FARE: Diagnostics for Fair Ranking using Pairwise Error Metrics

Caitlin Kuhlman
cakuhlman@wpi.edu
Worcester Polytechnic Institute

MaryAnn VanValkenburg
mevanvalkenburg@wpi.edu
Worcester Polytechnic Institute

Elke Rundensteiner
rundenst@wpi.edu
Worcester Polytechnic Institute

ABSTRACT

Ranking, used extensively online and as a critical tool for decision making across many domains, may embed unfair bias. Tools to measure and correct for discriminatory bias are required to ensure that ranking models do not perpetuate unfair practices. Recently, a number of error-based criteria have been proposed to assess fairness with regard to the treatment of protected groups (as determined by sensitive data attributes, e.g., race, gender, or age). However this has largely been limited to classification tasks, and error metrics used in these approaches are not applicable for ranking. Therefore, in this work we propose to broaden the scope of fairness assessment to include error-based fairness criteria for rankings. Our approach supports three criteria: *Rank Equality, Rank Calibration*, and *Rank Parity*, which cover a broad spectrum of fairness considerations from proportional group representation to error rate similarity. The underlying error metrics are formulated to be rank-appropriate, using pairwise discordance to measure prediction error in a model-agnostic fashion. Based on this foundation, we then design a fair auditing mechanism which captures group treatment throughout the entire ranking, generating in-depth yet nuanced diagnostics. We demonstrate the efficacy of our error metrics using real-world scenarios, exposing trade-offs among fairness criteria and providing guidance in the selection of fair-ranking algorithms.

CCS CONCEPTS

• **Information systems → Decision support systems**; **Data analytics**; **Content ranking**;

KEYWORDS

Fairness, Fair Ranking, Pairwise Fairness, Fairness Auditing

ACM Reference Format:
Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. FARE: Diagnostics for Fair Ranking using Pairwise Error Metrics. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313443

1 INTRODUCTION

As sophisticated machine learning increasingly impacts our lives on and offline, there is growing concern that discriminatory practices will be baked into automated decision models [3]. The potential for harm is vast, highlighting the need for open and transparent procedures to audit and correct for unfair bias. To address this, recent work on algorithmic fairness assesses the treatment of protected groups by examining the errors made by automated decision making procedures. The main focus of this prior research has been on classification tasks, where predictive models determine a binary outcome [7, 8, 15, 17, 23, 27, 31]. In our work, we broaden the scope of error-based fairness assessment to include rankings.

Motivation. Evaluating ranked data is the de facto process used today for decision making, in particular to make sense of the vast amount of information available online. Rankings simplify the information, helping us to make sense of options and limiting the scope of choices to consider. We rely on ranking models for everyday tasks such as purchasing products to pivotal life decisions such as applying to colleges or jobs. Companies rely on rankings to organize information, whether evaluating candidates to hire [1], or evaluating potential customers using credit scoring [20].

However, rankings can have major pitfalls. Large socio-economic datasets cannot easily be distilled into simple rankings without exploiting patterns that exist in the data, including subtle encodings of historical inequalities. This may lead to unfair decisions, in particular for regulated domains such as employment, education, and lending [3]. For instance, Amazon recently revealed a failed attempt to design a hiring algorithm to screen candidates which they found inadvertently encoded a gender bias against women [10]. Use of such a tool risks automating illegal discrimination. Yet, the development of ranking algorithms for hiring is widespread [1], while no systematic approaches to audit these methods are available to date. Or, consider college rankings published online and used by students and faculty. Highly ranked colleges often have poor outcomes for low-income students, such as lower graduation rates or excessive debt after graduation [29]. The rankings could be considered unfair for only providing utility to high-income students. Left unchecked, such harm compounds through the creation of negative feedback loops. Colleges consistently given a low rank will attract less talent, decreasing their potential to improve [16, 26]. These institutions are entrenched in the position determined by the ranking model - often proprietary and not disclosed.

State-of-the-Art in Fair Ranking. To avoid these kinds of unfair and potentially illegal discriminatory practices, mechanisms to audit ranking models that drive our decision making are essential. Initial methods for measuring unfairness in rankings have recently been put forth [5, 28, 30, 32]. However, they target only a single type of fairness criterion, namely, *statistical parity*. This criterion requires that members of different groups have the same proportional representation among desirable outcomes, i.e., in a top position in the ranking. In our hiring example, such a criterion would dictate that the top candidates in the ranking have a similar proportion of men and women as in the entire applicant pool.

However, this particular criterion may not be appropriate for all applications. Alternatively, a rich variety of fairness criteria available in the literature for classification are *error-based*, meaning

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313443

they require that the predictive model make the same "mistakes" about each group. In our college example, such a criterion might dictate that colleges not be erroneously ranked lower than their counterparts solely due to supporting low income students. We postulate that there is value in considering error-based fairness criteria for rankings, which to date have not yet been applied.

Challenges. For classification, *error-based* definitions of unfairness are measured by within-class error rates, such as True Positive Rate, True Negative Rate or probability of assignment to positive class, for members of different groups [17, 23]. However these metrics require binary class labels, and cannot be applied to rankings learned from training data with non-binary labels. Measuring error made by a ranking model therefore requires metrics appropriate to the task. Ranking models optimize for different types of often application-specific error. For instance, in their fair ranking method, Singh and Joachims [28] define criteria specific to online search, balancing the item exposure in a ranked list of search results with considerations such as document relevance and click-through rate. This narrow definition cannot be directly applied for other ranking applications such as our college ranking example. Instead, a general approach to fair error-based ranking is preferable.

Further, notions of fairness for classification depend on a *preferred outcome* determined by membership in the positive class. Recent methods define an analogous notion based on the prefix of the ranking. While this approach appears to naturally capture the preferred rank outcome, focusing only on the top items in a ranking has severe shortcomings. For many applications, accuracy throughout the entire ranking matters. If the rank position determines funding, bottom-ranked institutions could lose resources and be forced to close. If rankings are used to determine peer groups for tournament-style competitions, inaccuracies low in the ranking would result in mismatched opponents. *In short, accuracy is important at all positions of the ranking.*

Proposed Approach. In this work, we design the first comprehensive auditing methodology for error-based fairness assessment of rankings. In particular, we design analogues for the three core types of fairness criteria in the literature on fair classification [7, 8, 13, 15, 17, 23, 27, 31]. Our proposed criteria, *Rank Equality*, *Rank Calibration*, and *Rank Parity*, together give a comprehensive nuanced assessment of group unfairness in rankings. Our analysis rests on the development of appropriate error metrics for fair ranking. For this, pair inversions, as in foundational rank evaluation techniques [22], are leveraged as basic building block for measuring rank errors. This proposed fairness assessment is model-agnostic and applicable for a wide range of ranking applications.

Further, we propose an auditing mechanism to apply our proposed criteria to measure fairness throughout an entire ranking instead of only in the top results. Our approach, called FARE (for Fair Auditing based on Rank Error), uses a sliding window technique to measure ranges of ranking error and generates error sequences which capture each group's treatment throughout the entire ranking. Custom FARE diagnostics provide a nuanced summary of results, while remaining interpretable to stakeholders who rely on rankings for decision making.

We evaluate our fairness metrics in a comprehensive experimental analysis. Using our FARE framework, we evaluate rankings created using state-of-the-art fair ranking methods [32] on real-world datasets [2, 19]. Our analysis identifies strengths and weaknesses of these existing techniques with regard to fairness. We demonstrate how our analysis guides the choice of appropriate fair ranking correction method to apply.

Key contributions of our work include:

(1) We define a new set of fairness criteria customized for rankings: *Rank Equality, Rank Calibration, and Rank Parity* based on novel rank-specific error metrics. These criteria are model-agnostic and applicable for many ranking applications, capturing key notions of fairness previously applied only for classification.
(2) We present FARE, the first comprehensive framework to audit ranking models using error-based fairness criteria. FARE offers a suite of fairness diagnostics for ease of interpretation.
(3) We demonstrate the application of our FARE framework to assess the capability of state-of-art rank correction methods to achieve fairness, and illustrate the power of FARE's diverse criteria to account for unfair bias in rankings across domains.

2 RELATED WORK

Measuring Fairness in Classification. No single rule has been shown to definitively determine the fairness of an algorithm. The majority of criteria for evaluating fairness focus on the treatment of *groups* determined by some protected data attribute. *Error-based group fairness criterion* [7, 17, 23, 27] are concerned with whether the predictions of a classifier are fair with respect to the group membership of items in the training dataset. To verify this, fairness assessment checks whether groups are treated 'similarly,' meaning error rates for each group are within some threshold. It has been observed that all criteria cannot necessarily be satisfied at the same time [7, 23]. Therefore the appropriate choice of fairness metric is context-dependent. Group fairness criteria can be categorized as:

- **Equalized Odds**, coined by Hardt et al. [17], seeks to ensure that the probability of an object being assigned a particular label by the classifier is independent of its group membership, conditional on the true class label [7, 17, 23]. To verify this, Equalized Odds stipulates that the *false positive and true positive error rates* must be similar across all groups.
- **Calibration**, a group-wise measure of fairness for probabilistic classifiers, requires that the calibration error for each group is similar [7, 23, 27]. Calibration error indicates the difference between the true likelihood of membership in the positive class and the probability given by the classifier. For example, if there are 100 objects assigned a 90% chance of being in the positive class by the classifier, approximately 90 of these objects should actually be positive.
- **Statistical Parity** requires proportional representation of each group [8, 15, 31]. This is similar in spirit to the 80% rule for evaluating disparate impact in judicial rulings [15] and affirmative action quotas employed by many institutions. Typically the desired outcome is that some minimal proportion of minority group members are predicted as positive.

Measuring Fairness in Rankings. Recent work has begun to address group fairness concerns in rankings [5, 28, 30, 32]. Most works [5, 30, 32] focus on enforcing only *Statistical Parity* criterion, and detect unfairness only in the top-k prefix of a ranking. Criteria

based on prediction error have not yet been applied for fair ranking to our best knowledge.

Yang and Stoyanovich [30] measure fairness according to a discounted cumulative scoring metric that evaluates the proportional representation of groups. In [5] fairness is defined according to a threshold on the maximum proportion of the majority group allowed in the prefix. Zehlike et al. [32] extend these statistical parity approaches, designing a greedy algorithm to ensure the ranking meets the criterion while optimizing for utility. Singh and Joachims target fair ranking in information retrieval specifically [28]. Metrics favoring accuracy at the top of the ranking are well-suited here since, out of possibly thousands of documents returned for a query, only a few top results are likely to be clicked. While they broaden their fairness definitions beyond statistical parity, metrics proposed in [28] are application-specific, defining fairness in terms of item exposure to users, clickthrough rates, and document relevance.

3 FAIR RANKING PROBLEM FORMULATION

Ranking can have different meanings in different contexts, and ranking models can be trained over various types of ground truth information. Rank predictions can be learned from training data with binary labels (e.g., in bipartite ranking [9]) or discrete labels with ordered classes (i.e., ordinal regression [18] with labels such as "best", "neutral", "worst"). Traditional regression ranks according to continuous scoring functions. Learning-to-rank approaches also include pairwise and listwise models [25].

Therefore, to be widely applicable, we target general rankings with a model-agnostic approach. We assume only that an ordering is imposed over a set of objects X according to some function which assigns each $x_i \in X$ a position relative to all others. Following previous work on error-based criteria for fair classification [7, 8, 13, 15, 17, 23, 27, 31], we formulate the task of auditing the fairness of a ranking as a supervised learning problem in that the true ranking over X is known. [1] Unique to the context of fairness analysis, objects have an associated *protected attribute*, such as gender, age, or race, which partitions the objects into a set of two or more disjoint *groups* $A = \{A_1, \ldots, A_m \mid \cup_{i=1}^{m} A_i = X, A_i \cap A_j = \emptyset, \forall i \neq j\}$. We consider only a single protected attribute, leaving intersectional fairness considerations with multiple protected attributes to future work.

Let $\rho = [x_1 \succ x_2 \succ \ldots \succ x_n]$ be the true ranking over all objects $x_i \in X$, where $\succ$ is an ordering relation on X such that $x_i \succ x_j$ implies that x_i appears at a more preferred position in the ranking than x_j. The number of items in the set is denoted by $|X| = n$. The position of a single object x_i in the true ranking is denoted $\rho(x_i)$. Our task is to evaluate a learned ranking $\hat{\rho}$ according to a given *Fairness Criterion* which relies on a group error function L.

Definition 1. *Given a group error metric $L_{A_i}(\rho, \hat{\rho})$, a* ***Fairness Criterion*** *(FC) is an evaluation rule which designates a ranking $\hat{\rho}$ as fair in relation to a true ranking ρ if:*

$$L_{A_i}(\rho, \hat{\rho}) \cong L_{A_j}(\rho, \hat{\rho}) \,, \; \forall A_i, A_j \in A, \; i \neq j$$

Fairness is evaluated by checking whether the error for each group is similar, or within some threshold, indicated by the symbol $\cong$. The larger the difference in the errors for each group, the more unfair the ranking is considered to be. Our assessment therefore hinges on the choice of a rank-appropriate group error function L.

[1] While a "truly fair" ground truth ranking may not be available, this assumption allows us to distinguish whether a proposed ranking is more or less fair than an alternative.

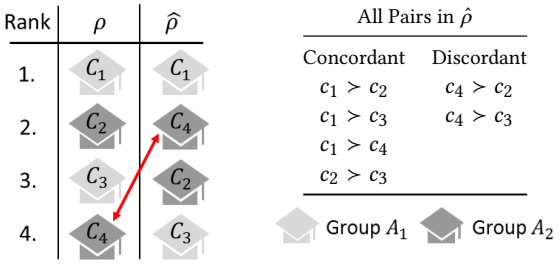

Figure 1: On the left is a true ranking of colleges ρ and predicted ranking $\hat{\rho}$ over two groups of colleges. The resulting discordant and concordant pairs are shown on the right.

4 PROPOSED ERROR METRICS FOR RANK FC

4.1 Pairwise Comparison of Group Errors

To design a general approach for evaluating group error in rankings, we consider foundational approaches [11]. One classic method is to sum the absolute difference in rank position between the true and predicted rankings for each object in the dataset (i.e., to use the Spearman footrule distance). Another popular methodology uses the pairwise error, or Kendall Tau distance [22], by counting the number of inverted pairs of objects in the predicted ranking compared to the true ranking. These two classic approaches to measuring rank similarity have been shown to be *equivalent*, meaning the Kendall Tau is always within a constant factor of the Spearman footrule distance [12, 24]. Given this insight, the two metrics have been used interchangeably for tasks such as rank aggregation [14, 24]. For fairness assessment, the same reasoning applies in that either rank error metric could be applied. However, since we are concerned with the comparative ranking outcomes for different groups, the pairwise approach provides a natural formulation.

Given the position in the true ranking of a pair of objects where $\rho(x_i) > \rho(x_j)$, they are said to be discordant if $\hat{\rho}(x_i) \prec \hat{\rho}(x_j)$. Figure 1 shows the sets of concordant and discordant pairs resulting from an error between two rankings of colleges. We observe that a ranking containing objects from two different groups, A_i and A_j, can be divided into three subsets of pairs: those containing only objects from group A_i, those containing only objects from A_j, and the set of "mixed" pairs containing one object from each group. The total number of unordered pairs in a ranking over X is $\phi(X) = |X|(|X|-1)/2$. The cardinality of the set of mixed pairs is $\phi(X) - \phi(A_i) - \phi(A_j)$. We denote the number of discordant pairs in a set X as $\phi^D(X)$ and concordant pairs as $\phi^C(X)$. The number of mixed pairs favoring objects from one group A_i over another group A_j is denoted by $\phi_{i>j}(X)$. For instance, the cardinality of the set of discordant pairs favoring A_i over A_j is indicated as $\phi^D_{i>j}(X)$.

For simplicity, we henceforth consider two groups. However, rank error based on discordant pairs can easily be extended to multiple groups. For instance, to compute the number of discordant pairs favoring group A_1 given m groups, we compute $\sum_{i=2}^{m} \phi^D_{1>i}(X)$.

4.2 Proposed Rank Equality Criterion

The Equalized Odds criterion for classification measures fairness in terms of the rate at which groups are falsely assigned to the preferred or non-preferred classes. When evaluating a ranking, there are no binary assignments by which to gauge preference. However, *position* in a ranking does indicate a preferred or undesirable outcome - the top of the ranking being analogous to the positive class. When an object is overestimated by the model it is incorrectly assigned a more preferred position than in the true ranking. This is similar in effect to a false positive error made by a classifier. Accordingly, underestimating the position of an object in the ranking incorrectly penalizes it, as in a false negative. Following this principle, we compute the Rank Equality error for group A_i in terms of the number of discordant pairs which favor A_i over items from another group A_j. This proposed metric (Definition 2) captures the number of times that an object from group A_i is incorrectly *overestimated* compared to objects in A_j. The error is then normalized by the total number of mixed pairs ensuring that the error falls in a range of [0, 1]. Normalization creates an interpretable measure of preference and accounts for any imbalance in the size of the groups. To apply the Rank Equality FC, we simply compare the *Req* error for each group.

Definition 2. ***Rank Equality Error***

$$Req_{A_i}(\rho, \hat{\rho}) = \frac{\phi^D_{i>j}(X)}{\phi(X) - \phi(A_i) - \phi(A_j)}$$

where $\phi^D_{i>j}(X)$ denotes the number of discordant pairs which favor the target group A_i over A_j, $i \neq j$.

Rank Equality dictates that no group should be unfairly privileged or penalized compared to another group. As an example, consider the rankings shown in Figure 1. To compute the Rank Equality error for group A_2, we count the number of discordant pairs where an item from A_2 is favored over an item from A_1. Four pairs contain an object from each group: $(c_1, c_2), (c_1, c_4), (c_2, c_3), (c_3, c_4)$. One of these pairs (c_3, c_4) is discordant, since $\hat{\rho}(c_4) \succ \hat{\rho}(c_3)$ and $\rho(c_3) \succ \rho(c_4)$, and favors A_2. Thus $Req_{A_2}(\rho, \hat{\rho}) = \frac{1}{4}$.

4.3 Proposed Rank Calibration Criterion

Calibration is used to evaluate probabilistic classifiers in terms of the confidence of the model, using the mean squared error between predicted likelihood of assignment in the positive class and an estimated "true" probability [4]. Applied as an FC, this criterion checks how well the classifier predicts objects in each group. To evaluate the calibration of a ranking $\hat{\rho}$ for a group A_i, we propose to measure error in predicted rank position by counting the number of discordant pairs which contain at least one member of A_i, as given in Definition 3. This captures the overall error made for items in the group. The value is normalized by the total number of pairs containing objects from A_i. For example, the pairs containing objects from group A_2 in Figure 1 are $(c_1, c_2), (c_1, c_4), (c_2, c_3), (c_2, c_4), (c_3, c_4)$. Pairs (c_2, c_4) and (c_3, c_4) are both discordant, therefore following Definition 3, the rank calibration error is $Rcal_{A_2}(\rho, \hat{\rho}) = \frac{2}{5}$.

Definition 3. ***Rank Calibration Error***

$$Rcal_{A_i}(\rho, \hat{\rho}) = \frac{\phi^D_i(X)}{\phi(X) - \phi(A_j)}$$

Where $\phi^D_i(X)$ denotes the number of discordant pairs containing at least one object from the target group A_i.

4.4 Proposed Rank Parity Criterion

Finally, we also apply pair inversion to design a statistical parity metric like those explored in previous work on fair ranking [5, 30, 32]. Here, the goal is to ensure fair representation of members of each group among objects given a favorable rank position. We propose to capture this idea by counting the pairs in which one group is favored over the other in the learned ranking, regardless of their positions in the true ranking. We again normalize by the total number of mixed pairs in the learned ranking.

Definition 4. ***Rank Parity Error***

$$Rpar_{A_i}(\rho, \hat{\rho}) = \frac{\phi_{i>j}(X)}{\phi(X) - \phi(A_i) - \phi(A_j)}$$

Where $\phi_{i>j}(X)$ is the number of pairs of objects which favor a member of group A_i over a member of A_j, $i \neq j$.

In Figure 1, two pairs in $\hat{\rho}$ favor group A_2 over group A_1: (c_2, c_3) and (c_4, c_3). Therefore, $Rpar_{A_2}(\rho, \hat{\rho}) = \frac{2}{4}$. This matches our intuition of parity, since the groups are still somewhat evenly distributed through the ranking $\hat{\rho}$ in spite of the incorrect placement of c_4.

4.5 Discussion: Ranking Criteria and Their Interrelationships

We now analyze our metrics to understand their interrelationships and scope of applicability. Given a ranking $\hat{\rho}$ over g groups, there are g^2 ways of choosing two objects from the ranking, allowing for group repetition. These pairs may be either *concordant* or *discordant*, resulting in $2g^2$ types of pairs. Table 1 shows the categories of pairs that can be formed for $g = 2$ groups. The colors in the table correspond to the colors in the Venn diagram in Figure 2, illustrating the relationship between the types of pairs used to compute errors for a single group, A_1.

Since our fairness analysis is concerned with the relative error made for each group, within-group concordant pairs are not considered when computing error metrics. All other types of pairs are included in the definition of at least one error metric. Discordant mixed pairs are used to compute all three error metrics. These pairs of objects intuitively capture the *major disparity between groups*: cases where one group is erroneously favored over the other. We define this as Rank Equality. Rank Calibration instead measures the *total error* for each group. This metric counts all pairs containing objects from a single group, capturing within-group as well as across-group errors. Finally, Rank Parity considers the *total advantage* of one group over the other.

FC are compatible in extreme cases. In a perfect prediction there is no error between $\hat{\rho}$ and ρ. In this case $Req = Rcal = 0$ for all groups, since no pairs are discordant. The corresponding FC deem $\hat{\rho}$ fair since the group errors are identical. The *Rpar* error in this case will simply measure the relative advantage of each group in the true ranking. It may be considered fair or unfair depending on the

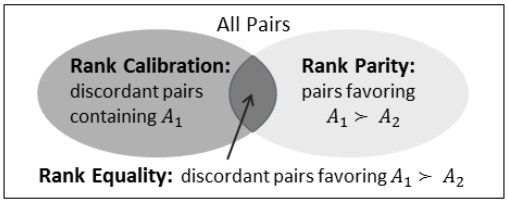

Figure 2: Relationship between the types of pairs used to compute the error for group A_1 (corresponding to Table 1).

Discordant	Concordant
$A_1 \succ {A_1}^D$	$A_1 \succ {A_1}^C$
$A_2 \succ {A_2}^D$	$A_2 \succ {A_2}^C$
$A_1 \succ {A_2}^D$	$A_1 \succ {A_2}^C$
$A_2 \succ {A_1}^D$	$A_2 \succ {A_1}^C$

Table 1: Categorization of pairs in $\hat{\rho}$ over two groups A_1, A_2.

distribution of the objects in each group. Since this is independent of the other metrics, it is therefore possible for a perfect prediction to satisfy all three criteria. In the worst case, $\hat{\rho}$ ranks the objects in the reverse order from ρ. In this case $Rcal = 1$ since all pairs are discordant. Each group is predicted with the same amount of error, therefore by the Rank Calibration FC, $\hat{\rho}$ is considered fair. In this case no pairs are concordant, therefore $Req = Rpar$. Whether the ranking is considered fair according to the corresponding FC again depends on the distribution of groups throughout the ranking.

5 FAIR AUDITING BASED ON RANK ERROR

The FARE Methodology. We next design a non-parametric approach to assess rankings using our proposed error metrics. We want to understand the entire ranking of items from the preferred positions at top of the ranking to the lowest ranked objects. Aggregating the entire treatment of each group using a single fairness score provides only a coarse assessment. Therefore our FARE methodology (for Fair Auditing based on Rank Error) generates *sequences of within-range errors* for each group. The differences in these sequences tell a richer story than would a single value for each group, revealing disparity throughout the entire ranking.

To start, FARE sorts the data according to the predicted ordering $\hat{\rho}$ and bins it into k subsets $\langle B_1, B_2, \ldots B_k \rangle$. Error metrics are then applied to the objects in each bin. In the case of two groups A_1 and A_2 we evaluate errors $l_{1i} = L_{A_1}(\beta_i, \hat{\beta}_i)$ and $l_{2i} = L_{A_2}(\beta_i, \hat{\beta}_i)$ for the data in each bin B_i. This produces two error sequences: $S_1 = \langle l_{11}, l_{12}, \ldots l_{1k} \rangle$ and $S_2 = \langle l_{21}, l_{22}, \ldots l_{2k} \rangle$. An equi-width binning strategy compares the top-k ranked items across both groups in the first bin, the next k in the next bin, and so on. An equi-depth strategy is also possible, where each bin measures how well the ranking predicts $\frac{|A_i|}{k}\%$ of items from each group.

If the number of bins k is so large that there are only a few objects in each, then the sequence of error measurements may exhibit a high degree of variance. This could exaggerate differences between groups in the case where one is a minority. On the other hand, if bins contain many objects, the result is a rather coarse estimate. To capture the error at a sufficient number of positions throughout the ranking while achieving a reasonable bin size, we adopt a *sliding window approach*. This introduces a smoothing transformation over the data to account for high variance across bins. Each consecutive bin of size w overlaps the previous, offset by a fixed step size $s < w$. The first bin B_1 contains objects $\{x_1, x_2, .., x_w\}$, ordered according to their predicted positions $\hat{\rho}(x_i)$, the second bin $B_2 = \{x_{s+1}, x_{s+2}, \ldots x_{s+w}\}$, and so on. Each error sequence S_i contains $\lfloor \frac{|A_i|}{s} \rfloor$ bins.

Diagnostics for Analyzing Fairness. The next step in the auditing procedure is to compare the error sequences S_1 and S_2 produced by our FARE framework to see if they are similar, and therefore meet the fairness criterion, or if they differ in ways which indicate an unfair ranking. To facilitate this, FARE offers *audit plots*. Similar to reliability diagrams for assessing the calibration of classifiers [4], these visual depictions reveal differences in the shapes, patterns and values of the error sequences. Since our proposed error metrics are normalized, the y-axis on each plot has a fixed range of $[0, 1]$, providing an easily interpretable snapshot view of the error sequences generated during the audit process.

Audit plots are augmented by compact statistics, or *fairness scores*, indicating whether the ranking model satisfies the FC. Conceptually, any diagnostic metrics comparing the sequences can be plugged into the framework. We employ a distance diagnostic $dist(S_1, S_2) = \frac{1}{k}\sum_{i=1}^{k} |l_{1i} - l_{2i}|$ to summarize the similarity of the error sequences as a single value. These scores can be thresholded to flag unfair cases where the average magnitude of error for one group is much larger than the other or to apply strict FC cutoffs.

Complexity. A simple pair counting algorithm can be used to compute each of our proposed error metrics in $O(nlog(n))$ time using an adaptation of the mergesort algorithm. Performing an audit using the FARE methodology can therefore also be done in logarithmic time, requiring $O(n/s(wlog(w))$ time for step size s and window size w to compute the error sequences for each group. In cases where performance is an issue, we can improve the pair counting procedure to run in $O(n\sqrt{log(n)})$ time [6].

6 EXPERIMENTAL EVALUATION: AUDITING RANK CORRECTION METHODS

We demonstrate the power of our proposed error metrics using FARE to audit post-processing techniques from the literature designed to correct existing rankings. Two state-of-the-art fair ranking methods which enforce statistical parity notions of fairness are compared. Using the FARE methodology, we reveal the resulting tradeoffs between this criterion and the prediction error introduced. We evaluate these fair ranking algorithms using FARE audit plots (Figure 3) and FARE distance diagnostics (Table 2).

State-of-Art Fair Correction Methods. We reproduce a subset of experiments presented by Zehlike et al. [32] using implementation and data made available by the authors."FA*IR" rankings are generated using a greedy algorithm proposed in [32] to create a fair top-k prefix ranking. The rankings target a user-specified minimum proportion of the minority group, subject to a statistical significance test. The proportion is indicated in the method name, e.g., FA*IR2 for 20%. Here we use the same proportions as the authors, chosen to be close to the actual group ratio over the entire dataset.

The "Feldman" method [15] is proposed as pre-processing step for fair classification in which data are ranked. In this method the utility scores for objects in the minority group are adjusted to match the distribution of the majority. We compare these rankings against a baseline of the true ranking with no correction.

Datasets. We evaluate the rank correction methods using two datasets. The Statlog German Credit Dataset [19] provides a "ground truth" ranking of people according to their credit-worthiness for our experiments. Three "fair" rankings are then created using *age* < 25, *age* < 35 and *gender* = female as protected group attributes. Prefix rankings with $k = 100$ are generated. Audit parameters for this dataset are $w = 30, s = 10$. The COMPAS recidivism dataset published by ProPublica in their investigation of racial bias in the criminal justice system is also utilized [2]. This dataset is ranked according to COMPAS scores indicating the likelihood of re-offending for the "true" ranking with $k = 1000$. "Fair" rankings are generated according to groups *race* = African American and *gender* = male. Audit parameters are $w = 100, s = 10$.

Metrics. We produce audit plots using our proposed metrics *Req*, *Rcal* and *Rpar*, and summarize the results using FARE distance diagnostics. In their experiments, Zehlike et al. [32] use a number of metrics to gauge the tradeoff between parity and prediction accuracy. We include two metrics for comparison: *NDCG*: normalized discounted cumulative gain [21] (commonly used in search), and *rank drop*: the maximum number of positions lost by an object.

Discussion. Table 2 summarizes the FARE diagnostics for our experiments. The rankings deemed most fair in this audit are highlighted in bold. Asterisks mark the conclusions which align with the analysis in [32]. For three out of five rankings FA*IR clearly outperforms Feldman, satisfying multiple fairness concerns.

The impact of both the FA*IR and Feldman rank correction techniques on statistical parity concerns is apparent, as measured by our Rank Parity FC. For instance, for the German Credit dataset using *age* < 25, *Rpar* distance is 0.25 in the baseline "true" ranking. Both methods are able to reduce this to 0.03. The degree to which error is introduced as a result of the correction algorithm is reflected in the *Req* and *Rpar* scores. By comparison, the NDCG metric is not sensitive to the rank correction methods, and therefore not expressive enough to capture unfairness. The rank drop values tend to align with the FARE diagnostics. However, this value is not very interpretable. We cannot observe which group had the farthest drop, or whether the position of many items dropped.

For such nuanced analysis we turn to our FARE audit plots, shown in Figure 3. For the German Credit dataset, we can visually discern that both FA*IR and Feldman systematically introduce *Req* and *Rcal* errors that indicate an unfair disparity in the treatment of the groups throughout the ranking. For the COMPAS dataset, we observe similar patterns and magnitude of error throughout the rankings for both groups. In this case, the correction methods are introducing error in a fair manner. Feldman shows a drop midway through the *Rcal* sequence while the FA*IR error appears more consistent. FARE compliments the use of these ranking methods by providing this in-depth view of the treatment of each group.

In the interest of reproducibility, all code, data, and a full analysis including additional diagnostics and datasets is available online:
Examples and analysis: https://github.com/caitlinkuhlman/fare
FARE python package: https://pypi.org/project/fare

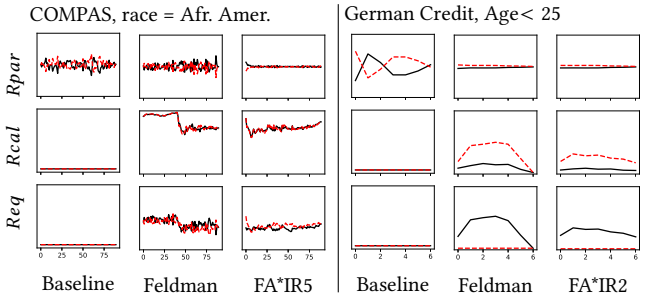

Figure 3: Audit plots for rank correction methods illustrate how errors (plotted on the y-axis and normalized between 0 and 1), manifest throughout the ranking. The x-axis represents the sliding window moving from highly ranked items on the left to the lowest on the right. Errors for group A_1 are shown as a solid black line, group A_2 as dashed red line.

			FARE				
Dataset	**Group**	**Method**	*Rpar*	*Rcal*	*Req*	**NDCG**	**Drop**
German Credit k=100	age< 25	Baseline	0.25	0.00	0.00	1.00	0
		Feldman	**0.03**	0.23	0.33	1.00	8
		FA*IR2	**0.03***	**0.18**	**0.27**	1.00	7
German Credit k=100	age< 35	Baseline	0.17	0.00	0.00	1.00	0
		Feldman	**0.04**	0.06	**0.26**	0.99	36
		FA*IR6	**0.04**	**0.04**	0.40	0.99	30
German Credit k=100	gen=f	Baseline	0.33	0.00	0.00	1.00	0
		Feldman	**0.05**	**0.10**	**0.27**	1.00	8
		FA*IR7	0.08	0.11	0.28	1.00	0
COMPAS k=1000	race	Baseline	0.09	0.00	0.00	1.00	0
		Feldman	0.08	**0.01**	0.08	0.98	393
		FA*IR5	**0.02***	**0.01**	**0.04**	0.99	319
COMPAS k=1000	gen=m	Baseline	0.13	0.00	0.00	1.00	0
		Feldman	0.09	0.03	0.09	1.00	294
		FA*IR8	**0.02***	**0.01**	**0.03**	1.00	161

Table 2: Fairness evaluation for rank correction methods. FARE distance diagnostics are shown in the center, and compared to standard error metrics.

7 CONCLUSION

In this work we present the first methodology for auditing rankings using group error metrics which capture popular notions of fairness. Our proposed fairness criteria together with our FARE auditing method comprise a powerful diagnostic tool for nuanced analysis of the treatment of groups being ranked. FARE can be applied in a general and model-agnostic fashion, for many applications where rankings are used to simplify complex socio-economic datasets, providing a crucial service of debunking unfair rankings.

ACKNOWLEDGMENTS

The authors thank the Computing Resources Association for Women for support through the CREU program. This work was also partially funded by NSF IIS-1815866, IIS-1560229, IIS-1815866 and US Department of Education GAANN Fellowship P200A150306.

REFERENCES

[1] Ifeoma Ajunwa, Sorelle Friedler, Carlos E Scheidegger, and Suresh Venkatasubramanian. 2016. Hiring by algorithm: predicting and preventing disparate impact. (2016).

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *Pro Publica* (2016).

[3] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Cal. L. Rev.* 104 (2016), 671.

[4] Rich Caruana and Alexandru Niculescu-Mizil. 2004. Data mining in metric space: an empirical analysis of supervised learning performance criteria. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 69–78.

[5] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2017. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840* (2017).

[6] Timothy M Chan and Mihai Pătraşcu. 2010. Counting inversions, offline orthogonal range counting, and related problems. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 161–173.

[7] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[8] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 797–806.

[9] Corinna Cortes and Mehryar Mohri. 2004. AUC optimization vs. error rate minimization. In *Advances in neural information processing systems*. 313–320.

[10] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (2018). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[11] Persi Diaconis. 1988. Group representations in probability and statistics. *Lecture Notes-Monograph Series* 11 (1988), i–192.

[12] Persi Diaconis and Ronald L Graham. 1977. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)* (1977), 262–268.

[13] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Mark DM Leiserson. 2018. Decoupled Classifiers for Group-Fair and Efficient Machine Learning. In *Conference on Fairness, Accountability and Transparency*. 119–133.

[14] 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*. ACM, 613–622.

[15] 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*. ACM, 259–268.

[16] Malcolm Gladwell. 2011. The order of things. *The New Yorker* 87, 1 (2011), 68–75.

[17] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*. 3315–3323.

[18] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. 1999. Support vector learning for ordinal regression. (1999).

[19] PDH Hofmann. 1994. Statlog (German Credit Data) Data Set. *UCI Repository of Machine Learning Databases* (1994).

[20] Mikella Hurley and Julius Adebayo. 2016. Credit scoring in the era of big data. *Yale JL & Tech.* 18 (2016), 148.

[21] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

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

[23] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceedings of the 8th Conference on Innovation in Theoretical Computer Science*.

[24] Ravi Kumar and Sergei Vassilvitskii. 2010. Generalized distances between rankings. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, 571–580.

[25] Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval* 3, 3 (2009), 225–331.

[26] Cathy O'Neil. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

[27] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*. 5680–5689.

[28] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2219–2228.

[29] Annie Waldman and SiSi Wei. 2015. Colleges Flush With Cash Saddle Poorest Students With Debt. *Pro Publica* (2015).

[30] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM, 22.

[31] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. ACM, 1171–1180.

[32] 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*. ACM, 1569–1578.