϵ	ϵ	ϵ	ϵ	0	0	1	ϵ	ϵ
x_8	ϵ	1	ϵ	ϵ	ϵ	ϵ	0	0	ϵ	1	ϵ
x_9	1	ϵ	ϵ	ϵ	ϵ	ϵ	1	ϵ	0	0	0
x_{10}	ϵ	1	ϵ	ϵ	ϵ	ϵ	1	0	0	0	0
x_{11}	ϵ	ϵ	1	ϵ	ϵ	ϵ	ϵ	ϵ	0	0	0

(b) Result of $\delta_{eq}(x_i, x_j)$, Eq. 2.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	0	0	0	0	0	0	1	ϵ	1	ϵ	ϵ
x_2	0	0	0	0	0	0	1	ϵ	1	ϵ	ϵ
x_3	0	0	0	0	0	0	1	ϵ	ϵ	1	ϵ
x_4	0	0	0	0	0	0	ϵ	1	ϵ	1	ϵ
x_5	0	0	0	0	0	0	ϵ	1	ϵ	ϵ	1
x_6	0	0	0	0	0	0	ϵ	1	ϵ	ϵ	1
x_7	1	1	1	ϵ	ϵ	ϵ	0	0	1	ϵ	ϵ
x_8	ϵ	ϵ	ϵ	1	1	1	0	0	ϵ	1	1
x_9	1	1	ϵ	ϵ	ϵ	ϵ	1	ϵ	0	0	0
x_{10}	ϵ	ϵ	1	1	ϵ	ϵ	ϵ	1	0	0	0
x_{11}	ϵ	ϵ	ϵ	ϵ	1	1	ϵ	1	0	0	0

(c) Result of $\delta_{prop}(x_i, x_j)$, Eq. 3.

similarity as a means of comparing groups. That is, $\mathbf{A}_{jj} = 0$, then the non-diagonal entries are determined using one of two similarity functions depending on ϕ as below.

$$\mathbf{A}_{ij} = \begin{cases} \delta_{eq}(x_i, x_j), & \text{for } \phi = \text{equal rank fairness} \\ \delta_{prop}(x_i, x_j), & \text{for } \phi = \text{proportional rank fairness} \end{cases} \quad (1)$$

Next, we describe how we determine the within-a-group similarly functions, e.g., $\delta_{eq}(x_i, x_j)$ and $\delta_{prop}(x_i, x_j)$. An example can be seen in Figure 3. The proposed within-a-group similarity functions are easy to calculate. This provides practitioners with a flexible yet user-friendly solution that is relatively simple to implement.

For the fairness objective of *equal rank fairness*, we operationalize the fair fusion principle by defining that candidates holding the same position within their respective groups are exactly similar and should thus be treated similarly. In other words, candidate item x_o that has the second-highest fusion score in their group g_u is "similar" to x_e, which also has the second-highest fusion score in their group g_v. In contrast, x_o is not similar to candidate x_q, which has the eighth-highest fusion score in their group g_u. This is formally expressed as:

$$\delta_{eq}(x_i, x_j) = \begin{cases} 1, & \text{if } |\tau(x_i) - \tau(x_j)| < 1 \\ \epsilon, & \text{otherwise} \end{cases} \quad (2)$$

where $\tau(x_i)$ is the one-indexed ranked position of x_i in a ranking of x_i's respective group, when ordered by decreasing fusion scores f, and $\epsilon = 0.00001$.

For realizing *proportional rank fairness*[2], we apply the fair fusion principle by proposing that two candidates are similar if their positions with their respective groups are in the same percentile. This effectively adjusts equal treatment, as in Eq. 2, by the potentially distinct sizes of groups to represent them proportionally. In proportional representation, unlike in equal representation, the sum of the rows of $\mathbf{A}$ can vary drastically due to group sizes. For this reason, we normalize $\mathbf{A}$, once it is determined. To make this more concrete, we define $min_{x_i,x_j} \in \{x_i, x_j\}$ as the candidate among x_i or x_j that belongs to the smaller group of size θ_{gmin}, and $max_{x_i,x_j} \in \{x_i, x_j\}$ to be the candidate that belongs to the larger group of size θ_{gmax}. If the sizes of the two compared groups are equal, candidates are

[2]Proportional is defined as proportional to the group's share of the population.

Table 3: Overview of real-world datasets.

Dataset	n	$\|X\|$	Protected Attribute	Groups	$\mathcal{R}$ Form	Bias Mitigation	Conjoint
Adult	4	11687	Race	5	$\mathcal{R}_p$	Equal	Yes
Happiness	17	153	Continent	5	$\mathcal{R}_p$	Prop.	No
Ibm-Hr	4	1462	Age	5	$\mathcal{R}_s$	Prop.	Yes
Economic	10	145	Bank Region	7	$\mathcal{R}_s$	Equal	No

chosen at random. As a formal expression,

$$\delta_{prop}(x_i, x_j) = \begin{cases} 1, & \text{if } |\lceil \frac{\tau(max_{x_i,x_j})}{(\theta_{gmax}/\theta_{gmin})} \rceil - \tau(min_{x_i,x_j})| < 1 \\ \epsilon, & \text{otherwise} \end{cases} \quad (3)$$

where $\tau(x_i)$ and ϵ are the same as defined in Eq. 2.

Wise uses Eq. 2 and Eq. 3 repectively to generate matrix $\mathbf{A}$ that is then utilized to debias the standard fusion scores f into f^* based on parameter λ (line 10). The regularization expression in line 10 is the previously discussed closed-from regularization formula introduced in Zhou et al. [100] and Liu et al. [64]. These matrix operations regularize a set of candidate scores (in our case, f) by candidate pairwise information (in our case, the proposed matrix A). In line 10, $\mathbf{D}$ is a diagonal normalizing matrix so that $\mathbf{D}_{ii} = \sum_{j=1}^{m} \mathbf{A}_{ij}$, $\mathbf{I}$ is an $m \times m$ identity matrix, and λ is our input fairness control parameter. Wise produces the final fused ranking by sorting items by their decreasing now debiased fusion scores f^*.

5 Experimental Evaluation

5.1 Real-World and Synthetic Datasets

5.1.1 Real-world Datasets. The real-world datasets consist of base rankings with only positional orderings of candidates (e.g., base rankings $\mathcal{R}_p$) and base rankings with positional orderings and relevance scores (e.g, base rankings $\mathcal{R}_s$). Table 1 presents a summary of dataset characteristics and their use.

For positional base rankings $\mathcal{R}_p$, we use the Adult [53] fairness benchmark dataset, containing to-be fused rankings of people whose income is $\geq \$50K$, the protected attribute is race. This dataset contains *conjoint base rankings*. The second dataset Happiness [43] contains annual rankings of countries that are fused into an overall ranking; the protected attribute is geographical continents. The base rankings have a strong bias toward European countries and are *non-conjoint* not every country is in each year's ranking.

For score-enhanced base rankings, $\mathcal{R}_s$, we use Ibm-Hr [45] which ranks and scores employees by performance indicators. These are *conjoint base rankings*. Groups are five distinct age categories; employees in their 30s are most preferred. Lastly, Economic [46] provides annual rankings of countries by economic freedom over 10 years. These base rankings are *non-conjoint*. Groups are World Bank Regions, and base rankings contain a strong European bias.

5.1.2 Synthetic Datasets. We create two synthetic datasets to perform controlled studies on (1.) the spectrum of disjoint to conjoint base rankings in Section 5.5 and (2.) the influence of base ranking agreement in Section 5.6.

First, we generate base rankings with the popular *Mallows model* [47, 65]. Given a reference ranking, the model's spread parameter α, is increased to generate base rankings that contain more agreement with the provided reference ranking. For $n = 50$ base rankings, and $|X| = 100$ candidates, we create a reference ranking stacking four groups of 25 candidates each on top of each other, thereby creating an extremely biased ranking. Since groups are equally sized, NDKL *eq.* is equal to NDKL *prop.* [40]. We refer to the set of six base rankings parametrized by α as the Mallows dataset.

Second, we utilize the German Credit dataset [62], which has $|X| = 1000$ candidates to generate base rankings ranging from disjoint to conjoint (varying degrees of candidate overlap). We use a parameter δ representing the proportion of overlapping candidates, i.e., shared amongst all base rankings. Then, for $\delta \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, we sample without replacement δ candidates to determine the fixed set of candidates that will appear in all base rankings. For each of the 10 base rankings, we sample without replacement $100 * (1 - \delta)$ candidates. The final base rankings sort candidates by their credit amount. The groups are males and females. We refer to the set of six base rankings parametrized by δ as Overlap Credit.

Table 4: Runtimes, in seconds, of all methods for the study described in Section 5.4. Base is Combmnz on Ibm-Hr and Economic, and Borda on Happiness and Adult datasets.

Method	Ibm-Hr	Economic	Happiness	Adult
Wise	19.510	0.244	0.313	2533.452
Base	0.156	0.031	0.005	0.055
Post-Eg	0.202	0.043	0.012	2.054
Post-Fq	282.156	5.988	8.727	40086.445
Pre-Eg	0.325	0.070	0.049	8.110
Pre-Fq	1114.646	40.595	77.751	161013.532
Rapf	9.493	0.068	0.043	3886.102
Epira	226.095	-	-	37539.840

5.2 Compared Methods

We compare Wise with six state-of-the-art methodologies, including (1) fairness-unaware score-powered rank fusion (SRF) methods, (2) fair-ranking augmented SRF methods and (3) fair rank aggregation (FRA) methods.

Core Fairness-Unaware SRF Baselines. We use two prominent SRF methods, one for fusing positional rankings $\mathcal{R}_p$ and the other for score-based rankings $\mathcal{R}_s$.

i.) **Borda** [6]: This SRF receives positional $\mathcal{R}_p$ base rankings.

ii.) **Combmnz** [39]: This SRF method uses both candidate positions and scores to produce the fused ranking.

Fair-Ranker Augmented SRF Methods. Since Wise integrates with a given SRF method (Section 4.2), we compose comparative methods by using fair ranking algorithms to either *fairness re-rank base rankings prior* to conventional SRF fusion or *fairness re-rank the SRF-produced fused ranking*.

We select two commonly used fair rankers: Epsilon-Greedy [37] and Fair Queues [42]. Epsilon-Greedy repeatedly swaps pairs of items so that each item has probability ε of swapping with a random item below it. As suggested in [37], we use $\varepsilon = 0.6$. Fair Queues [42], a multi-group version of the FA*IR [96] algorithm, ensures that the difference in fairness of exposure between groups is $\leq DDP$. Following [42], we use $DDP = 0.1$. Using these methods, we compare with the four approaches below.

iii.) **Pre-Eg**: Epsilon-Greedy [37] re-ranks rankings pre-fusion.

iv.) **Post-Eg**: Epsilon-Greedy [37] re-ranks the fused ranking.

v.) **Pre-Fq**: Fair Queues [42] re-ranks rankings pre-fusion.

vi.) **Post-Fq**: Fair Queues [42] re-ranks the fused ranking.

Post-Eg, Pre-Eg, Post-Fq, Pre-Fq, and Wise integrate with an SRF method. For a level comparison, when the input is positional-only base rankings $\mathcal{R}_p$, we utilize Borda. We denote this by appending $^*B$ to the method name. When the input also includes scores, i.e., $\mathcal{R}_s$ base rankings, we utilize Combmnz, and append $^*C$.

Fair Rank Aggregation (FRA) methods. We also study two recent FRA methods - Epira [17] and Rapf [89]. As described in Section 2, these methods ignore any scores included with base rankings. Both explicitly assume base rankings are conjoint ranked lists. However, only Epira truly breaks on non-conjoint base rankings [18]; thus, we only use it on base rankings of identical candidates.

vii.) **Epira** [17]: From the distance-based FRA methods (Table 1), this method handles multiple groups and is a fast approximation of its ILP-based siblings, which are restricted to < 200 candidates [17, 18, 54]. Epira ensures the min-max ratio in fairness of exposure between groups is $\geq \gamma$. As suggested in [17], we use $\gamma = 0.9$.

viii.) **Rapf** [89]: A randomized FRA method (Table 1) that ensures the fused ranking satisfies the proportional p-fairness notion proposed by Wei et al. [89]. Rapf randomly selects a single base ranking and re-ranks it to satisfy as much p-fairness as possible. Thus, it has no fairness-controlling parameter.

Implementation Details. We utilize implementations provided by the authors of Fair Queues (Pre-Fq, Post-Fq), Rapf, and Epira [3]. We implement the other methods for which no public implementation exists ourselves. Unless otherwise specified, as in the parameter tuning study in Section 5.7, we set $\lambda = 0.9$ for Wise. We make all our source and experimental code available at https://github.com/KCachel/wisefuse.

5.3 Metrics for Evaluation

Fairness is measured by NDKL as defined in Table 2. The feasible ranges of NDKL *eq.* and NDKL *prop.* vary based on the candidate set X and its group sizes [40, 81]. They should be interpreted *within a dataset* by comparing how different methods change the score up or down as opposed to comparing these scores across datasets.

Utility is measured through ARBO (Table 2). As related work suggests, preserving the initial orderings *within each group* can be a desirable form of utility in some contexts [92, 96]. We also measure how well a fused ranking preserves within-a-group orderings. To get a single score for this *group utility* notion, we measure:

$$\text{WG-RBO}(R^*, R') = \frac{1}{|G|} \sum_{\forall g_i \in G} rbo(\text{order of } g_i \in R^*, \text{order of } g_i \in R') \tag{4}$$

where r^* is the evaluated fused ranking, r' Borda for $\mathcal{R}_p$ datasets, and Combmnz for $\mathcal{R}_s$ datasets. Its highest (best) value is 1.

Inspired by prior work [44, 76], using the above metrics, we also study *how predictably methods trade off utility with fairness* when method-specific fairness parameters are adjusted.

Computational efficiency is measured as the average time of five runs per method. While further optimizations and computing setups might affect the reported runtimes, it is expected that relative trends in runtime would be observed.

[3]The code repository contains code references.

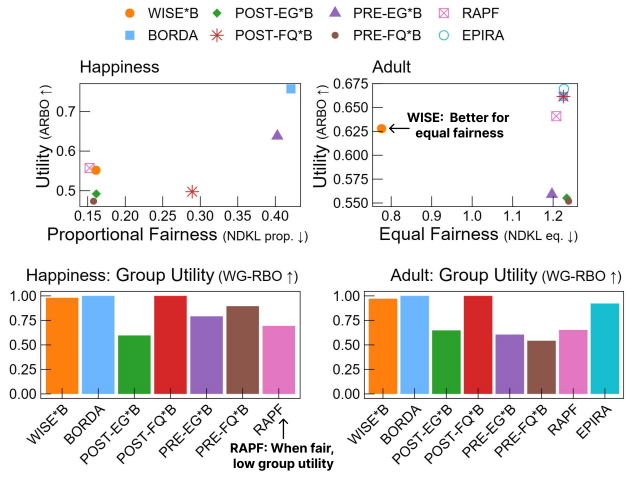

Figure 4: Fusing positional-only rankings $\mathcal{R}_p$. Wise outperforms the compared methods for equal rank fairness, and Rapf does well in proportional settings but generally has one of the lowest group utility values.

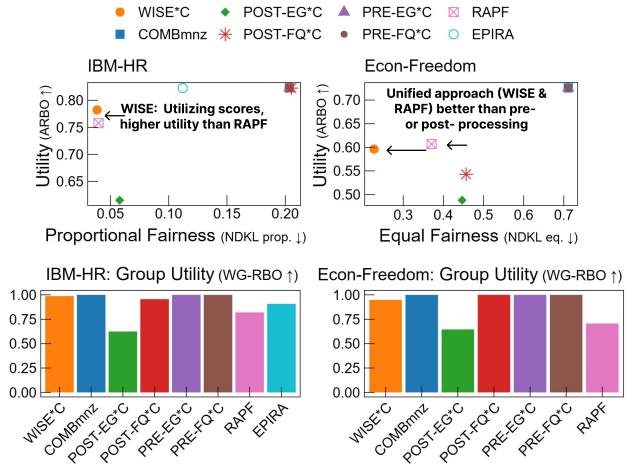

Figure 5: Fusing score-enhanced rankings $\mathcal{R}_s$. By utilizing candidate scores Wise provides more utility for similar fairness provided by Rapf. Unified approaches like Wise, Rapf, and Epira offer better fairness than pre- (Pre-Eg) or post-processing (Pre-Fq) an SRF method with a fair ranker.

5.4 Comparative Study on Real-world Datasets

Figures 4 and 5 plot the fairness, utility, and within-a-group ordering preservation results of each method for $\mathcal{R}_p$ and $\mathcal{R}_s$ base rankings, respectively. Then each setting contains one dataset where the objective is either equal or proportional fairness. Table 4 displays the runtime for each method. In the tested experimental settings, when compared to the baseline methods, Wise achieves the best fairness performance without drastically degrading utility.

Notably, Wise always occupies the most desirable top-left position in the Figure 4 and Figure 5 scatterplots. In the fairness-utility

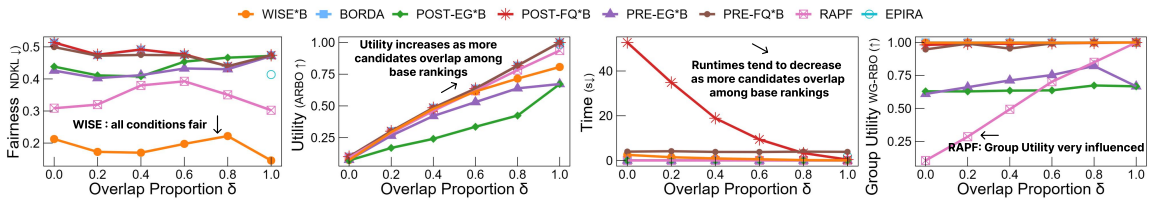

Figure 6: Influence of candidate overlap in base rankings. Wise can achieve the best fairness performance under all levels of candidate overlap. All methods have lower utility and higher runtimes when base rankings have fewer overlapping candidates.

scatterplots, it is always furthest left, indicating the most bias mitigation, for both equal and proportional fairness. When another method such as Rapf in the Ibm-Hr dataset or Post-Eg in the Happiness dataset has a similar fairness level, Wise's utility value (y-axis) is higher – indicating it provides more overall utility for the same level of bias mitigation. Wise also has high WG-RBO values in Figures 4 and 5 – only surpassed by methods that offer no bias mitigation (e.g., Borda or Combmnz). Lastly, Table 4 shows Wise's effective debiasing yields reasonable runtimes.

After Wise, the next best methods are Rapf, Epira, and Post-Eg, depending on the metric and input. Rapf is closest to Wise in fairness performance – it tends to have the second lowest NDKL values. However, as seen in Adult in Figure 4, it performs poorly when the objective is equal fairness. Moreover, it tends to have drastically lower WG-RBO values; meaning items are unnecessarily shuffled around within groups. For instance, in the Economic dataset, Rapf has WG-RBO of 0.71 compared to 0.95 in Wise. Next, Epira also provides bias mitigation – though, as expected, since it is geared toward proportional rank fairness, it does little to aid equal rank fairness. Its primary drawbacks are inefficiency and that it only handles the special case of base rankings with conjoint ranked lists. Lastly, Epsilon-Greedy was designed to be fast [37]. Thus, as one would expect, it is the fastest method after the fairness-unaware methods. It contributes significantly towards mitigating bias but, due to its candidate-swapping approach, has almost always the worst utility (e.g., as in Ibm-Hr and Happiness).

The last tier of methods includes Pre-Eg, Post-Fq, Pre-Fq and fairness-unaware Borda and Combmnz. As expected Borda and Combmnz have the best WG-RBO, ARBO, and runtimes. But they provide no bias mitigation. It is difficult to establish a clear order in performance among Pre-Eg, Post-Fq, and Pre-Fq. Between Post-Fq and Pre-Fq, Post-Fq appears slightly better. It mitigates some bias in two out of the four datasets (Economic and Happiness). Whereas Pre-Fq only has debiasing effects in Happiness.

In general, the fact that Pre-Eg performs worse than Post-Eg, which is also true for Pre-Fq and Post-Fq, indicates that post-processing is a significantly better approach to fair rank fusion than pre-processing. This is because post-processing alters a utility-maximized but unfair fused ranking to predictably decrease utility until the desired fairness is enforced. In contrast, pre-processing alters base rankings prior to fusion, yielding unpredictable results in both objectives. Nonetheless, unlike Wise, these methods are not near the top-left corner (ideal position) of the fairness-utility scatterplots in Figures 4 and 5.

5.5 Varying Item Overlap in Base Rankings

In Figure 6, using the Overlap Credit dataset, we study how varying levels of candidate overlap among base rankings affect Wise and its compared methods. Parameter δ moves from base rankings ordering completely disjoint candidates ($\delta = 0$) to all base rankings ordering the exact same set of candidates ($\delta = 1$). An implicit effect of increasing overlap is that as δ increases, the overall number of candidates $|X|$ decreases. We conclude, as in Section 5.4, there are two classes of methods. Those that (1.) provide the most bias mitigation, (which in Figure 6 includes all explicit fair rank fusion methods Wise, Rapf, and Epira), and (2.) the remaining methods that provide very little to negligible bias mitigation.

In the first class, Wise offers more bias mitigation than Rapf. Specifically, when $\delta \leq 0.6$, Wise provides more bias mitigation for the same amount of utility. Then $\delta \geq 0.6$ (rankings have more candidate overlap), Wise's lower ARBO can be attributed to the fact that Wise achieves more fairness. Thus, the fairness-utility tradeoff explains its lower utility. Since Rapf turns a single base ranking into the fused ranking, it has a steep drop in WG-RBO as base rankings become less conjoint (smaller δ). Then Epira has the weakest performance in the first class. While it only handles conjoint base rankings ($\delta = 1$), even in this setting, it has less bias mitigation and a longer runtime compared to Rapf or Wise.

Lastly, because increasing candidate overlap decreases the number of candidates, we conclude that the utility and runtimes of all methods are positively affected as the candidate overlap increases in the base rankings. This is best seen in Post-Fq in which the Fair Queues method takes less time to run as the number of candidates decreases (e.g., δ increases).

5.6 Varying Agreement Among Base Rankings

In Figure 7, using the Mallows dataset, we study how varying levels of agreement in to-be-fused base rankings impact Wise and its compared methods. Mallows contains 6 different sets of base rankings parameterized by α. When α increases, the base rankings align more and more with a central biased ranking. Thus, Figure 7 displays how each method is influenced by varying base ranking agreement from low to high.

Under the studied agreement conditions in to-be-fused base rankings, Wise has the most bias-mitigation with the highest utility. Of the other best debiasing methods Epira, Post-Eg, and Rapf, only Rapf has comparable fairness values, but Wise has higher utility values than Rapf for all agreement conditions. Wise also has the highest WG-RBO utility values. Second, while the ability of Wise to mitigate bias is unaffected by base ranking agreement conditions,

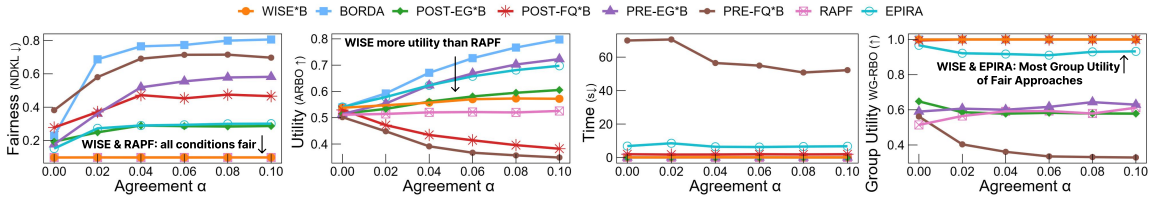

Figure 7: Influence of base ranking agreement. Wise is fair under all agreement conditions in the base rankings. Moreover, it has higher utility than the also fair Rapf. Fairness of Borda, Pre-Fq, and Pre-Eg is strongly affected by the levels of agreement.

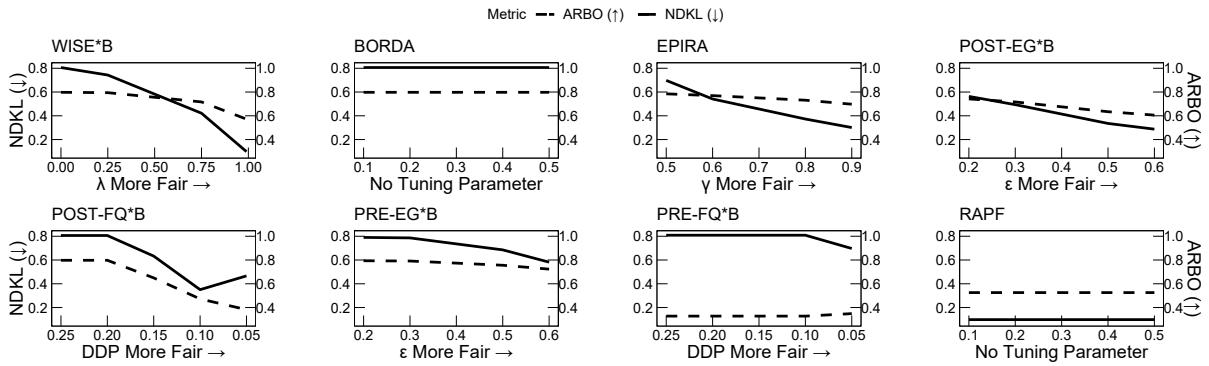

Figure 8: Fairness controllability of all methods. Each method's chart displays fairness (NDKL) on the left y-axis and utility on the right y-axis (ARBO). The x-axis displays variations in the fairness-control parameter of the respective method.

Wise's actual ARBO scores are affected by agreement. This is a natural effect of fusion. When base rankings are aligned and more identical, there is less "compromise" and thus higher utility in the fused ranking. Whereas when base rankings disagree, there is more "compromise" in the fused ranking. This can be seen in the fairness-unaware Borda method – when agreement increases along the x-axis, the utility also increases. Regarding runtimes, we observe that methods are mostly unaffected by agreement except for Pre-Fq.

5.7 Controllability of Fairness-Utility Tradeoff

Lastly, we compare to what degree methods can control the tradeoff between fairness and utility. Using the Mallows $\alpha = 0.1$ dataset, Figure 8 plots fairness (left y-axis) and utility (right y-axis) scores for five fairness-tuning parameter settings for each method. Neither Borda nor Rapf have a parameter for controlling bias-mitigation.

Predominantly for methods that feature fairness-tuning parameters, e.g, Wise, Epira, Post-Eg, Post-Fq, Pre-Eg, Pre-Fq, we observe, as expected, that as the tuning parameter is adjusted, unfairness and utility decrease accordingly. However, some methods (e.g., Post-Eg, Post-Fq, Pre-Eg, Pre-Fq) provide less tuning power, meaning they don't span quite as large a range of NDKL values as Wise or Epira. In the case of Pre-Fq, the *DDP* parameter does not apply much bias mitigation as seen by the relatively straight lines. We also see this, although to a lesser degree, in Post-Eg and Pre-Eg. Post-Fq appears somewhat unpredictable as unfairness dips down and then up. Lastly, Epira and Wise both provide predictable fairness-utility trade-off tuning over the broadest range. However, the drawbacks of Epira, as presented in our prior experimental Sections 5.4 and 5.5, lead us to conclude that Wise offers better bias-mitigation performance, in addition to the ability to easily control the fairness-utility tradeoff.

6 Conclusion and Future Work

In this work, we introduce Wise a new plug-and-play fair rank fusion approach. Wise can fairly fuse a variety of to-be-fused rankings previously unaddressed by existing fair rank aggregation methods (e.g., non-conjoint and score-augmented base rankings). Wise, operationalizes our proposed fair fusion principle. Wise works in any fusion setting where we have closed-box access to a score-powered rank fusion (SRF) method, making it relatively easy to fairness-enhance existing fusion pipelines. We demonstrate that, compared to existing alternatives from the literature, Wise achieves the best bias-mitigation and utility performance while providing practitioners with an easy-to-control and efficient approach.

*Future Work.*In the design of Wise, we assumed that the protected group attribute is accurately given. To relax this assumption, more work is needed to extend Wise to incorporate either noisy protected attribute information (probabilities) or function without protected attributes. To offer additional functionality, future work could study ways to bound the level of desired utility or fairness in the Wise methodology, as well as the effect of ties within rankings.

Acknowledgments

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

References

[1] Nir Ailon. 2010. Aggregation of Partial Rankings, p-Ratings and Top-m Lists. *Algorithmica* 57, 2 (feb 2010), 284–300.

[2] Nir Ailon, Moses Charikar, and Alantha Newman. 2008. Aggregating inconsistent information: Ranking and clustering. *J. ACM* 55, 5, Article 23 (nov 2008), 27 pages. https://doi.org/10.1145/1411509.1411513

[3] Alnur Ali and Marina Meilă. 2012. Experiments with Kemeny ranking: What works when? *Mathematical Social Sciences* 64, 1 (2012), 28–40. https://doi.org/10.1016/j.mathsocsci.2011.08.008 Computational Foundations of Social Choice.

[4] Natasha Allden and Lisa Harris. 2013. Building a positive candidate experience: towards a networked model of e-recruitment. *The Journal of business strategy* 34, 5 (2013), 36–47.

[5] Avi Arampatzis and Jaap Kamps. 2009. A signal-to-noise approach to score normalization. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (Hong Kong, China) *(CIKM '09)*. Association for Computing Machinery, New York, NY, USA, 797–806. https://doi.org/10.1145/1645953.1646055

[6] Javed A. Aslam and Mark H. Montague. 2001. Models for Metasearch. In *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, W. Bruce Croft, David J. Harper, Donald H. Kraft, and Justin Zobel (Eds.). ACM, 275–284. https://doi.org/10.1145/383952.384007

[7] 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

[8] Aparna Balagopalan, Abigail Z. Jacobs, and Asia J. Biega. 2023. The Role of Relevance in Fair Ranking. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (<conf-loc>, <city>Taipei</city>, <country>Taiwan</country>, </conf-loc>) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2650–2660. https://doi.org/10.1145/3539618.3591933

[9] Linas Baltrunas, Tadas Makcinskas, and Francesco Ricci. 2010. Group recommendations with rank aggregation and collaborative filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (Barcelona, Spain) *(RecSys '10)*. Association for Computing Machinery, New York, NY, USA, 119–126. https://doi.org/10.1145/1864708.1864733

[10] Elias Bassani and Luca Romelli. 2022. ranx.fuse: A Python Library for Metasearch. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 4808–4812. https://doi.org/10.1145/3511808.3557207

[11] Nadja Betzler, Michael R. Fellows, Jiong Guo, Rolf Niedermeier, and Frances A. Rosamond. 2009. Fixed-parameter algorithms for Kemeny rankings. *Theoretical Computer Science* 410, 45 (2009), 4554–4570. https://doi.org/10.1016/j.tcs.2009.08.033 Algorithmic Aspects in Information and Management.

[12] 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

[13] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*. 405–414.

[14] Avrim Blum and Kevin Stangl. 2020. Recovering from Biased Data: Can Fairness Constraints Improve Accuracy?. In *1st Symposium on Foundations of Responsible Computing (FORC 2020) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 156)*, Aaron Roth (Ed.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 3:1–3:20. https://doi.org/10.4230/LIPIcs.FORC.2020.3

[15] Amanda Bower, Hamid Eftekhari, Mikhail Yurochkin, and Yuekai Sun. 2021. Individually Fair Rankings. In *International Conference on Learning Representations*.

[16] 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

[17] 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

[18] 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

[19] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems* 30 (2017).

[20] 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

[21] Stefan Chanas and Przemysław Kobylanski. 1996. A new heuristic algorithm solving the linear ordering problem. *Comput. Optim. Appl.* 6, 2 (sep 1996), 191–205. https://doi.org/10.1007/BF00249646

[22] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the Impact of Gender on Rank in Resume Search Engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Montreal QC</city>, <country>Canada</country>, </conf-loc>) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174225

[23] 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.

[24] Vincent Conitzer and Tuomas Sandholm. 2005. Communication complexity of common voting rules. In *Proceedings of the 6th ACM Conference on Electronic Commerce* (Vancouver, BC, Canada) *(EC '05)*. Association for Computing Machinery, New York, NY, USA, 78–87. https://doi.org/10.1145/1064009.1064018

[25] 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

[26] 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

[27] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 473–484. https://doi.org/10.1145/3531146.3533113

[28] Persi Diaconis and R. L. Graham. 1977. Spearman's Footrule as a Measure of Disarray. *Journal of the Royal Statistical Society: Series B (Methodological)* 39, 2 (1977), 262–268. https://doi.org/10.1111/j.2517-6161.1977.tb01624.x arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1977.tb01624.x

[29] Fernando Diaz. 2005. Regularizing ad hoc retrieval scores. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management* (Bremen, Germany) *(CIKM '05)*. Association for Computing Machinery, New York, NY, USA, 672–679. https://doi.org/10.1145/1099554.1099722

[30] 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

[31] 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

[32] Miles Efron. 2009. Generative model-based metasearch for data fusion in information retrieval. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries* (Austin, TX, USA) *(JCDL '09)*. Association for Computing Machinery, New York, NY, USA, 153–162. https://doi.org/10.1145/1555400.1555426

[33] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. Fairness in Information Access Systems. *Found. Trends Inf. Retr.* 16, 1–2 (jul 2022), 1–177. https://doi.org/10.1561/1500000079

[34] Edith Elkind, Piotr Faliszewski, Piotr Skowron, and Arkadii Slinko. 2014. Properties of multiwinner voting rules. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems* (Paris, France) *(AAMAS '14)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 53–60.

[35] Edith Elkind, Piotr Faliszewski, and Arkadii Slinko. 2010. On the role of distances in defining voting rules. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1* (Toronto, Canada) *(AAMAS '10)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 375–382.

[36] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge*

discovery and data mining. 259–268.

[37] 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

[38] Andres Ferraro, Xavier Serra, and Christine Bauer. 2021. What is fair? Exploring the artists' perspective on the fairness of music streaming platforms. In *IFIP conference on human-computer interaction*. Springer, 562–584.

[39] Edward A. Fox and Joseph A. Shaw. 1993. Combination of Multiple Searches. In *TREC (NIST Special Publication, Vol. 500-215)*. National Institute of Standards and Technology (NIST), 243–252.

[40] 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

[41] Avijit Ghosh, Ritam Dutt, and Christo Wilson. 2021. When Fair Ranking Meets Uncertain Inference. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (<conf-loc>, <city>Virtual Event</city>, <country>Canada</country>, </conf-loc>) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1033–1043. https://doi.org/10.1145/3404835.3462850

[42] Ananya Gupta, Eric Johnson, Justin Payan, Aditya Kumar Roy, Ari Kobren, Swetasudha Panda, Jean-Baptiste Tristan, and Michael Wick. 2021. Online post-processing in rankings for fair utility maximization. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 454–462.

[43] 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).

[44] Maria Heuss, Daniel Cohen, Masoud Mansoury, Maarten de Rijke, and Carsten Eickhoff. 2023. Predictive Uncertainty-based Bias Mitigation in Ranking. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (<conf-loc>, <city>Birmingham</city>, <country>United Kingdom</country>, </conf-loc>) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 762–772. https://doi.org/10.1145/3583780.3615011

[45] IBM. 2016. *IBM HR Analytics Employee Attrition & Performance*. Technical Report. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

[46] The Fraser Institute. 2023. *Economic Freedom of the World 2023 Annual Report*. Technical Report. https://www.fraserinstitute.org/economic-freedom/dataset

[47] 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.

[48] N. Jardine and C.J. van Rijsbergen. 1971. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval* 7, 5 (1971), 217–240. https://doi.org/10.1016/0020-0271(71)90051-9

[49] 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

[50] John G Kemeny. 1959. Mathematics without numbers. *Daedalus* 88, 4 (1959), 577–591.

[51] Anna Khudyak Kozorovitsky and Oren Kurland. 2011. Cluster-based fusion of retrieved lists. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 893–902. https://doi.org/10.1145/2009916.2010035

[52] Jon Kleinberg and Manish Raghavan. 2018. Selection Problems in the Presence of Implicit Bias. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 94)*, Anna R. Karlin (Ed.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 33:1–33:17. https://doi.org/10.4230/LIPIcs.ITCS.2018.33

[53] Ronny Kohavi and Barry Becker. 1996. UCI machine learning repository: adult data set. *Avaliable: https://archive. ics. uci. edu/ml/machine-learning-databases/adult* (1996).

[54] 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

[55] S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22, 1 (1951), 79 – 86. https://doi.org/10.1214/aoms/1177729694

[56] Oren Kurland and J. Shane Culpepper. 2018. Fusion in Information Retrieval: SIGIR 2018 Half-Day Tutorial. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (2018).

[57] 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.

[58] Joon Ho Lee. 1997. Analyses of multiple evidence combination. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Philadelphia, Pennsylvania, USA) *(SIGIR '97)*. Association for Computing Machinery, New York, NY, USA, 267–276. https://doi.org/10.1145/258525.258587

[59] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. 2023. Fairness in Recommendation: Foundations, Methods, and Applications. *ACM Trans. Intell. Syst. Technol.* 14, 5, Article 95 (oct 2023), 48 pages. https://doi.org/10.1145/3610302

[60] Shangsong Liang and Maarten de Rijke. 2015. Burst-aware data fusion for microblog search. *Information Processing & Management* 51, 2 (2015), 89–113. https://doi.org/10.1016/j.ipm.2014.10.008

[61] Shangsong Liang, Ilya Markov, Zhaochun Ren, and Maarten de Rijke. 2018. Manifold Learning for Rank Aggregation. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1735–1744. https://doi.org/10.1145/3178876.3186085

[62] M Lichman. 2013. UCI Machine Learning Repository: Statlog (German Credit Data) Data Set.

[63] 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

[64] Wei Liu, Jun Wang, and Shih-Fu Chang. 2012. Robust and Scalable Graph-Based Semisupervised Learning. *Proc. IEEE* 100, 9 (2012), 2624–2638. https://doi.org/10.1109/JPROC.2012.2197809

[65] Colin L Mallows. 1957. Non-null ranking models. *Biometrika* 44, 1/2 (1957), 114–130.

[66] R. Manmatha and H. Sever. 2002. A formal approach to score normalization for meta-search. In *Proceedings of the Second International Conference on Human Language Technology Research* (San Diego, California) *(HLT '02)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 98–103.

[67] Ilya Markov, Avi Arampatzis, and Fabio Crestani. 2012. Unsupervised linear score normalization revisited. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Portland, Oregon, USA) *(SIGIR '12)*. Association for Computing Machinery, New York, NY, USA, 1161–1162. https://doi.org/10.1145/2348283.2348519

[68] Mark Montague and Javed A. Aslam. 2001. Relevance score normalization for metasearch. In *Proceedings of the Tenth International Conference on Information and Knowledge Management* (Atlanta, Georgia, USA) *(CIKM '01)*. Association for Computing Machinery, New York, NY, USA, 427–433. https://doi.org/10.1145/502585.502657

[69] Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Serena Wang. 2020. Pairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5248–5255.

[70] Samuel E. L. Oliveira, Victor Diniz, Anisio Lacerda, Luiz Merschmanm, and Gisele L. Pappa. 2020. Is Rank Aggregation Effective in Recommender Systems? An Experimental Analysis. *ACM Trans. Intell. Syst. Technol.* 11, 2, Article 16 (jan 2020), 26 pages. https://doi.org/10.1145/3365375

[71] 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

[72] Christine Pinney, Amifa Raj, Alex Hanna, and Michael D. Ekstrand. 2023. Much Ado About Gender: Current Practices and Future Recommendations for Appropriate Gender-Aware Information Access. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (<conf-loc>, <city>Austin</city>, <state>TX</state>, <country>USA</country>, </conf-loc>) *(CHIIR '23)*. Association for Computing Machinery, New York, NY, USA, 269–279. https://doi.org/10.1145/3576840.3578316

[73] Chuan Qin, Hengshu Zhu, Tong Xu, Chen Zhu, Chao Ma, Enhong Chen, and Hui Xiong. 2020. An Enhanced Neural Network Approach to Person-Job Fit in Talent Recruitment. *ACM Trans. Inf. Syst.* 38, 2, Article 15 (feb 2020), 33 pages. https://doi.org/10.1145/3376927

[74] 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

[75] 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

[76] 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* (<conf-loc>, <city>Virtual Event</city>, <country>Canada</country>, </conf-loc>) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 306–316. https://doi.org/10.1145/3404835.3462949

[77] Jad Salem, Deven Desai, and Swati Gupta. 2022. Don't let Ricci v. DeStefano Hold You Back: A Bias-Aware Legal Solution to the Hiring Paradox. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (<conf-loc>, <city>Seoul</city>, <country>Republic of Korea</country>, </conf-loc>) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 651–666. https://doi.org/10.1145/3531146.3533129

[78] Jad Salem and Swati Gupta. 0. Secretary Problems with Biased Evaluations Using Partial Ordinal Information. *Management Science* 0, 0 (0), null. https://doi.org/10.1287/mnsc.2023.4926 arXiv:https://doi.org/10.1287/mnsc.2023.4926

[79] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attentionon Fair Group Representation in Ranked Lists. In *Companion Proceedings of The 2019 World Wide Web Conference*. 553–562.

[80] Frans Schalekamp and Anke van Zuylen. [n. d.]. *Rank Aggregation: Together We're Strong*. 38–51. https://doi.org/10.1137/1.9781611972894.4 arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611972894.4

[81] Tobias Schumacher, Marlene Lutz, Sandipan Sikdar, and Markus Strohmaier. 2022. Properties of Group Fairness Metrics for Rankings. *ArXiv* abs/2212.14351 (2022).

[82] Ashudeep Singh and Thorsten Joachims. 2017. Equality of opportunity in rankings. In *Workshop on Prioritizing Online Content (WPOC) at NIPS*. 31.

[83] 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

[84] 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.

[85] Jessie J. Smith, Lex Beattie, and Henriette Cramer. 2023. Scoping Fairness Objectives and Identifying Fairness Metrics for Recommender Systems: The Practitioners' Perspective. In *Proceedings of the ACM Web Conference 2023* (<conf-loc>, <city>Austin</city>, <state>TX</state>, <country>USA</country>, </conf-loc>) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 3648–3659. https://doi.org/10.1145/3543507.3583204

[86] Tom Sühr, Sophie Hilgard, and Himabindu Lakkaraju. 2021. Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 989–999. https://doi.org/10.1145/3461702.3462602

[87] Anke van Zuylen and David P. Williamson. 2008. Deterministic Algorithms for Rank Aggregation and Other Ranking and Clustering Problems. In *Approximation and Online Algorithms*, Christos Kaklamanis and Martin Skutella (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 260–273.

[88] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.* 28, 4, Article 20 (nov 2010), 38 pages. https://doi.org/10.1145/1852102.1852106

[89] 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

[90] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 666–677.

[91] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 666–677. https://doi.org/10.1145/3442188.3445928

[92] Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. 2019. Balanced Ranking with Diversity Constraints. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 6035–6042. https://doi.org/10.24963/ijcai.2019/836

[93] 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

[94] H. Peyton Young. 1974. An axiomatization of Borda's rule. *Journal of Economic Theory* 9 (1974), 43–52.

[95] Peyton Young. 1995. Optimal Voting Rules. *Journal of Economic Perspectives* 9, 1 (March 1995), 51–64. https://doi.org/10.1257/jep.9.1.51

[96] 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

[97] 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

[98] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in Ranking, Part I: Score-based Ranking. *ACM Computing Surveys (CSUR)* (2022).

[99] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in Ranking, Part II: Learning-to-Rank and Recommender Systems. *ACM Computing Surveys (CSUR)* (2022).

[100] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. 2003. Learning with local and global consistency. In *Proceedings of the 16th International Conference on Neural Information Processing Systems* (Whistler, British Columbia, Canada) *(NIPS'03)*. MIT Press, Cambridge, MA, USA, 321–328.