Fairer Together: Mitigating Disparate Exposure in Kemeny Rank Aggregation

Kathleen Cachel
Worcester Polytechnic Institute
kcachel@wpi.edu

Elke Rundensteiner
Worcester Polytechnic Institute
rundest@wpi.edu

ABSTRACT

In social choice, traditional Kemeny rank aggregation combines the preferences of voters, expressed as rankings, into a single consensus ranking without consideration for how this ranking may unfairly affect marginalized groups (i.e., racial or gender). Developing fair rank aggregation methods is critical due to their societal influence in applications prioritizing job applicants, funding proposals, and scheduling medical patients. In this work, we introduce the Fair Exposure Kemeny Aggregation Problem (FairExp-kap) for combining vast and diverse voter preferences into a single ranking that is not only a suitable consensus, but ensures opportunities are not withheld from marginalized groups. In formalizing FairExp-kap, we extend the fairness of exposure notion from information retrieval to the rank aggregation context and present a complimentary metric for voter preference representation. We design algorithms for solving FairExp-kap that explicitly account for position bias, a common ranking-based concern that end-users pay more attention to higher ranked candidates. epik solves FairExp-kap exactly by incorporating non-pairwise fairness of exposure into the pairwise Kemeny optimization; while the approximate epira is a candidate swapping algorithm, that guarantees ranked candidate fairness. Utilizing comprehensive synthetic simulations and six real-world datasets, we show the efficacy of our approach illustrating that we succeed in mitigating disparate group exposure unfairness in consensus rankings, while maximally representing voter preferences.

CCS CONCEPTS

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

KEYWORDS

fair exposure rank aggregation, group fairness, rank aggregation, voting rules

ACM Reference Format:
Kathleen Cachel and Elke Rundensteiner. 2023. Fairer Together: Mitigating Disparate Exposure in Kemeny Rank Aggregation. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3593013.3594085

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3594085

1 INTRODUCTION

In social choice, the goal of rank aggregation rules, such as the widely popular Kemeny rule [7], is to combine the vast individual preferences of voters, modeled as rankings, into a single consensus ranking. Both the initial collection of voter input rankings, called a preference profile, and the consensus ranking, order the same set of candidates. Rank aggregation is used to collectively, from diverse preferences, prioritize job applicants for employers [26], patients for medical care [18], and proposals for funding [13]. For each of these contexts it is paramount to generate high-quality consensus rankings that not only represent the individual preferences of voters, but are also unbiased. Namely, rank aggregation (i.e. voting) rules bear responsibility to ensure their consensus ranking does not withhold opportunities or resources from candidates belonging to marginalized or protected groups (e.g., racial or gender).

In this work, we thus focus on modeling and achieving the fair treatment of ranked candidates in consensus rankings. The fair ranking of candidates is complicated by the phenomenon, known as position bias [25], i.e., that the end user's attention is concentrated on higher positioned candidates, and their attention decays or altogether stops before the entire consensus ranking has been seen. The attention a ranked candidate receives, called *exposure*, is tied to the candidate's position in the ranking. Mitigating disparate group exposure ensures fairness for ranked candidates in the presence of position bias, which commonly plagues the practical use of consensus rankings.

Unfortunately, prior work studying fairness concerns in Kemeny rank aggregation [9, 29, 43] does not guarantee the mitigation of disparate exposure amongst ranked candidate groups. As we show in Section 7.3, the representation-based method of [43] and the cross-group pairwise method of [9, 28] sometimes even increase disparate exposure relative to the traditional Kemeny rule. Moreover, in [43] the consensus ranking is generated from a *single* voters ranking. Bridging this gap, we address the challenging problem of developing rank aggregation approaches that combine the preferences of numerous voters in a way that mitigates disparate exposure unfairness and maximally represents voter preferences. As the Kemeny rule is pairwise, we must integrate exposure (by design, a listwise concept) into pairwise methods. Further, Kemeny rank aggregation is NP-hard, without considering integrating fairness objectives.

In particular, we extend the commonly accepted fairness of exposure notion from information retrieval [39] to the rank aggregation context, and design a complimentary metric to quantify voter preference representation in a consensus ranking. We formulate these metrics to have intentionally aligned interpretations: range in [0, 1], whereby 1 is best. This ensures easy comparison between the fairness and preference representation objectives they quantify. Utilizing these metrics, we introduce the Fair Exposure Kemeny

Rank Aggregation Problem (for short, FairExp-kap). Addressing this new problem, we design exact and approximate techniques, explicitly handling the concern of position bias and including an easy-to-use parameter controlling how much disparate exposure mitigation is applied to a given scenario.

This controlled ability to tune bias mitigation provides practitioners with the flexibility to incorporate their domain knowledge when deciding on the appropriate amount of fairness intervention. Our exact epik method, explicitly integrates fairness of exposure into the Kemeny rule by finding the ranking that has minimum Kendall tau distance from voter preferences subject to fairness of exposure. Our approximate method epira is highly flexible and provably optimal at representing within-a-group voter preferences (Theorem 6), and we demonstrate it works not only with the Kemeny voting rule, but can be used with any other voting rule. Our work demonstrates that we can succeed in generating consensus rankings in a way that both mitigates disparate exposure and maximally represents the preference of all voters. This research can benefit groups of people who are adversely impacted by traditional consensus ranking processes. This work makes the following technical contributions:

- We formalize the problem of mitigating disparate exposure in Kemeny rank aggregation as FairExp-kap, along with metrics for quantifying its candidate fairness and voter preference representation objectives (Section 4).
- We introduce epik, a technique for solving FairExp-kap exactly, that incorporates non-pairwise disparate exposure constraints into the pairwise Kemeny integer program (Section 5).
- We design epira, an approximation technique for solving FairExp-kap efficiently by repositioning candidates within a pre-computed consensus ranking (our implementation supports five voting rules). We prove that subject to fairness of exposure, it maximally represents within-a-group voter preferences (Section 6).
- Through comprehensive simulations using the Mallows model and experiments on six real-world datasets, we show the efficacy of our approach. We observe that our algorithms outperform both the standard Kemeny rule (Section 7.2) and prior fair Kemeny rule methods for solving FairExp-kap (Section 7.3). We confirm Theorem 5.1 in that epira optimally represents within-a-group voter preferences (Section 7.4) and show the generality of epira using five alternate voting rules. Based on our study we recommend its use with the Copeland rule.

2 RELATED WORK

Kemeny rank aggregation is a well-studied problem in social choice [7, 12, 32, 37]. We present fairness-aware algorithms for Kemeny rank aggregation that mitigate potential disparate exposure among protected groups of candidates (harms of allocation). There is a rich line of related work that also proposes variations to Kemeny rank aggregation. Unlike our work, this literature does not ensure the fair treatment (exposure) of ranked candidates. This includes designing web search engines [3, 16], addressing settings where voters provide partial rankings [48], settings where candidates become unavailable [4], where candidates are voters [26], a privacy-preserving variant [1] and a set-wise variant [19].

Recently methodologies for Kemeny rank aggregation considering other fairness definitions have been proposed [9, 28, 43]; they do not directly translate to our target problem of mitigating disparate exposure amongst candidate groups (as shown in Section 7.3). [9] propose pairwise Fair-Kemeny (PFair-Kem) which extends the procedure from [28] to multiple groups.

PFair-Kem addresses cross-group pairwise comparison fairness in the consensus ranking. Fairness is ensured only when the entire ranking is considered [9]. This is not suitable for applications where the consensus ranking is a list that is used top down or not entirely viewed or displayed [20, 21]. [43] proposes Rank Aggregation p-Fairness (RAPF). RAPF randomly selects a single voter's ranking and corrects it to satisfy p-fairness (a representation-based criteria that each group is proportionally represented) [43], which does not address position bias. Moreover, RAPF lacks a meaningful consensus process, by using a single voter's preference it prioritizes the preference of one voter, whereas the opposite (namely non-dictatorship) is a common and desirable voting rule property [7]. PFair-Kem and RAPF achieve altogether different fairness objectives. Thus, unsurprisingly, when we compare against them in our evaluation study, Section 7.3, they significantly under-perform our introduced techniques in mitigating disparate exposure.

Also, at the nexus of social choice and algorithmic fairness, is fair multi-winner voting [8, 10, 33]. This context differs from our setting since an unordered subset, sometimes called committee, of candidates is produced from voter rankings or chosen committees. This setting selects, but does not order, k candidates from $k < n$ candidates. Additionally, [8, 10, 33] is not applicable to fair rank aggregation due to the distinct difference in fairness concerns between the two problems. In multi-winner voting, unfairness results from demographic groups being underrepresented in or excluded from the chosen candidate committee. Instead, in our rank aggregation context (e.g., ordering all n candidates), unfairness results from groups being denied the attention of the viewer(s) of the consensus ranking.

Lastly, since its introduction in [39], fairness of exposure has underpinned much related literature in fair information retrieval. However, unlike our work, this literature does not consider voting, in particular the setting of multiple rankings, nor the problem of explicit preference aggregation. Examples include fairness metrics [15, 39, 44], learning-to-rank applications [34, 40, 45, 46], outlier-aware fair ranking [23, 36], online fair ranking [22], and multi-stakeholder fair recommendation [42]. For recent surveys of fairness in ranking and recommendation, see [17, 35, 47].

3 PRELIMINARIES

Here, we introduce standard terminology and notation [7], setting up our rank aggregation context, as well as the mathematical notion of fair group exposure.

3.1 Notation for Rankings and Rank Aggregation

Our setting involves finite sets $V = \{v_1, ..., v_n\}$ of n voters and $C = \{c_1, ..., c_m\}$ of m candidates (also called alternatives or items).

The preferences of each voter are represented as a ranking over C. The collection of these rankings expressing voter preferences defines a *preference profile* R. That is, $R = \{r_1, ..., r_n\}$, whereby r_j is the ranking by voter j. Let $r_j^{pos}(c_i)$ denote the position of candidate c_i in the ranking produced by voter j, where position 1 is the highest (best) and position m is the lowest (worst). Let Π_C denote the set of all possible rankings over candidate set C.

As we seek to generate a consensus ranking from R, we are interested in voter aggregation rules that, given profile R, find a *consensus ranking* that is a suitable compromise representing the preferences R. We employ two such rules:

- *Kemeny Rule [27]:* selects a ranking with minimal Kendall tau distance to R. The *Kendall tau distance* d_{KT} between any two rankings is the number of candidate pairs on which the two rankings disagree. Given preference profile R, the Kemeny rule returns a ranking r in $\text{argmin}_{r \in \Pi_C} \sum_{i=1}^{n} d_{KT}(r, r_i)$.
- *Copeland Rule [14]:* orders candidates by decreasing Copeland score summed $\forall\ r \in R$. The Copeland score for candidate c_i in ranking r is $Copeland(c_i, r) = |\{c_j \in C \mid r^{pos}(c_i) < r^{pos}(c_j)\}| - |\{c_j \in C \mid r^{pos}(c_j) < r^{pos}(c_i)\}|$. It orders candidates by the total number of pairwise contests they win over other candidates in the profile R.

3.2 Group Fairness of Exposure in Rankings

In our setting, candidates C have an associated *protected attribute*, such as gender, race, or their combination. Let A be the protected attribute and $G_{A:l}$ be the set of candidates with value l in A. We refer to this candidate set $G_{A:l}$ as a "group". We aim to ensure *group fairness* in the resulting consensus ranking, that is, *a fair group ordering in the ranking.*

In information retrieval, Singh and Joachims [39] introduce the concept of *group exposure*, that is, the attention a group receives in a ranking. First, the *exposure* of a candidate c_i in ranking r is:

$$exposure(r, c_i) = 1/log_2(r^{pos}(c_i) + 1)). \tag{1}$$

For instance, if the ranking is being used for hiring decisions, we can think about the exposure of a candidate as the probability that the candidate is hired. Then the *group exposure* of a group $G_{A:j}$ in ranking r is:

$$group\ exposure(r, G_{A:l}) = |G_{A:l}|^{-1} \sum_{\forall c_i \in G_{A:l}} exposure(r, c_i). \tag{2}$$

This is the average of exposure (Equation 1) of all group members. We can think of group exposure as the likelihood someone from group $G_{A:l}$ is hired. Singh and Joachims [39], for the case of two groups, define a minimal difference in exposure (Equation 2) to indicate a fair ranking. Next, we use the concept of exposure to define our fairness metric.

4 FAIR-EXPOSURE KEMENY AGGREGATION

Given fair rank aggregation is a dual objective task, namely, voter preference representation and ranked candidate fairness, measuring both objectives is critical for formulating our problem. Thus, we introduce metrics to quantify its consensus and fairness objectives before defining our Fair Exposure Kemeny Aggregation problem (Section 4.2).

4.1 Proposed Measurement of Preference Representation and Fairness Objectives

To wholistically quantify fairness across *multiple* groups (as opposed to per group as in Equation 2), we can measure disparate group exposure in two ways. One, which directly adopts the approach of Singh and Joachims [39] by using the maximum absolute difference in exposure across pairs of groups. Or, two, where we use the ratio between the minimum and maximum group exposure. We adopt the latter, since it has two advantages. First, it nicely aligns with contemporary fairness objectives such as the US EEOC [1], "four-fiths" rule which states the least privileged group must receive a proportion of the positive outcome (in our case, exposure) that is at least 80% of the proportion received by the most privileged group [11]. And second, its interpretation is not sensitive to the number of ranked candidates. Definition 4.1 presents the exposure ratio metric, with range [0, 1], where lower values indicate groups have disparate exposure (thus unfairness) and 1 indicates each group has the same exposure (complete fairness).

Definition 4.1. (Exposure Ratio). Given ranking r and protected attribute A, the **exposure ratio** *(ER) of the groups defined by A in ranking r is $ER(r, A) = \frac{\min\{group\ exposure(r, G_{A:l})\}}{\max\{group\ exposure(r, G_{A:k})\}} \forall\ G_{A:l}, G_{A,k}$.*

Next, we turn to our second objective: maximally representing the preference profile R in the consensus ranking. While, traditional Kemeny rank aggregation minimizes the Kendall tau distance d_{KT} between consensus ranking r and profile R, utilizing, $d_{KT}(r, R)$, to measure preference representation in our context has significant drawbacks. First, a lower Kendall tau distance is always preferred, which is the opposite of exposure ratio where a higher value is a better (indicating a more fair ranking). And second, a "good" Kendall tau distance is surprisingly difficult to pinpoint, for instance $d_{KT}(r, R)$ of 10 could be low preference representation if there are 6 candidates, but high preference representation if there are 60. To address these issues, we introduce the notion and measure of consensus accuracy (Definition 4.2), which captures the opposite of the Kendall tau distance, namely, agreement.

Definition 4.2. (Consensus Accuracy). Given consensus ranking r and preference profile R, the **consensus accuracy** *(CA) of r is $CA(r, R) = (\binom{m}{2}n - d_{kt}(r, R))/\sum Q(R)_{xy}$, where Q is a pairwise comparison matrix representing the profile R.*

Specifically, each (x, y) entry of the matrix Q denotes the number of voters agreeing with c_x ranked above c_y, that is:

$$Q(R)_{xy} = \sum_{\forall r_i \in R} |r_i^{pos}(c_x) < r_i^{pos}(c_y)|. \tag{3}$$

In words, consensus accuracy corresponds to the total pairwise agreements between the preference profile R and the consensus ranking r, divided by the sum total of pairwise agreements in R. It ranges from $[0, 1]$, where 1 is the best value, meaning *every single pairwise preference* in R is represented in r. Consensus accuracy quantifies preference representation in a way that is both consistent across the number of ranked candidates, and aligned with exposure ratio (i.e., higher and closer to 1 is "better").

[1] United States Equal Employment Opportunity Commission

4.2 Fair Exposure Kemeny Aggregation Problem

With our fairness and preference representation objectives in place, along with their respective measures (*ER* and *CA*), we present the Fair Exposure Kemeny Aggregation Problem.

Problem 1. *(FairExp-kap). Given preference profile R (of n voters ranking candidate set C), where each candidate belongs to a group defined by protected attribute A, and desired exposure ratio $\lambda \in (0, 1]$ the* **Fair Exposure Kemeny Aggregation Problem (FairExp-kap)** *is to find consensus ranking $r \in \Pi_C$, such that :*

(1.) *the exposure ratio (Definition 4.1) of consensus ranking r is greater than parameter λ, i.e, $ER(r, A) \geq \lambda$, and*
(2.) *the consensus accuracy (Definition 4.2) of r, CA(r, R), is maximized.*

In short, the FairExp-kap problem is to determine the consensus ranking via the Kemeny rule, where the primary objective is the consensus ranking has fairness $ER(r, A) \geq \gamma$. Maximizing consensus accuracy, is in fact, the Kemeny rule, as maximizing pairwise agreements is the same as minimizing pairwise disagreements (i.e, the minimized Kendall tua distance in the traditional Kemeny rule). This approach, compared to maximizing fairness subject to maintaining the consensus accuracy of the fairness-unaware Kemeny ranking, ensures that the resulting ranking satisfies a provided degree of fairness (γ), critically facilitating consensus rankings fairer than the Kemeny ranking.

By emphasizing fairness for ranked candidates in FairExp-kap, we observe that the positions of some candidates in the resulting consensus ranking will be different from had we only applied the Kemeny rule to the preference profile. These changes could lower the consensus accuracy of the resulting consensus ranking. Thus, there is a tradeoff between exposure ratio (fairness) and consensus accuracy (profile representation). This tradeoff is unique to each preference profile as it is influenced by the relative agreement amongst voters and the underlying fairness of their rankings.

5 EPIK: EXACT INTEGER PROGRAM SOLUTION

We present our integer program solution for FairExp-kap called epik (Exposure Parity in Kemeny) in Algorithm 1.

Our key insight is that we can translate pairwise information, namely the number of pairs each candidate wins in the consensus ranking, into ordinal rank positions. Then we utilize this expression to formulate the non-pairwise exposure ratio as a constraint on the pairwise Kemeny integer program. We explain in our remarks below.

The first remark shows we can translate the number of pairwise wins candidate c_i has over all other candidates in ranking r, $WP(r, c_i)$, into its ordinal rank position, $r^{pos}(c_i)$.

Remark 1. *We can express $r^{pos}(c_i)$, the position of candidate c_i in ranking r, in terms of the number of pairs that candidate has won in r, $WP(r, c_i)$, as $r^{pos}(c_i) = |r| - WP(r, c_i)$.*

Thus, remark 1 translates rank positions, the building block of exposure, into pairwise wins, the building block of the Kemeny rule. Next, remark 2 directly transforms the exposure of a candidate into its pairwise wins.

Algorithm 1 EPIK - Exposure Parity In Kemeny

Input: Preference profile *R* (*n* voters and *m* candidates), with candidates defined by protected attribute *A*, and parameter γ representing the minimum exposure ratio (Definition 4.1) for the consensus ranking.

Output: Binary variables $C_{a,b} \forall\, c_a, c_b \in C$ inducing consensus ranking *r* with $ER(r{,}A) \geq \gamma$.

$$\text{Maximize: } \sum_{\forall a,b} Q_{a,b} C_{a,b} \tag{4}$$

$$\text{Subject to: } C_{a,b} \in \{0, 1\} \tag{5}$$

$$C_{a,b} + C_{b,a} = 1 \;\forall c_a, c_b \tag{6}$$

$$C_{a,b} + C_{b,d} + C_{d,a} \leq 2 \;\forall c_a, c_b, c_d \tag{7}$$

$$\forall\; G_{A:j}, G_{A:k}$$

$$\frac{\min\{|G_{A:j}|^{-1} \sum_{\forall c_i \in G_{A:j}} \frac{1}{log_2(|r| - \sum_{h=1}^{m} C_{i,h} + 1))}\}}{\max\{|G_{A:k}|^{-1} \sum_{\forall c_l \in G_{A:k}} \frac{1}{log_2(|r| - \sum_{q=1}^{m} C_{l,q} + 1))}\}} \geq \gamma \tag{8}$$

Remark 2. *We can write the exposure of a candidate c_i in ranking r in terms of the number of pairs that candidate has won as $exposure(r, c_i) = 1/log_2(|r| - WP(r, c_i) + 1)$.*

epik exactly solves our FairExp-kap problem. In epik (Algorithm 1), binary decision variables $C_{i,j}$ are utilizied to represent if candidate c_i is placed above candidate c_j in the output consensus ranking. While epik operates in the space of m^2 binary variables, this is fortunately the same computational space as the fairness-unaware Kemeny integer program [12]. In epik, the objective function (Equation 4) maximizes consensus accuracy (i.e, minimizes $d_{KT}(r, R)$) by maximizing pairwise agreements in matrix Q (Equation 3). Subsequent constraints (Equations 5 - 7) ensure that the output is a valid ranking and contains no cycles. These equations find the Kemeny consensus ranking [12]. The last constraint (Equation 8) bounds the exposure ratio (*ER*) of the resulting consensus ranking to be $\geq \gamma$.

To bound the exposure ratio of the consensus ranking, we use Remark 1 and Remark 2. First, we express the exposure of a group $G_{A:j}$ in ranking r, $group\ exposure(r, G_{A:j})$, as:

$$group\ exposure(r, G_{A:j}) = |G_{A:j}|^{-1} \sum_{\forall c_i \in G_{A:j}} exposure(r, c_i) \tag{9}$$

$$= |G_{A:j}|^{-1} \sum_{\forall c_i \in G_{A:j}} \frac{1}{log_2(r(c_i) + 1))} \tag{10}$$

$$= |G_{A:j}|^{-1} \sum_{\forall c_i \in G_{A:j}} \frac{1}{log_2(|r| - \sum_{*=1}^{m} C_{i,*} + 1))}. \tag{11}$$

Where the second equality is from Equation 1 of exposure and the third equality uses remark 2 expressed in terms of the binary variables $C_{i,j}$. We express the exposure ratio for groups defined by

Algorithm 2 EPIRA - Exposure Parity In Rank Aggregation (default is `Copeland` voting rule.)

Input: Profile R (n voters and m candidates), with candidates defined by protected attribute A, and parameter γ representing the minimum exposure ratio (Definition 4.1) for the consensus ranking. Unless specified to be `Kemeny`, `Borda`, `Schulze`, or `Maximin`, `VotingRule` is defaulted to `Copeland`.
Output: Consensus ranking r with $ER(r,A) \geq \gamma$.

```
 1: Let $r_c$ = VotingRule($R$), with default Copeland, $r$ = deep copy
    of $r_c$.
 2: while $ER(r,A) < \gamma$ do
 3:     $Gmin$ = group with lowest $group\ exposure$ (Eq. 2) in $r$.
 4:     $Gmax$ = group with highest $group\ exposure$ (Eq. 2) in $r$.
 5:     $c_{Gmin}$ = highest ranked candidate of $Gmin$, that is still
        below a candidate of $Gmax$, provided the candidate was not
        re-positioned in a prior iteration.
 6:     $UnderE = group\ exposure(r, Gmin)$ -
        $exposure(r, lowest\ (bottom)\ ranked\ candidate\ \in Gmin)$.
 7:     $Boost$ = $|Gmin| *$ AverageExp(r,Gmax) $* \gamma - UnderE$.
 8:     Determine position $p$ in $r$ with exposure closest to $Boost$
        not occupied with a member of $Gmin$.
 9:     Insert $c_{Gmin}$ into $p$ and the item at previously at $p$ to the
        position of the former.
10: end while
11: Preserving assignment of groups to positions, re-assign items
    to positions in $r$ by item order within a group in $r_c$.
12: return $r$
```

attribute A in consensus ranking r as:

$$ER(r,A) = \frac{\min\{group\ exposure(r, G_{A:j})\}}{\max\{group\ exposure(r, G_{A:k})\}} \forall\ G_{A:j}, G_{A,k} \tag{12}$$

$$= \frac{\min\{|G_{A:j}|^{-1} \sum_{\forall c_i \in G_{A:j}} \frac{1}{log_2(|r| - \sum_{*=1}^{m} C_{i,*} + 1))}\}}{\max\{|G_{A:k}|^{-1} \sum_{\forall c_l \in G_{A:k}} \frac{1}{log_2(|r| - \sum_{*=1}^{m} C_{l,*} + 1))}\}}, \forall G_{A,j}, G_{A,k}. \tag{13}$$

The second equality is from Equation 2 of group exposure and Remark 2. Thus, we can formulate a constraint on the exposure ratio of a consensus ranking r in terms of the binary variables $C_{i,j}$. Hence, we can use exposure ratio *ER* as a fairness constraint even though it is not pairwise. Because we integrate fairness of exposure into the Kemeny optimization, EPIK is an exact solution.

6 EPIRA: CANDIDATE SWAPPING ALGORITHM FOR SOLVING FAIREXP-KAP

It bears consideration that optimization can be computationally intensive both for fairness-unaware consensus generation [2, 12] and for fairness of exposure in recommendation ystems [6, 39]. And while Kemeny is extremely popular it is just one approach to consensus generation [7]. Thus we think it is of great significance to address fair rank aggregation (FairExp-kap) for *any* voting rule without requiring access to optimization solvers. Therefore, we design the EPIRA (Exposure Parity in Rank Aggregation) strategy, which is extremely flexible in that it can be used with any voting rule favored in a given application context. EPIRA is a post-processing candidate swapping algorithm that deterministically swaps candidates to ensure, first, that the exposure ratio of the resulting consensus ranking is at least γ. And second, that based on the candidate swaps performed to ensure $ER(r, A) \geq \gamma$, consensus accuracy does not degrade. This is done by ensuring the ordering of candidates within a group is preserved from the fairness-unaware consensus ranking to the fair consensus ranking.

We utilize Copeland as the default voting rule with EPIRA since it a known high-fidelity Kemeny rule approximation [2], and our empirical results (Section 7.5) confirm this design choice. EPIRA finds the Copeland consensus ranking ${r_c}^2$, then while the exposure ratio of r_c is below the desired value of γ it re-positions groups in ranking r, a copy of r_c. The repositioning, lines 3 – 9, where each is linear in m, is done by moving up the highest ranked candidate, c_{Gmin}, in the group with lowest exposure, G_{min}. However, as the highest candidate in G_{min} could be at the top of r (i.e., position 1) we restrict c_{Gmin} to be ranked below at least one member of G_{max}, the group with the highest exposure. Further, during each swap we mark a candidate as "swapped", and we do not swap the same candidate twice to avoid falling into a loop where the same candidate is re-positioned back and forth. The actual swap is done by first determining *Boost*, which is the exposure value needed to satisfy γ provided c_{Gmin} replaces its contribution to G_{min}'s exposure with exposure equal to *Boost*. Then, c_{Gmin} is moved into position p, which is the position with exposure closest to *Boost*. Once $ER(r, A) \geq \gamma$, we use ranking r as a blueprint to re-assign candidates to positions in r by preserving their original **within-a-group** (WiG) order from the Copeland ranking r_c. The WiG order is the order of candidates within the same group, i.e. signifying preference preservation within groups.

A natural question is why we insist on what appears to be an extra final step in EPIRA, instead of returning r when $ER(r, A) \geq \gamma$ is satisfied. The reason is that our proposed final step yields the following guarantee.

Theorem 5.1 *EPIRA (Algorithm 2) has smaller Kendall tau distance to R, and thus higher consensus accuracy (CA), than EPIRA without preserving the within-a-group order (WiG) of the initial Copeland (or alternate voting rule) ranking.*

To develop a foundation for Theorem 5.1, we first establish and then prove Lemma 5.1 below.

Lemma 5.1 *Given the result r_v of a rank aggregation rule, and two rankings r^{WiG} and r^{NoW}, such that each have the same assignment of groups to rank positions, but r^{WiG} preserves the within-a-group (WiG) order of r and r^{NoW} does not, then $d_{KT}(r, r^{NoW}) < d_{KT}(r, r^{WiG})$ cannot hold.*

Proof. Let r be a consensus ranking, and r^{WiG} and r^{NoW} be the same ranking except that in r^{NoW} there are candidates c_i and c_j belonging to the same group, whereby in r and r^{WiG}, $r^{pos}(c_i) < r^{pos}(c_j)$ and $r^{pos}_{WiG}(c_i) < r^{pos}_{WiG}(c_j)$ (i.e., c_i is ranked higher towards the top than c_j in r and r^{WiG}), but $r^{pos}_{NoW}(c_j) < r^{pos}_{NoW}(c_i)$ (i.e., c_j is ranked higher towards the top than c_i in r^{NoW}). Using

[2] We use Copeland in this text as it is the default `VotingRule`. Algorithm 2 input displays the existing voting rules supported in our implementation, other rules can be used as well.

$d_{KT:c_i,c_j}(r, r^{WiG})$ to denote the Kendall tau distance between candidates c_i and c_j in rankings r and r^{WiG}, let $\mathcal{D} = d_{KT}(r, r^{WiG}) - d_{KT:c_i,c_j}(r, r^{WiG})$. Then from Kendall tau, it follows $\mathcal{D}$ also equals $d_{KT}(r, r^{NoW}) - d_{KT:c_i,c_j}(r, r^{NoW})$ since r^n and r^{WiG} are otherwise the exact same ranking. By contradiction, we will show that $d_{KT}(r, r^{NoW}) \geq d_{KT}(r, r^{WiG})$. Assume $d_{KT}(r, r^{NoW}) < d_{KT}(r, r^{WiG})$. Observe that:

$$\begin{aligned} d_{KT}(r, r^{NoW}) &= d_{KT}(r, r^{NoW}) - d_{KT:C_i,C_j}(r, r^{NoW}) \\ &\quad + d_{KT:C_i,C_j}(r, r^{NoW}) \\ &= d_{KT}(r, r^{NoW}) - d_{KT:C_i,C_j}(r, r^{NoW}) + 1 \\ &= \mathcal{D} + 1 \end{aligned}$$

since $d_{KT:C_i,C_j}(r, r^{NoW}) = 1$ by Kendall tau. Then

$$\begin{aligned} d_{KT}(r, r^{WiG}) &= d_{KT}(r, r^{WiG}) - d_{KT:C_i,C_j}(r, r^{WiG}) \\ &\quad + d_{KT:C_i,C_j}(r, r^{WiG}) \\ &= d_{KT}(r, r^{WiG}) - d_{KT:C_i,C_j}(r, r^{WiG}) \\ &= \mathcal{D} \end{aligned}$$

since $d_{KT:C_i,C_j}(r, r^{WiG}) = 0$ by Kendall tau. Further, observe that, $\mathcal{D}$+1 cannot be $< \mathcal{D}$. Thus, we have a contradiction with $d_{KT}(r, r^{NoW}) < d_{KT}(r, r^{WiG})$. Hence, we have shown $d_{KT}(r, r^{NoW})$ cannot be $< d_{KT}(r, r^{WiG})$. □

Thus, EPIRA is guaranteed to be equal or better at representing profile R than EPIRA without the within-a-group property. Thus, for the specific candidate swaps performed to achieve $ER(r, A) \geq \gamma$ EPIRA maximizes consensus accuracy.

7 EXPERIMENTS

In the first two parts of this section we evaluate the proposed EPIK and EPIRA algorithms, studying their behavior compared to the Kemeny rule (Section 7.2), along with providing empirical assessment for their performance compared to using alternate techniques for solving FAIREXP-KAP (Section 7.3). In the final two parts, we evaluate EPIRA in terms of its WiG property (Section 7.4) and its performance with voting rules beyond the Copeland rule (Section 7.5).

7.1 Experimental Setup

7.1.1 Compared Methods. As we newly propose FAIREXP-KAP, no prior work supports finding consensus rankings subject to group fair exposure. We thus compare against the alternate techniques below.

- **KEMENY** [12]: the standard Kemeny rule integer program.
- **PRE-FE:** we study pre-processing by using the re-ranking piece of EPIRA to pre-process each ranking in profile R to be fair prior to aggregation, i.e., to satisfy $ER \geq \gamma$. Then the standard Kemeny rule is applied.
- **PFAIR-KEM** [9]: this fair rank aggregation method applies constraints on cross-group pairs in the Kemeny integer program to ensure each group wins a proportional share of cross-group pairwise comparisons in the consensus ranking (enforced by parameter δ). As done in [9], $\delta = 0.1$.
- **RAPF** [43]: this fair rank aggregation method selects a random ranking in the profile and corrects it to satisfy what the authors call p-fairness, which represents groups proportionally at each position and is inspired by the Chairman assignment problem. It does not employ voter aggregation rules, yet we include it for completeness.

Reference ranking	a	b	c	d	e
Exposure ratio (ER)	0.484	0.573	0.721	0.797	0.844

Table 1: Fairness of reference rankings used in the *Mallows* datasets. Fairness, by exposure ratio (ER), increases "alphabetically".

7.1.2 Datasets. In our experiments, we consider seven datasets.

- ***Mallows***: for our controlled study, we input profiles according to the popular **Mallows** model [24, 30]. The model has a reference ranking of candidates and a spread parameter ϕ, which as it increases the profile contains more consensus (agreement) with the provided reference ranking. For $n = 12$, and $m = 20$, we control the fairness of five profiles based on the exposure ratio (ER) of the reference ranking. Table 1 summarizes the ER values of the five reference rankings (a - e) we use in our Mallows datasets. Fairness increases "alphabetically". We refer to the 30 profiles (5 references rankings with 6 dispersion parameters each) as the *Mallows* dataset.

From Social Choice we use five datasets that contain strictly ordered profiles from the Preflib repository [31].

- ***AGH 2003*** [41]: AGH University of Science and Technology course selection from 2003. Where 146 students ranked all courses. We create groups by splitting courses so $1 - 6$ comprise group A and courses $7 - 9$ group B.
- ***AGH 2004*** [41]: AGH University of Science and Technology course selection from 2004. Where 153 students ranked all courses. We create groups by splitting courses so $1, 3, 4$ comprise group A and courses $2, 5, 6, 7$ group B.
- ***Dublin North*** [31]: 3,662 voters ranking all 12 candidates in a Irish election (2002), with male and female groups.
- ***Dublin West*** [31]: 3,800 voters ranking all 9 candidates in a Irish election (2002), with male and female groups.
- ***Meath*** [31]: 2,490 voters ranking all 9 candidates in a Irish election (2002), with male and female groups.

From the fairness literature [9], we use the following rank aggregation dataset.

- **CSRankings** [5]: consists of rankings of 65 US CS departments from $2010 - 2020$, using relative order from csrankings.org as the yearly ranking. The protected attribute is the combination of the geographic region (Northeast, Midwest, West, and South) and whether the institution is public or private. This forms eight groups.

Our source code and experiment implementation is available at: https://github.com/KCachel/Fairer-Together-Mitigating-Disparate-Exposure-in-Kemeny-Aggregation.

7.2 How do our FairExp-kap solutions compare to the fairness-unaware Kemeny rule solution?

To compare our FairExp-kap solutions with the Kemeny rule solution, we study how both are affected by the underlying fairness (modeled by reference rankings a - e) and consensus (modeled by ϕ) of the profile. Figure 1a compares the consensus rankings found by epik and kemeny across preference profiles in the *Mallows* dataset. Immediately, we observe that kemeny almost always returns very unfair consensus rankings (seen in pale green squares with low *ER* values). The one exception is the profile with very low underlying fairness and consensus (e.g. a and $\phi = 0$). Here, low underlying consensus with an unfair reference means most voters have relatively fair rankings, so when they are combined Kemeny creates a fair consensus ranking. Observing the kemeny exposure ratios we provide two takeaways. First, the conditions in the profile that are likely to yield the most unfair Kemeny consensus rankings are high consensus with an unfair reference ranking (bottom right, e.g., a, b and $\phi = 0.8, 1$) and low consensus with a fair reference ranking (top left, e.g., d, e and $\phi = 0, 0.2, .4$). Second, unsurprisingly, the more consensus in the profile, the greater the consensus accuracy is for kemeny.

Turning to FairExp-kap solutions shown by epik in Figure 1a and epira in Figure 1b our algorithms always achieve the desired exposure ratio of ≥ 0.9, demonstrating they mitigate disparate exposure in the Kemeny rule across diverse preference profile conditions. We see that due to the swapping strategy of epira it has slightly higher exposure ratios than epik (e.g., *ER* is often around 0.95), whereas the optimization approach of epik ensures its exposure ratios hover right at the desired $\gamma = 0.9$. This also results in the consensus accuracy of epira generally being around $0.01 - 0.02$ lower than epik. Overall, we see epira is a very good approximation of epik.

We generally observe that the consensus accuracy of both methods is not significantly lower than that of kemeny. Nonetheless, we eschew statements that our algorithms provide comparable consensus accuracy as kemeny, since as it is clear to see, consensus accuracy is influenced by the preference profile. However, based on our two takeaways from kemeny above, we would expect, and empirically confirm, that consensus ranking produced by epik and epira have higher relative consensus accuracy when there is more consensus in the profile (e.g., $\phi = 0.8, 1$). Also, when the profile has high consensus with low underlying fairness the drop in consensus accuracy from kemeny to our algorithms is greatest (e.g., for a and $\phi = 0.8$ *CA* drops 0.09 between kemeny and epik). Interestingly, this statement is *not* analogous to "when kemeny is most unfair then epik and epira have the lowest consensus accuracy". A counterexample is the profile c and $\phi = 0.6$, which has the same exposure ratio as profile a and $\phi = 1$ In this case, both have *ER* = 0.49. In the former profile, the consensus accuracy of epik is 0.02 less than kemeny, but in the latter the consensus accuracy of epik is 0.09 less than kemeny.

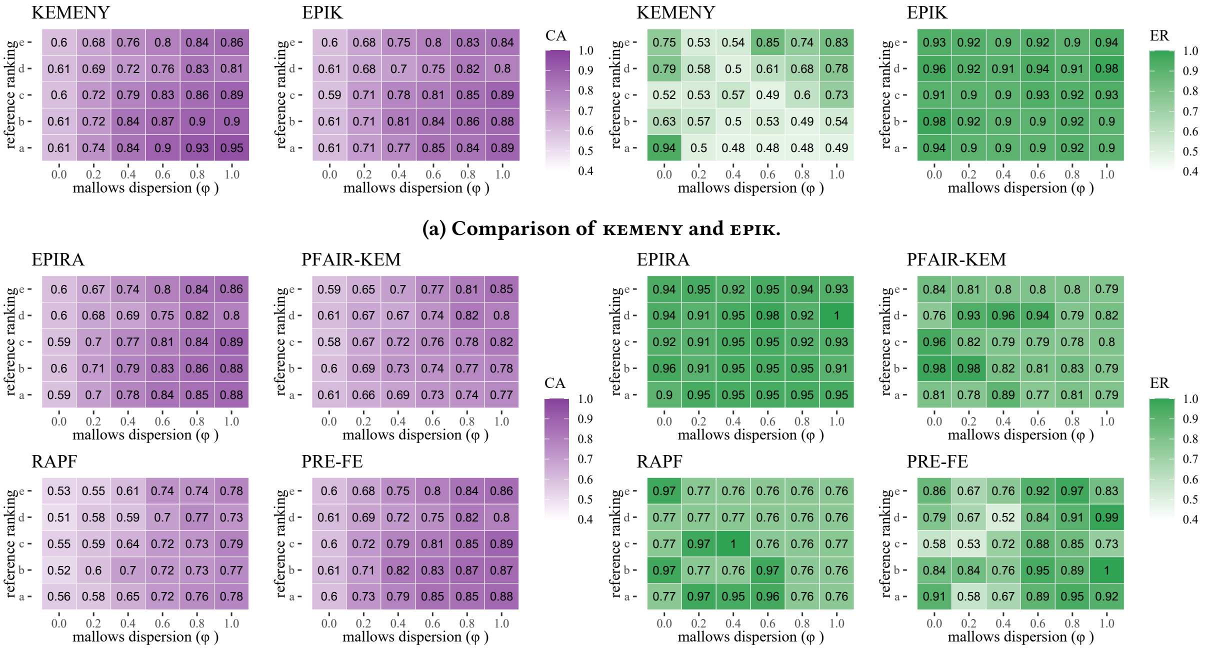
7.3 How do epik and epira compare to alternate techniques for solving the FairExp-kap problem?

We now assess the ability of our proposed algorithms and comparative methods (detailed in Section 7.1.1) to solve FairExp-kap with exposure ratio $ER \geq \gamma$, along with preference representation measured by consensus accuracy (*CA*).

7.3.1 Mallows results. In Figure 1, in addition to the comparison of epik and kemeny (Figure 1a) and epira (Figure 1b) discussed earlier, Figure 1b also shows the performance of the comparative methods for the *Mallows* dataset. We observe several expected behaviors in Figure 1. First, epik and epira always return a consensus ranking with $ER \geq 0.9$ (seen with dark green squares). Second, kemeny, almost always returns very unfair consensus rankings, but naturally has the highest consensus accuracy values (seen in dark purple). Third, as expected due to pfair-kem and rapf having altogether different fairness notions than the FairExp-kap problem, we see that they do not effectively mitigate disparate exposure. For both, we observe a spectrum of outcomes including slight fairness improvement (e.g., e and $\phi = 0.8$), significant improvement (e.g., a and $\phi = 1$), and even the introduction of more disparate exposure (unfairness) than kemeny (e.g., e and $\phi = 1$). And fourth, epira often has higher exposure ratios than epik, since epira performs candidate swaps to meet γ, so it sometimes creates consensus rankings that are fairer than they need to be.

Interestingly, we observe that pre-fe solves FairExp-kap well when the preference profile has a high underlying consensus (e.g., $\phi = 0.8, 1$). In this case, voters are in agreement (regardless of the underlying fairness), thus increasing the exposure ratio to $\geq \gamma$ for each ranking in the profile prior to the Kemeny rule achieves fairness since the rankings were similar to begin with. Nonetheless, pre-fe is not a consistently good solution to FairExp-kap (as seen when $\phi = 0.2, 0.4$). In contrast to pre-fe, pfair-kem has no discernible condition in the preference profile as to when the pfair-kem is expected to mitigate disparate exposure. Further, pfair-kem consistently achieves lower consensus accuracy values than our algorithms, even when it is less fair (e.g. $a, \phi = 0.8$). However, rapf has the lowest consensus accuracy values of all methods. We also find that rapf frequently results in the same consensus ranking regardless of the preference profile (as seen by the frequent .76 *ER* value). Since rapf re-ranks a single voter's ranking in the profile to satisfy a different fairness objective, only using one ranking performs poorly on consensus accuracy and its fairness notion does not mitigate disparate exposure. Thus, across diverse fairness and consensus conditions in preference profiles, only our proposed epik and epira consistently solve FairExp-kap, outperforming all comparative methods.

7.3.2 Real dataset results. Table 2 shows results for all methods for the preflib data sets (Table 2a), and *CS Rankings* data set (Table 2b), respectively. Unsurprisingly, exactly as in the *Mallows* dataset, we again observe the same four expected behaviors as discussed earlier (Section 7.3.1). Further, across all six data sets, epik and epira outperform the comparative methods in achieving fair group exposure (*ER*) and preference representation (*CA*). Alternative fairness techniques in rank aggregation, pfair-kem and rapf, do not



(a) Comparison of KEMENY and EPIK.

(b) Comparison of EPIRA, PFAIR-KEM, RAPF, and PRE-FE.

Figure 1: *Mallows* data set results. Preference representation, *CA*, is in purple (more consensus accuracy is darker) and fairness, *ER*, is in green (fairer is darker). For exposure ratio in EPIK, EPIRA, and PRE-FE we set $\gamma = 0.9$.

mitigate disparate exposure. In fact, as seen in the *Dublin North* data set, they can be more unfair than simply performing KEMENY. Post-processing is sometimes more effective at mitigating disparate exposure than either PFAIR-KEM and RAPF (e.g. in *AGH 2003*, *AGH 2004*, *Dublin North*, and *CS Rankings*), but even so it does not reliably ensure an exposure ratio $\geq \gamma$ while delivering high consensus accuracy.

Interestingly, the fourth expected behavior (EPIRA may sometimes have a higher exposure ratio than EPIK for the same γ parameter) does create a potential tradeoff in deciding which algorithm to use in mitigating Kemeny rule disparate exposure. For this, we observe that EPIK provides the better balance of satisfying exposure ratio at or extremely close to γ along with high consensus accuracy. However, trading a bit of consensus accuracy to use EPIRA does yield the benefit of the flexibility of EPIRA to integrate into existing consensus ranking processes.

7.4 Ablation Study of WiG (Within-a-Group) Property EPIRA Algorithm

Assessing the effectiveness of the Within-A-Group (WiG) property, we compare our EPIRA with WiG and without applying WiG. Figure 2 plots the percentage increase in EPIRA's consensus accuracy (*CA*) across *Mallows* profiles. This empirically confirms, as all squares are $\geq 0\%$, the results of Theorem 5.1, that our proposed EPIRA has a higher consensus accuracy than EPIRA without WiG preservation. Further, when there is a higher agreement (e.g., $\phi = 0.6, 0.8, 1$) in the profile and a reference ranking with a low underlying fairness (e.g a, b, c), the WiG property allows EPIRA to represent substantially more preferences from the profile. Our key takeaway is that WiG for EPIRA achieves a better preference representation.

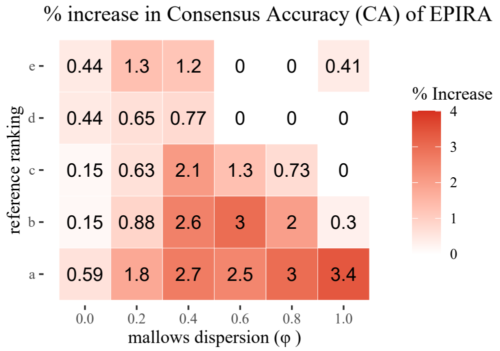


Figure 2: Percent increase in consensus accuracy (*CA*) of EPIRA from the within a group property, WiG.

7.5 Studying Efficacy of EPIRA Algorithm with Alternate Voting Rules

In our previous expiremental results (Sections 7.3.1 - 7.5), we used EPIRA with our default `Copeland` voting rule. Table 3, for the *AGH 2003* and *CS Rankings* datasets, demonstrates that EPIRA can be used effectively with alternate voting rules. See Appendix A for the background descriptions of each voting rule and results from all six

Dataset	Metric	KEMENY	EPIK	EPIRA	PFAIR-KEM	RAPF	PRE-FE
			$\gamma = .95$	$\gamma = .95$	$\delta = .1$		$\gamma = .95$
AGH 2003	consensus accuracy (CA)	0.7536	0.6897	0.6714	0.7456	0.7190	0.7536
	group A avg. exposure	0.4343	0.4796	0.4680	0.4305	0.4329	0.4343
	group B avg. exposure	0.5496	0.4590	0.4821	0.5572	0.5524	0.5496
	exposure ratio (ER)	0.7902	0.9572	**0.9707**	0.7725	0.7836	0.7902
AGH 2004	consensus accuracy (CA)	0.7877	0.6772	0.6545	0.7314	0.7006	0.6545
	group A avg. exposure	0.6121	0.5301	0.5191	0.5550	0.5801	0.5191
	group B avg. exposure	0.3965	0.5059	0.5205	0.4726	0.4392	0.5205
	exposure ratio (ER)	0.6478	0.9545	**0.9972**	0.8515	0.7570	0.9972
Dublin West	consensus accuracy (CA)	0.6482	0.6392	0.6385	0.6482	0.6427	0.6295
	men avg. exposure	0.5130	0.4807	0.4755	0.5130	0.4207	0.3719
	women avg. exposure	0.4224	0.4628	0.4693	0.4224	0.5377	0.5987
	exposure ratio (ER)	0.8233	0.9628	**0.9869**	0.8233	0.7824	0.6212
Dublin North	consensus accuracy (CA)	0.6524	0.6524	0.6524	0.6515	0.5821	0.6493
	men avg. exposure	0.4259	0.4259	0.4259	0.4373	0.3823	0.4162
	women avg. exposure	0.4167	0.4167	0.4167	0.3601	0.6351	0.4653
	exposure ratio (ER)	**0.9782**	**0.9782**	**0.9782**	0.8235	0.6019	0.8944
Meath	consensus accuracy (CA)	0.6423	0.6415	0.6415	0.6423	0.5235	0.6180
	men avg. exposure	0.3931	0.4034	0.4034	0.3931	0.3887	0.3873
	women avg. exposure	0.4468	0.3851	0.3851	0.4468	0.4732	0.4821
	exposure ratio (ER)	0.8799	**0.9546**	**0.9546**	0.8799	0.8215	0.8032

(a) Results for all Preflib data sets with desired exposure ratio being .95 ($\gamma = .95$).

Dataset	Metric	KEMENY	EPIK	EPIRA	PFAIR-KEM	RAPF	PRE-FE
			$\gamma = .8$	$\gamma = .8$	$\delta = .1$		$\gamma = .8$
CS Rankings	consensus accuracy (CA)	0.9274	0.8962	0.8985	0.8560	0.8166	0.9056
	northeast private avg. exposure	0.3285	0.2281	0.2675	0.2901	0.2749	0.2654
	northeast publicavg. exposure	0.2125	0.2238	0.2142	0.2149	0.2135	0.2138
	midwest private avg. exposure	0.1833	0.2170	0.2413	0.2025	0.2005	0.2125
	midwest public avg. exposure	0.2385	0.2457	0.2144	0.2557	0.2567	0.2433
	west private avg. exposure	0.2973	0.2258	0.2231	0.2698	0.2718	0.2188
	west public avg. exposure	0.2246	0.2333	0.2244	0.2309	0.2242	0.2252
	south private avg. exposure	0.1992	0.2561	0.2346	0.2099	0.2090	0.2150
	south public avg. exposure	0.1807	0.2590	0.2584	0.1993	0.2286	0.2589
	exposure ratio (*ER*)	0.5500	**0.8378**	0.8008	0.6871	0.7293	0.8007

(b) Results for *CS Rankings* data sets with desired exposure ratio being four-fiths ($\gamma = 0.8$).

Table 2: Consensus accuracy, exposure ratio (best marked in bold), and average exposure of each group results for all methods.

datasets further confirming the trends in Table 3. Collectively, we observe the following four takeaways.

First, epira always find a consensus ranking with exposure ratio $\geq \gamma$, *regardless of the voting rule used.* Second, epira's consensus accuracy is sensitive to the voting rule. This can be seen with Maximin which generally does not do as well in the preference representation objective as other voting rules. This is expected as Maximin diverges the most from Kemeny. Third, Copeland generally performs the best with high consensus accuracy, indicating it is a great alternative to Kemeny. And, fourth, using Schulze and Borda finds exceptionally similar fair consensus rankings as using Copeland.

8 LIMITATIONS

While we do not foresee clear negative outcomes of this work, but rather positive benefits to disadvantaged groups in consensus processes, we are mindful that we have only begun the study of mitigating disparate exposure in rank aggregation. Thus, our approach has potential limitations we must consider. First, the central fairness concern of our work is the disparate exposure of *ranked candidate* groups, which limits our ability to address maximally representing voter preferences. In social choice, preference representation itself has long been considered a form of fairness with respect to voters [7]. By centering ranked candidate fairness, even though we maximize consensus accuracy, our consensus rankings will inherently have less consensus accuracy (potential voter fairness) than the Kemeny rule solution. And second, we cannot provide

Dataset	Metric	epira $\gamma = .95$				
		`Kemeny`	`Copeland`	`Schulze`	`Borda`	`Maximin`
AGH 2003	consensus accuracy (CA)	0.6714	0.6714	0.6714	0.6688	0.5546
	exposure ratio (ER)	0.9707	0.9707	0.9707	0.9707	0.9855

(a) *AGH 2003* dataset with desired exposure ratio being .95 **(**γ = .95**).**

Dataset	Metric	epira $\gamma = .8$				
		`Kemeny`	`Copeland`	`Schulze`	`Borda`	`Maximin`
CS Rankings	consensus accuracy (CA)	0.9002	0.8985	0.8986	0.8983	0.8725
	exposure ratio (ER)	0.8017	0.8008	0.8032	0.8134	0.8033

(b) *CS Rankings* dataset with desired exposure ratio being four-fiths (γ = 0.8**).**

Table 3: Consensus accuracy, exposure ratio, and average exposure of each group results for epira with alternate `VotingRules`.

an optimal exposure ratio parameter γ, as from a practical standpoint, how much disparate exposure "intervention" is reasonable will be context specific. Ultimately, we leave it to practitioners to control how much fairness intervention is needed and suggest they assess rankings for both ranked candidate fairness and preference representation (exposure ratio and consensus accuracy metrics are helpful in this regard). And lastly, while our methodology supports intersectional protected attributes, we do not inherently study intersectional fairness concerns in rank aggregation. We believe studying intersectional concerns (beyond a simple cross-product of protected attributes) would be exciting grounds for future work.

9 CONCLUSION

We study the new problem of creating a consensus ranking of candidates from a preference profile that is both fair by group exposure and as representative of the profile as possible. Extending the notion of fairness of exposure from information retrieval to social choice, we formally introduce this problem, FairExp-kap, and present an exact integer program epik and an approximate and voting rule agnostics method epira. In our empirical results, we find that utilizing prior work that addresses different fairness notions in rank aggregation, does not achieve the efficacy of our proposed techniques in both the fair exposure and consensus objectives of FairExp-kap.

ACKNOWLEDGMENTS

This work was supported in part by NSF-IIS 2007932.

REFERENCES

[1] Daniel Alabi, Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. 2021. Private Rank Aggregation in Central and Local Models. In *AAAI Conference on Artificial Intelligence*.

[2] Alnur Ali and Marina Meilă. 2012. Experiments with Kemeny ranking: What works when? *Mathematical Social Sciences* 64, 1 (2012), 28–40.

[3] Alon Altman and Moshe Tennenholtz. 2008. Axiomatic foundations for ranking systems. *Journal of Artificial Intelligence Research* 31 (2008), 473–495.

[4] Katherine A Baldiga and Jerry R Green. 2013. Assent-maximizing social choice. *Social Choice and Welfare* 40, 2 (2013), 439–460.

[5] Emery Berger. 2018. CSRankings: Computer Science Rankings.

[6] 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.

[7] Felix Brandt, Vincent Conitzer, and Ulle Endriss. 2012. Computational social choice. *Multiagent systems* 2 (2012), 213–284.

[8] Robert Bredereck, Piotr Faliszewski, Ayumi Igarashi, Martin Lackner, and Piotr Skowron. 2018. Multiwinner elections with diversity constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[9] 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)*. IEEE.

[10] L Elisa Celis, Lingxiao Huang, and Nisheeth K Vishnoi. 2017. Multiwinner voting with fairness constraints. *arXiv preprint arXiv:1710.10057* (2017).

[11] US Equal Employment Opportunity Commission et al. 1979. Questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures. *US Equal Employment Opportunity Commission: Washington, DC, USA* (1979).

[12] 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.

[13] Wade D. Cook, Boaz Golany, Michal Penn, and Tal Raviv. 2007. Creating a consensus ranking of proposals from reviewers' partial ordinal rankings. *Comput. Oper. Res.* 34 (2007), 954–965.

[14] Arthur H Copeland. 1951. *A reasonable social welfare function.* Technical Report. Mimeo, University of Michigan USA.

[15] 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*. 275–284.

[16] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. 2001. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*. 613–622.

[17] Michael D. Ekstrand, Anubrata Das, Robin D. Burke, and Fernando Diaz. 2021. Fairness in Information Access Systems. *Found. Trends Inf. Retr.* 16 (2021), 1–177.

[18] Erica B. Fields, Gül E. Okudan Kremer, and Omar M. Ashour. 2013. Rank aggregation methods comparison: A case for triage prioritization. *Expert Syst. Appl.* 40 (2013), 1305–1311.

[19] Hugo Gilbert, Tom Portoleau, and Olivier Spanjaard. 2020. Beyond Pairwise Comparisons in Social Choice: A Setwise Kemeny Aggregation Problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1982–1989.

[20] Laura A. Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[21] Zhiwei Guan and Edward Cutrell. 2007. An eye tracking study of the effect of target rank on web search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2007).

[22] 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.

[23] Maria Heuss, Fatemeh Sarvi, and M. de Rijke. 2022. Fairness of Exposure in Light of Incomplete Exposure Estimation. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2022).

[24] 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.

[25] Thorsten Joachims, Laura A. Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. *SIGIR Forum* 51 (2005), 4–11.

[26] Anson Kahng, Yasmine Kotturi, Chinmay Kulkarni, David Kurokawa, and Ariel Procaccia. 2018. Ranking wily people who rank each other. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

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

[28] Caitlin Kuhlman and Elke Rundensteiner. 2020. Rank aggregation algorithms for fair consensus. *Proceedings of the VLDB Endowment* 13, 12 (2020).

[29] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. Fare: Diagnostics for fair ranking using pairwise error metrics. In *The World Wide Web Conference*. 2936–2942.

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

[31] Nicholas Mattei and Toby Walsh. 2013. Preflib: A library for preferences http://www. preflib. org. In *International conference on algorithmic decision theory*. Springer, 259–270.

[32] Marina Meilă, Kapil Phadnis, Arthur Patterson, and Jeff A. Bilmes. 2007. Consensus ranking under the exponential model. *ArXiv* abs/1206.5265 (2007).

[33] Farhad Mohsin, Ao Liu, Pin-Yu Chen, Francesca Rossi, and Lirong Xia. 2022. Learning to Design Fair and Private Voting Rules. *J. Artif. Intell. Res.* 75 (2022), 1139–1176.

[34] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling Fairness and Bias in Dynamic Learning-to-Rank. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020).

[35] 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

[36] Fatemeh Sarvi, Maria Heuss, Mohammad Aliannejadi, Sebastian Schelter, and M. de Rijke. 2022. Understanding and Mitigating the Effect of Outliers in Fair Ranking. *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (2022).

[37] Frans Schalekamp and Anke van Zuylen. 2009. Rank Aggregation: Together We're Strong. In *Workshop on Algorithm Engineering and Experimentation*.

[38] Markus Schulze. 2011. A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method. *Social choice and Welfare* 36, 2 (2011), 267–303.

[39] 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*. 2219–2228.

[40] Ashudeep Singh and Thorsten Joachims. 2019. Policy learning for fairness in ranking. *Advances in Neural Information Processing Systems* 32 (2019).

[41] Piotr Skowron, Piotr Faliszewski, and Arkadii Slinko. 2013. Achieving fully proportional representation is easy in practice. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. 399–406.

[42] Nicolas Usunier, Virginie Do, and Elvis Dohmatob. 2022. Fast online ranking with fairness of exposure. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2157–2167.

[43] 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*. 262–275.

[44] Haolun Wu, Bhaskar Mitra, Chen Ma, Fernando Diaz, and Xue Liu. 2022. Joint Multisided Exposure Fairness for Recommendation. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2022).

[45] Himank Yadav, Zhengxiao Du, and Thorsten Joachims. 2019. Fair Learning-to-Rank from Implicit Feedback. *ArXiv* abs/1911.08054 (2019).

[46] Meike Zehlike and Carlos Castillo. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference 2020*. 2849–2855.

[47] Meike Zehlike, Tom Sühr, Ricardo Baeza-Yates, Francesco Bonchi, Carlos Castillo, and Sara Hajian. 2022. Fair Top-k Ranking with multiple protected groups. *Information Processing & Management* 59, 1 (2022), 102707.

[48] William S Zwicker. 2018. Cycles and intractability in a large class of aggregation rules. *Journal of Artificial Intelligence Research* 61 (2018), 407–431.

A DESCRIPTIONS AND EXPLANATION OF VOTING RULES

Below we describe all voting rules used in this work.

- *Kemeny Rule* [27]: selects a ranking with minimal Kendall tau distance to *R*. The *Kendall tau distance* d_{KT} between any two rankings is the number of candidate pairs on which the two rankings disagree. Given preference profile *R*, the Kemeny rule returns a ranking r in $\text{argmin}_{r \in \Pi_C} \sum_{i=1}^{n} d_{KT}(r, r_i)$.
- *Copeland Rule* [14]: creates a ranking by ordering candidates by decreasing Copeland scores. The Copeland score for candidate c_i is $Copeland(c_i) = |\{c_j \in C \mid r^{pos}(c_i) < r^{pos}(c_j)\}| - |\{c_j \in C \mid r^{pos}(c_j) < r^{pos}(c_i)\}|$. In words, it orders candidates by the total number of pairwise contests they win over other candidates in the profile *R*.
- *Schulze Rule* [38]: also known as the path or beatpath method, creates a ranking by ordering candidates by decreasing strength of their beatpaths in profile *R*. Here, a beatpath is an ordered sequence of candidates such that each candidate in the sequence wins a pairwise contest over the next one, and it's strength is the number of candidates in this sequence.
- *Borda Rule* [7] : creates a ranking by ordering candidates by decreasing Borda score. The Borda score for candidate c_i is $Borda(c_i) = n - r^{pos}(c_i)$. Borda results in candidates ordered by the total number of candidates below then in the profile *R*.
- *Maximin Rule* [7] : also known as Simpson and or Minimax, creates a ranking by ordering candidates by decreasing maximin score. Let $N(c_i, c_j)$ be the number of voters for whom $r^{pos}(c_i) < r^{pos}(c_i)$, then the maximin score for candidate c_i is $Maximin(c_i) = min_{\forall c_j \in C} N(c_i, c_j)$. In words, it orders candidates by their worst scores in pairwise contents in the profile *R*.

B EPIRA ALGORITHM WITH ALTERNATE VOTING RULES ON ADDITIONAL DATASETS

In our main paper we compare epira with alternate voting rules on the *AGH 2003* and *CS Rankings* datasets, here we provide additional results for all the remaining datasets. The takeaways in the main paper, Section 7.5, are representative of these results.

C EXPLICIT STUDY OF THE FAIRNESS TUNING PARAMETER γ

While in the main paper, we implicitly show epira and epik utilizing different γ values. Here we provide a dedicated experiment showing that epik and epira always find a consensus ranking with exposure ratio $\geq \gamma$. Figure 3 shows the results of running epik and epira on all the *Mallows* profiles for $\gamma = 0.6, 0.7$, and 0.8. We observe that both methods always have an exposure ratio greater or equal to the γ value, as seen in individual heatmaps. Then moving across heatmaps we see as the γ values increase so do the resulting exposure ratios for the consensus rankings.

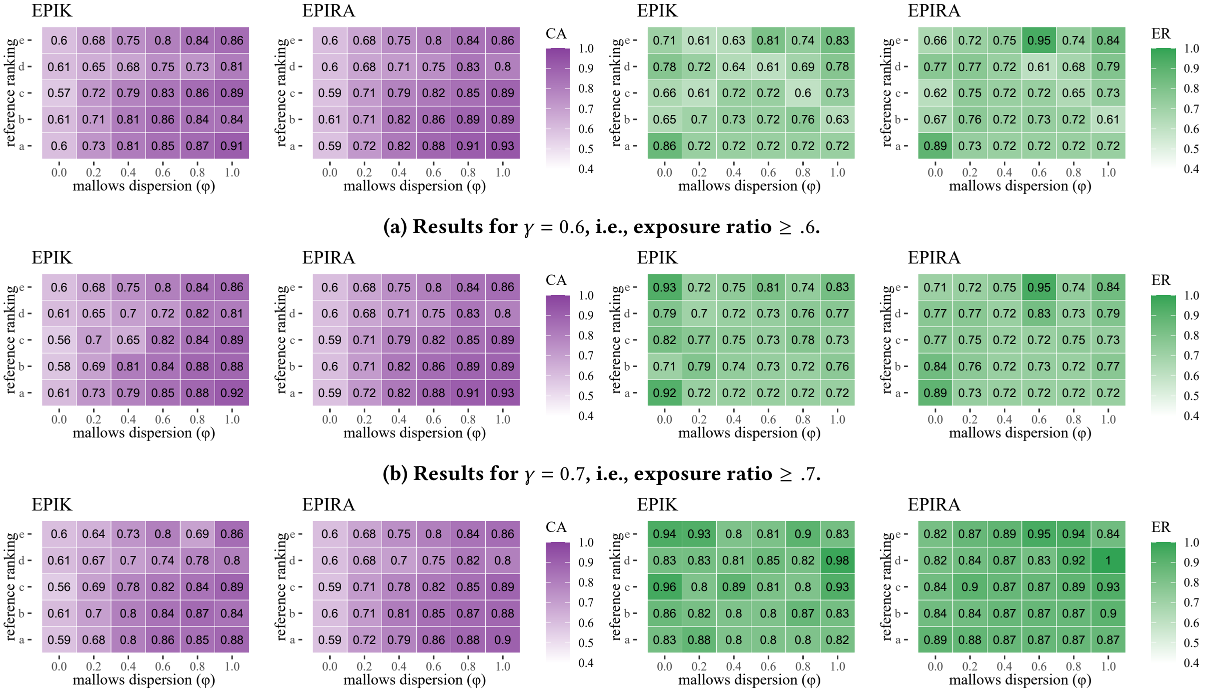
Dataset	Metric	EPIRA $\gamma = .95$				
		`Kemeny`	`Copeland`	`Schulze`	`Borda`	`Maximin`
AGH 2003	consensus accuracy (CA)	0.6714	0.6714	0.6714	0.6688	0.5546
	group A avg. exposure	0.4680	0.4680	0.4680	0.4680	0.4704
	group B avg. exposure	0.4821	0.4821	0.4821	0.4821	0.4773
	exposure ratio (ER)	0.9707	0.9707	0.9707	0.9707	0.9855
AGH 2004	consensus accuracy (CA)	0.6467	0.6545	0.6545	0.6545	0.6371
	group A avg. exposure	0.5191	0.5191	0.5191	0.5191	0.5301
	group B avg. exposure	0.5205	0.5205	0.5205	0.5205	0.5059
	exposure ratio (ER)	0.9972	0.9972	0.9972	0.9972	0.9545
Dublin West	consensus accuracy (CA)	0.6385	0.6385	0.6385	0.6319	0.6385
	male avg. exposure	0.4755	0.4755	0.4755	0.4755	0.4755
	female avg. exposure	0.4693	0.4693	0.4693	0.4693	0.4693
	exposure ratio (ER)	0.9869	0.9869	0.9869	0.9869	0.9869
Dublin North	consensus accuracy (CA)	0.6524	0.6524	0.6524	0.6515	0.6496
	male avg. exposure	0.4259	0.4259	0.4259	0.4259	0.4259
	female avg. exposure	0.4167	0.4167	0.4167	0.4167	0.4167
	exposure ratio (ER)	0.9782	0.9782	0.9782	0.9782	0.9782
Meath	consensus accuracy (CA)	0.6415	0.6415	0.6415	0.6405	0.6356
	male avg. exposure	0.4034	0.4034	0.4034	0.4034	0.4034
	female avg. exposure	0.3851	0.3851	0.3851	0.3851	0.3851
	exposure ratio (ER)	0.9546	0.9546	0.9546	0.9546	0.9546

(a) Preflibdata sets with desired exposure ratio being .95 **(**$\gamma = .95$**).**

Dataset	Metric	EPIRA $\gamma = .8$				
		`Kemeny`	`Copeland`	`Schulze`	`Borda`	`Maximin`
CS Rankings	consensus accuracy (CA)	0.9002	0.8985	0.8986	0.8983	0.8725
	northeast private avg. exposure	0.2679	0.2675	0.2685	0.2645	0.2503
	northeast publicavg. exposure	0.2149	0.2142	0.2160	0.2159	0.2168
	midwest private avg. exposure	0.2409	0.2413	0.2402	0.2226	0.2083
	midwest public avg. exposure	0.2147	0.2144	0.2157	0.2436	0.2444
	west private avg. exposure	0.2236	0.2231	0.2267	0.2324	0.2593
	west public avg. exposure	0.2232	0.2244	0.2214	0.2152	0.2329
	south private avg. exposure	0.2343	0.2346	0.2343	0.2227	0.2086
	south public avg. exposure	0.2589	0.2584	0.2582	0.2586	0.2589
	exposure ratio (ER)	0.8017	0.8008	0.8032	0.8134	0.8033

(b) ***CS Rankings*** **datasets with desired exposure ratio being four-fiths (**$\gamma = 0.8$**).**

Table 4: Consensus accuracy, exposure ratio, and average exposure of each group results for EPIRA with alternate `VotingRules`.



(a) Results for $\gamma = 0.6$, i.e., exposure ratio $\geq .6$.

(b) Results for $\gamma = 0.7$, i.e., exposure ratio $\geq .7$.

(c) Results for $\gamma = 0.8$, i.e., exposure ratio $\geq .8$.

Figure 3: *Mallows* data set results for different γ values in EPIK and EPIRA. Preference representation, *CA*, is in purple (more consensus accuracy is darker) and fairness, *ER*, is in green (fairer is darker).