Measuring Group Advantage: A Comparative Study of Fair Ranking Metrics

Caitlin Kuhlman*
Walter Gerych*
Elke Rundensteiner
{cakuhlman,wgerych,rundenst}@wpi.edu
Worcester Polytechnic Institute
Worcester, USA

ABSTRACT

Ranking evaluation metrics play an important role in information retrieval, providing optimization objectives during development and means of assessment of deployed performance. Recently, *fairness of rankings* has been recognized as crucial, especially as automated systems are increasingly used for high impact decisions. While numerous fairness metrics have been proposed, a comparative analysis to understand their interrelationships is lacking. Even for fundamental statistical parity metrics which measure group advantage, it remains unclear whether metrics measure the same phenomena, or when one metric may produce different results than another. To address these open questions, we formulate a conceptual framework for analytical comparison of metrics. We prove that under reasonable assumptions, popular metrics in the literature exhibit the same behavior and that optimizing for one optimizes for all. However, our analysis also shows that the metrics vary in the degree of unfairness measured, in particular when one group has a strong majority. Based on this analysis, we design a practical statistical test to identify whether observed data is likely to exhibit predictable group bias. We provide a set of recommendations for practitioners to guide the choice of an appropriate fairness metric.

CCS CONCEPTS

• **Human-centered computing**;

KEYWORDS

Fairness metrics, fair ranking, group advantage, statistical parity.

ACM Reference Format:
Caitlin Kuhlman, Walter Gerych, and Elke Rundensteiner. 2021. Measuring Group Advantage: A Comparative Study of Fair Ranking Metrics. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3461702.3462588

*Both authors contributed equally to this research.

AIES '21, May 19–21, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462588

1 INTRODUCTION

Increasingly, rankings not only mediate people's access to information online, but also screen and filter candidates in high impact domains such as hiring and university admissions [14, 29]. In these high-stakes settings, many factors may impact *fairness for the candidates being ranked*, including historic bias or misrepresentation of groups in training data [27], biased processing of images [6] and text [5], and implicit bias inherent in users' interaction behavior [8]. Therefore, evaluation metrics for measuring *fairness of rankings* are critically important for evaluating information retrieval (IR) procedures underlying such socio-technical systems.

IR systems typically optimize for multiple concurrent goals, motivating a rich body of work on meta-analysis of evaluation metrics for ranking [1, 7, 26]. In this vein, initial studies have investigated the relationship of fairness to factors such as diversity and novelty [13] and tradeoffs between fairness and relevance [11]. However, the relationships between different fairness metrics themselves have yet to be fully understood. An in-depth understanding of existing metrics is needed to facilitate oversight and guide practitioners in choosing a metric for their application. for measuring unfairness. Therefore, we conduct this comparative analysis of ways of measuring fairness in rankings.

State-of-the-Art Fairness Metrics. Proposed definitions for fair ranking target *group fairness* according to sensitive or legally protected data attributes (e.g., race, gender or age) [9, 14, 19, 27, 31–33]. One straightforward definition of group fairness is *statistical parity*, which simply dictates that each group receive a fair proportion of favorable outcomes. Other fairness definitions have also been proposed for rankings including individual fairness [4], fairness based on equal error-rates [19, 21, 27], and causal definitions [20, 30]. However, the basic notion of how to measure the advantage of one group over another still poses open questions for exploration.

Limitations of Metric Design. Evaluating group fairness hinges on the notion that one group may receive favorable or unfavorable outcomes in some systematically biased way. However for rankings *favorable outcomes are relative* – depending on many factors including the quality of the rest of the items being ranked and the importance of specific positions in the ranking [16]. Proposed fairness metrics account for this by measuring the relative favorability of outcomes for one group over another, or what we refer to as *group advantage*. A number of established strategies in IR have been employed: top-k analysis [9, 31, 32], pairwise inversions [3, 19, 21], and cumulative discounted metrics [14, 27]. Notions of user attention [4] and exposure of items being ranked [27, 33] have

also been explored. Unfortunately, a comprehensive comparative analysis of these approaches is lacking. Guidance is unavailable for deciding when, say, a pairwise metric is preferred to a top-k metric. Without this foundation being well understood, the design of fair ranking metrics is fraught with uncertainty.

Proposed Approach. In this work we study statistical parity metrics for a single protected group in-depth. This definition encapsulates a basic foundation for much fair ranking evaluation metric design. We identify three main approaches to quantifying *group advantage* in a ranking: top-k approaches, exposure metrics, and pairwise metrics. We propose a conceptual framework for evaluating the behavior of metrics in expectation over distributions of rankings. We prove that under a set of reasonable assumptions which characterize the generating function of group advantage, different fairness metrics exhibit the same trends and minima with respect to changing group advantage. Therefore, optimizing for one effectively optimizes for all. However we find that the metrics vary in degree of unfairness, with only one metric assigning a maximum unfair score when a group has a total advantage. This case deserves careful consideration, as it has implications for scenarios where a minority group is strongly disadvantaged.

Applying these observations in practice depends on whether observed data meets our theoretical assumptions. Therefore we design a statistical test to determine the likelihood that observed group advantage exhibits predictable bias. We demonstrate this using real-world sports rankings. Finally we suggest practical strategies for choosing fairness metrics. **Our contributions include:**

(1) We categorize ways of measuring group advantage in statistical parity metrics for fair ranking.
(2) We present a conceptual framework to compare the behavior of fairness metrics in expectation over distributions of rankings characterized by functions of group advantage.
(3) Our analytical evaluation identifies a set of fairness metrics that under reasonable assumptions share the same minima, follow the same trends, and capture fairness well.
(4) We design a statistical test to understand group advantage in rankings, and present recommendations to guide practitioners in selecting the appropriate fairness metric for their application.

2 REVIEW OF STATISTICAL PARITY METRICS

Statistical parity is a simple group fairness definition which requires that a *protected group* receives a fair proportion of favorable outcomes. "Fair" could mean equal shares among groups (as we use in our definitions for simplicity), or an alternative formulation such as a minimum for each according to some apriori specified target, or a proportion based on the size of the groups in the overall population or dataset, depending on the use case.

Given a dataset of candidates $x_i \in X$, the protected group G_p is determined by a *protected attribute* associated with each candidate, traditionally corresponding to legally protected data attributes such as race, gender, or age. The problem setting may also be more expansive, where candidates are text or images representing people [27], or include for instance attributes such as the political leaning of news sources [24]. Statistical parity was first proposed for classification setting where a binary classifier $f(x) = \hat{y}$ assigns each item $x \in X$ to a class in $\{0, 1\}$, and $\hat{y} = 1$ denotes the preferred outcome. Typical definitions of statistical parity [10, 12, 15] then require that $P(\hat{y} = 1 \mid x \in G_p) = P(\hat{y} = 1 \mid x \notin G_p)$.

In rankings there is no binary class assignment with which to evaluate the outcomes for the groups. A ranking is a permutation $\rho = [x_1 \prec x_2 \prec ... \prec x_n]$ over all candidates $x_i \in X$. Here $\prec$ is a total ordering relation on X such that $x_i \prec_\rho x_j$ implies that x_i appears at a *more* preferred position than x_j in the ranking ρ. The position of a single candidate x_i in the ranking ρ is denoted $\rho(x_i)$. We adopt the convention that low number positions are favored over higher ones, i.e. $\rho(x_i) = 1$ is the best rank position. Clearly, being ranked toward the top is a better outcome than being ranked near the bottom, but this determination is inherently relative. Therefore proposed metrics for statistical parity in rankings draw on traditional rank evaluation methods for measuring the *relative advantage* of each group being ranked. Next we review and categorize proposed metrics according to different ways of measuring group advantage.

2.1 Top-k Metrics

A popular method for identifying a favorable outcome in a ranking is by inclusion in a top-k prefix of the ranking [2, 9, 14, 31, 32]. This approach is intuitive since for many tasks the top k rank positions directly correspond to a good outcome for the candidates being ranked, e.g. the top 5 job applicants are invited to interview for a position. We focus on those definitions of top-k statistical parity that give a numerical measure of fairness [14, 31]. Since top-k metrics are highly dependent on the choice of k, a cumulative strategy is typically used to give more emphasis to outcomes at the top of the ranking. Metric scores are weighted by some discounting function $v(k)$ and aggregated. Scores may also be normalized to lie between $[0, 1]$ by computing the ideal (maximum) value R. For simplicity we use a logarithmic discounting function over fixed intervals of top-k prefixes proposed by Yang and Stoyanovich (2017) for all top-k metrics[1], such that the final fairness score for a given metric M is computed:

$$\frac{1}{R} \sum_{k=10,20,30...}^{n} \frac{1}{log_2(k)} M \tag{1}$$

Proposed top-k metrics include:

Normalized discounted difference (rND) [31] measures fairness as the difference between the representation of G_p in the top-k and in the entire ranking.

$$rND(\rho) = |P(x \in G_p \mid \rho(x) \leq k) - P(x \in G_p)| \tag{2}$$

Normalized Discounted Ratio (rRD) [31] compares ratios of outcomes for the protected group G_p and non-protected candidates.

$$rRD(\rho) = \left| \frac{P(x \in G_p \mid \rho(x) \leq k)}{P(x \notin G_p \mid \rho(x) \leq k)} - \frac{P(x \in G_p)}{P(x \notin G_p)} \right| \tag{3}$$

Skew@k [14] computes the logarithmic ratio of outcomes for G_p in the top-k versus the entire list.

$$skew_{G_p}@k(\rho) = \log\left(\frac{P(x \in G_p \mid x \leq k)}{P(x \in G_p)}\right) \tag{4}$$

[1] In the paper by Geyik et al. (2019) skew is computed for a fixed top-k, and rKL aggregated over all k.

Kullback-Leibler Divergence (*rKL*) was first proposed for evaluating statistical parity by Yang and Stoyanovich for the case of a single protected group and then extended to the more general case of multiple groups by Geyik *et al.* In the general case, *rKL* metric is computed as:

$$rKL(\rho) = KL(P||Q) \tag{5}$$

where $P = P(\rho(x) \leq k \mid x \in G_i)\ \forall G_i$, the proportion of each group in the top-k items, and $Q = P(x \in G_i)\ \forall G_i$, the proportion of each group in the entire ranking.

2.2 Exposure Metrics

Singh and Joachims (2018) proposed a statistical parity metric in the context of an IR-focused framework. In this setting, they consider the favorable outcome to be "exposure", a measure of the attention given to a candidate at a particular rank position. The attention score can be given by a discounting function $v(k)$ or some other measure of the importance of each position, for instance learned from implicit user feedback. Equation 6 gives a measure of group advantage for G_p based on the average importance of the rank positions assigned to each candidate $x \in G_p$.

$$exp_{G_p}(\rho) = \frac{S}{|G_p|} \sum_{x \in G_p} v(\rho(x)) \tag{6}$$

The S in Equation 6 is a normalizing constant such that the sum of the exposure of the protected group and the non-protected group is 1. The overall fairness of a ranking is then computed as the absolute difference in exposure for protected and non protected candidates, such that:

$$exp(\rho) = abs\big(1 - 2exp_{G_p}(\rho)\big) \tag{7}$$

In our evaluation we use a reciprocal rank function as our measure of attention where $v(k) = \frac{1}{k}$ for each position k in the ranking and denote the metric *expRR*. However, our analysis is general and does not hinge on any specific function.

2.3 Pairwise Metrics

Finally, several recent works use the pairwise advantage of each group to evaluate parity [3, 19, 21]. Pairwise metrics compute the advantage for one group based on the number of pairwise comparisons it wins against another group in the ranking. In Equation 8 we adopt the convention of normalizing by the number of pairs in the ranking containing candidates from different groups. Here $I(\cdot)$ is the indicator function which evaluates to 1 if $\cdot$ is true and 0 otherwise.

$$pair_{G_p}(\rho) = \frac{1}{|G_p|(|X| - |G_p|)} \sum_{x_i \in G_p} \sum_{x_j \notin G_p} I(\rho(x_i) \prec \rho(x_j)) \tag{8}$$

The overall fairness of a ranking is then determined as the absolute difference in pairwise advantage for the protected group and the non protected candidates, such that:

$$pair(\rho) = abs\big(1 - 2pair_{G_p}(\rho)\big) \tag{9}$$

3 PROPOSED COMPARISON FRAMEWORK

Rather than focus on any discrete single ranking, we describe the behavior of the metrics *in expectation* over distributions of rankings. Such stochastic frameworks are common in information retrieval, for instance they are used to collect unbiased click feedback [17]. Analysis of distributions of rankings is also in line with recent research on producing fair rankings [11, 27].

We propose the use of a matrix $R \in \mathbb{R}^{NxN}$ to represent the distributions over rankings, where N denotes the number of candidates to be ranked. Each row of R represents a position in the ranking, and each column a candidate. Each entry $R_{i,j}$ gives the probability that candidate x_j is assigned rank position i as expressed in Equation 10. As each element of R represents the probability that a given element is in a given position, the rows and columns must each add up to 1 and thus we call R doubly stochastic. For single discrete ranking, R is a binary matrix where $R_{i,j} = 1$ iff $\rho(x_j) = i$, and $R_{i,j} = 0$ otherwise.

$$R_{i,j} = P\big(\rho(x_j) = i\big) \tag{10}$$

Modelling Group Advantage. Fairness metrics are predicated on the belief that one ranking can give a more preferred outcome to one group than another. Therefore we propose to model this unfair advantage as a random variable α which can take on some range of values. For convenience we can choose $\alpha \in [0, 1]$ where a score of 0 means that a group is at a complete disadvantage, and 1 indicates a total advantage over other groups. For simplicity in our analysis we consider a single value for α representing the advantage of the protected group G_p, understanding that this equivalently implies an advantage or disadvantage for non-protected candidates. To model this in our framework, we now impose additional structure on R to represent the probability of each candidate being assigned each position as a function of *group advantage* α. Let us assume that each entry in R is a function such that:

$$R_{i,j} = f_{i,j}(\alpha), \quad f : [0, 1] \rightarrow [0, 1] \tag{11}$$

This framing reflects the fact that group advantage does not impact an entire ranking uniformly – it varies if evaluated at each position in the ranking. For instance, at position $i = 1$, if G_p has a large advantage and many possible candidates to choose from, it is highly likely that a protected candidate will be assigned to the top spot. However, for a lower rank position most protected candidates will have been assigned to positions above, and a non-protected item will now be more likely to be assigned. Therefore although the overall advantage does not change, $f_{i,j}(\alpha)$ at different rank positions i gives different likelihoods of assignment for x_j.

3.1 Reformulating Fair Ranking Metrics

Next we represent the statistical parity metrics using our framework. Each type of metric can be represented in a commensurable way as functions of the group advantage α.

Top-k Metrics. Each top-k metric measures distance or divergence between the proportion of the protected group G_p in the top-k, and the proportion of G_p in the entire ranking. We compute these values with the doubly stochastic matrix R as:

$$P = P(x \in G_p \mid \rho(x) \leq k) = \frac{1}{k} \sum_{i=1}^{k} \sum_{j \in G_p} f_{i,j}(\alpha) \tag{12}$$

$$Q = P(x \in G_p) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in G_p} f_{i,j}(\alpha) \tag{13}$$

Given this, the top-k metrics can easily be expressed as follows:

$$rND(\rho) = |P - Q| \tag{14}$$

$$rRD(\rho) = \Big|\frac{P}{1-P} - \frac{Q}{1-Q}\Big| \tag{15}$$

$$skew_{G_p}@k(\rho) = log(\frac{P}{Q}) \tag{16}$$

$$\begin{aligned} rKL(\rho) &= P * ln\Big(\frac{P}{Q}\Big) \\ &\quad - (1-P) * ln\Big(\frac{1-P}{1-Q}\Big) \end{aligned} \tag{17}$$

Exposure Metrics. The exposure metrics are calculated using the rank position of individual candidates $\rho(x)$. When considering distributions over rankings, the candidates no longer have only one rank - instead R gives a probability of assignment in a position. Therefore we replace $\rho(x)$ with its expected value $\mathbb{E}(\rho(x))$ and define the exposure metrics in terms of distributions of rankings.

$$\begin{aligned} exp_{G_j}(\rho) &= \frac{1}{|G_j|} \sum_{x_j \in G_j} \mathbb{E}(\rho(x_j)) \\ &= \frac{1}{|G_j|} \sum_{x_j \in G_j} v(i) \sum_{i=1}^{N} i \cdot f_{i,j}(\alpha) \end{aligned}$$

Pairwise Metrics. To represent the *pair* metric, we observe that the likelihood that $x_i \prec x_j$ in expectation is given by the difference in the expected rank position for each candidate x_i, x_j where:

$$\mathbb{E}(\rho(x_j)) = \sum_{i=1}^{N} i\, f_{i,j}(\alpha) \tag{18}$$

To compute the expected *pair* value we take the signed difference of the expected positions of each pair of protected and non-protected candidates:

$$pair = \frac{1}{N}\Big|\sum_{x_i \in G_p} \sum_{x_j \notin G_p} sign(\mathbb{E}(\rho(x_i)) - \mathbb{E}(\rho(x_j)))\Big| \tag{19}$$

3.2 Modelling Group Advantage

With our comparison framework in place, we next define a family of advantage functions f according to reasonable assumptions for group advantage. These assumptions describe an intuitive notion of group advantage controlling the distribution of groups throughout a ranking. Figure 1 illustrates this scenario for sets of randomly generated rankings. The protected group is represented by black squares, and we see that for different values of α the protected group is distributed in different ways.

Assumption 1.

$$\sum_{j \in G_p} f_{i,j}(\alpha) \geq \sum_{j \in G_p} f_{i+1,j}(\alpha) \quad \text{if } \alpha \geq \frac{|G_p|}{N}$$

Assumption 1 states that if α is greater than the overall probability of observing the protected group, then candidates in G_p will have a higher probability of being assigned favorable positions toward the top of the ranking, with uniformly decreasing probability for the lower positions (i.e. G_p has an advantage over other groups). We can see this case when $\alpha > 0.2$ in Figure 1.

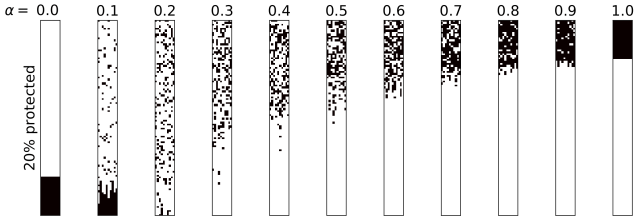


Figure 1: Sets of 10 random rankings that conform to Assumptions 1 and 2. 20% of candidates are in the protected group and α values are varied.

Assumption 2.

$$\sum_{j \in G_p} f_{i,j}(\alpha) \leq \sum_{j \in G_p} f_{i+1,j}(\alpha) \quad \text{if } \alpha \leq \frac{|G_p|}{N}$$

Assumption 2 conversely states that if the advantage is less than this value, then G_p is uniformly more likely to be observed as you move down the ranking (i.e. there is a protected group disadvantage). Together these assumptions imply that if α equals the proportion of protected candidates in the entire ranking, then all candidates are equally likely to be assigned to any position. We thus would expect our fairness metrics to deem such rankings as perfectly fair on average. Figure 1 illustrates this case when $\alpha = 0.2$.

4 COMPARATIVE METRIC ANALYSIS

From these intuitive assumptions, we now characterize the behavior of the fairness metrics with respect to group advantage in the following theorems (see Appendix for full proofs).

Optimizing for One Metric Optimizes for All.

Theorem 1. *Given a ranking ρ with a protected group of candidates G_p and associated advantage α, if Assumptions 1 and 2 hold, then the rND, rRD, rKL, expRR, and pair metrics share the same minima.*

Proof Intuition. We show that by definition, when $P = Q$, all metrics equal 0, otherwise the metric values are greater than 0. We note that the one exception is *skew*. When $P < Q$, *skew* is negative, therefore it does not share the same minima as the other metrics.

Improving Fairness According to One Metric Improves the Other Metrics As Well.

Theorem 2. *Given a ranking ρ with a protected group of candidates G_p and associated advantage α, if Assumptions 1 and 2 hold, then signs of the derivative with respect to α of the rND, rKL, rRD, and expRR metrics are the same.*

Proof Intuition. The slopes of each of metrics are the same everywhere other than the critical points when the metrics are expressed

as functions of α. In particular, for each metric M:

$$sign(\frac{d}{d\alpha}M) = sign(\frac{d}{d\alpha}\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha)) \quad \text{if } \alpha > Q$$

$$sign(\frac{d}{d\alpha}M) = -1 \cdot sign(\frac{d}{d\alpha}\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha)) \quad \text{if } \alpha < Q$$

We do not include the *pair* metric in this analysis since *pair* is computed using the discrete set of possible pairs in the ranking, and therefore the function has a non-continuous range. In our empirical evaluation in Section 6 we observe that the pairwise metric does indeed exhibit similar behavior to the rest of the metrics.

Pairwise Metrics Yield a Maximum Value When Either Group Has a Total Advantage.

Theorem 3. *Given a ranking ρ with a protected group of candidates G_p, $pair(\rho)$ has its maximum value when $\alpha = 0$ or $\alpha = 1$, meaning when one group has a total advantage.*

Proof. If either group has total advantage, then

$$sign(\mathbb{E}(\rho(x_i)) - \mathbb{E}(\rho(x_j))) = sign(\mathbb{E}(\rho(x_m)) - \mathbb{E}(\rho(x_n)))$$

$\forall x_i, x_m \in G_p$ and $x_j, x_n \notin G_p$. Thus,

$$pair = \frac{1}{N}\Big|\sum_{x_i\in G_p}\sum_{x_j\notin G_p} sign(\mathbb{E}(\rho(x_i)) - \mathbb{E}(\rho(x_j)))\Big| = 1$$

This is the maximum *pair* value, because

$$sign(\mathbb{E}(\rho(x_i)) - \mathbb{E}(\rho(x_j))) \neq sign(\mathbb{E}(\rho(x_m)) - \mathbb{E}(\rho(x_n))) \implies \frac{1}{N}\Big|\sum_{x_i\in G_p}\sum_{x_j\notin G_p} sign(\mathbb{E}(\rho(x_i)) - \mathbb{E}(\rho(x_j)))\Big| < 1$$

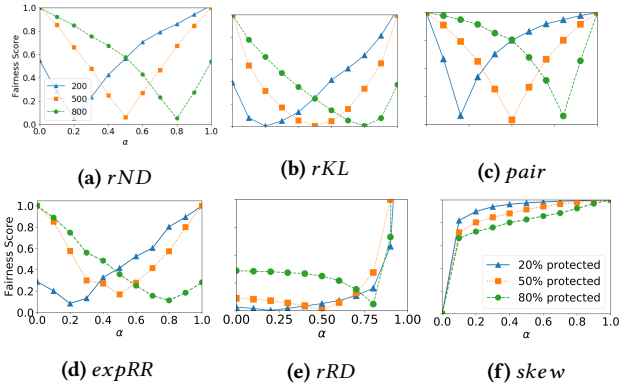
□

5 MONOTONICITY TEST

In our analysis so far, we consider functions of advantage f which are applied *monotonically* throughout the ranking. However, in the real world other factors may impact the probability of candidates being assigned to positions. Observed sample data is likely to be noisy and inconsistent. Or bias may be injected adversarially. In practice it is unlikely that the generating function of bias is known, necessitating a practical strategy to assess whether the results in Section 4 are likely to hold. For this, we now design a statistical test that can be applied to determine whether a given set of n observed rankings meets Assumptions 1 and 2. [2]

We first calculate for each rank position the probability that a member of the protected group is observed in that position by counting the instances in the sample data. Let us refer to these calculated probabilities as $\hat{M} : \sigma \Rightarrow [0, 1]$, where σ denotes the set of rank positions. If $\hat{M}$ is monotonically increasing or decreasing, then our assumptions are met. However, empirical probabilities may not be strictly monotonic due to noise. To overcome this, we apply a goodness of fit test between the empirical distribution and a monotonic target distribution in order to determine the likelihood that *the true generating distribution* M is monotonic.

For the target distribution, we find the closest-fitting monotonic function $M : \sigma \Rightarrow [0, 1]$, such that M is a function from ranking positions σ to probabilities. M is found using isotonic regression [25] fit to the observed data. Then given this ideal generating function and the previously described empirical distribution, we perform a Chi-Square goodness of fit test [22] between $\hat{M}$ and M. If the p-value returned by the test is greater than some desired level of significance (e.g., 0.05), then there is no statistical difference between the observed ranking probabilities and a monotonic function and we conclude that our assumptions will hold with high probability.



(a) rND (b) rKL (c) $pair$

(d) $expRR$ (e) rRD (f) $skew$

Figure 2: Fairness metrics applied to random rankings that conform to Assumptions 1 and 2.

6 EMPIRICAL EVALUATION

Next we verify our theoretical observations on group advantage in a controlled study, and then demonstrate the applicability of our monotonicity test in a case study on sports ranking data.

Methodology. We produce random rankings with different sized protected groups: minority (20% protected), balanced (50% in each group), and majority (80% protected). Group advantage following our Assumptions in Section 3.2 is generated following the methodology proposed in [31]. For each experiment results are averaged over 10 runs. We evaluate metrics rND, rRD, rKL, $skew$, $pair$ and $expRR$. Metric values are evaluated for varying levels of advantage $\alpha \in [0.0, 1.0]$. To facilitate comparison, we normalize each metric to lie in a range of zero to one (scaling based on the minimum and maximum possible values given the size of the groups).

Observations. Figure 2 compares metric behavior across different values of group advantage α for rankings with different size groups. There are clearly observable similar patterns for metrics rND, rKL, $pair$, and $expRR$ which share trends and minima. This aligns with our analysis in Theorems 1 and 2. We can also see a key difference among these metrics. Following Theorem 3 the *pair* metric always assigns a maximum unfair score in the extreme cases where one group is completely advantaged over the other (when

[2] We provide additional analysis of alternative advantage functions that do not conform to our assumptions in a supplemental report, along with all code and data to reproduce our experiments: https://github.com/waltergerych/AIES_2021_Measuring_Group_Advantage

$\alpha = 1.0$ or $\alpha = 0$), while the other metrics do not. We observe that *rND*, *rKL*, and *expRR* only consider the case of total advantage to be highly unfair when the size of the groups are balanced, or when it benefits a majority protected group. If a small minority group is totally disadvantaged these metrics give fairness scores around 0.5. For the *rRD* metric the minima are the same as the other metrics, but when the protected group is totally favored, the *rRD* metric explodes. We have scaled the figure based on the max value for $\alpha = 0.9$ for readability. The *skew* metric can be observed to follow a totally different pattern with respect to group advantage.

Case Study on Sports Ranking Data. To study the applicability of our analytic results on real data, we consider a dataset of rankings of National Football League (NFL) players [18]. In fantasy sports games, players score points based on the performance of real athletes. Rankings of the athletes are provided by experts to give guidance to fantasy players. For the first 15 weeks of the 2019 football season, we analyzed rankings from roughly 90 different experts who ranked the top 10 quarterbacks each week. We consider these as samples from a distribution of rankings. Players were assigned to groups based on team conferences: the American Football Conference (AFC) which we consider the protected group and the National Football Conference (NFC).

We apply the monotonicity test proposed in Section 5 to evaluate whether the data are likely to conform to our standard assumptions of bias. We use isotonic regression to determine the closest monotonically increasing or decreasing function to the observed data. Figure 3 shows the observed and ideal functions for each week. Our test finds 10 out of the 15 sets of weekly rankings likely to meet our assumptions of bias – meaning even on this small sample size, most rankings are likely to reflect monotonic bias functions. However, we also see from Figure 3 that it is possible for bias to follow a variety of different patterns.

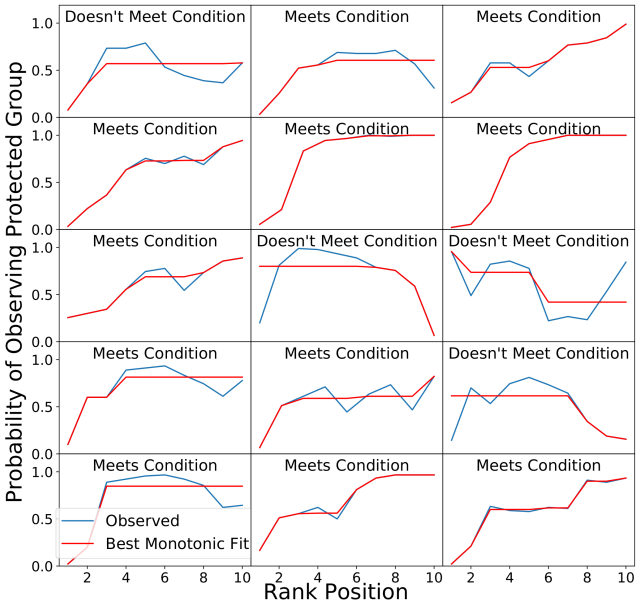


Figure 3: Monotonic functions fit to observed advantage functions for weekly sports rankings.

7 DISCUSSION AND RECOMMENDATIONS

Together our theoretical analysis and our empirical study provide insights into the relationships among different fair ranking metrics. We see that when advantage can be expected to conform to the assumptions laid out, the top-*k* metrics (with the exception of *skew*) share the same behavior as the pairwise and exposure based metrics. Therefore, **minimizing one will minimize all, and improving any metric will improve the others.**

However, the metrics do vary in the degree of unfairness measured. We prove in Theorem 3 that the *pair* metric has a maximum value when either group has a total advantage over the other. This matches the intuitive notion that a total advantage is the most extreme violation of statistical parity possible. On the other hand, as we observe empirically in Section 6, **most metrics will not flag rankings as unfair which strongly disadvantage a minority group**. One reason for this could be that if ranked at random, a group with many more candidates is more likely to have an advantage by chance. However, when the protected group is a small minority this case may crucially be when fairness evaluation is needed most. Therefore, these competing notions of what constitutes unfairness deserve careful consideration when selecting a fairness metric in any applied setting.

Recommendations. Following from these key observations, and bearing in mind that not all unfair bias in rankings will match our assumptions, we put forth the following recommendations for the practical application of fair ranking metrics.

If the monotonicity test indicates that observed data *is not* likely to conform to standard bias, metrics can have unexpected behavior. In that case we suggest evaluating the output of all metrics on the observed data, especially considering unusual or edge cases. It is imperative to formulate a well-defined notion of fairness for your application in order to choose the metric which aligns best.

If the observed data passes the monotonicity test, and the application will aim to *minimize* any bias detected, then we recommend using any of the *rKL*, *rND*, *pair*, or *exp* metrics. Based on Theorem 1, they are equivalent. *rRD* and *skew* metrics are not recommended, as they can exhibit unbounded values. Also, *skew* does not align with the other definitions. Choice among the recommended metrics depends on the application - for instance, if the attention function is known then an exposure based metric may be preferred. A smooth function will be easiest to mathematically optimize, so therefore *rKL* may be preferred out of the top-*k* metrics as observed in [31]. Finally, if bias is not guaranteed to be minimized, but rather measured for evaluation, we recommend using the *pair* metric – especially if the protected group is a minority. Other measures may not capture the bias well when the majority group has an advantage.

8 RELATED WORK

Recent research evaluates the applicability of fairness metrics specifically for information retrieval [11, 13, 23, 28]. Pitoura *et al.* consider general categories of user and content bias and frame fairness

metrics as similarity measures. Verma *et al.* provide a comprehensive survey of metrics proposed to-date. Gao and Shah empirically compare fairness metrics with diversity and novelty metrics. They consider exposure-based statistical parity along with a number of diversity metrics. Diaz *et al.* propose stochastic distributions of rankings be used as an evaluation framework for evaluating exposure-based fairness metrics. This work follows an individual fairness paradigm and considers expected exposure in relation to the relevance of documents. These works are complementary to ours, particularly [11, 13] which relate exposure-based strategies for measuring group advantage to other (non-fairness) metrics for ranking.

Our analysis builds on work [31] which proposes multiple top-k fair ranking metrics and compares them with respect to group size and advantage. We broaden the scope of this initial within-class comparison to now include pairwise and exposure based metrics, evaluated for distributions of rankings.

9 CONCLUSION

In this work, we offer a comparative analysis of key statistical parity evaluation metrics for rankings. Our analysis reveals fundamental similarities among metrics from the literature that on the surface appear diverse, under common assumptions about the relative advantage of the groups. This is important to the field, in that work optimizing for one of these metrics now can be shown to equally optimize for all. However, a key distinction is noted in the magnitude of fairness scores when a minority group is at a total disadvantage.In addition, we propose a statistical test to evaluate group advantage and include a case study and strategies for choosing a fair ranking metric in practice.

ACKNOWLEDGMENTS

This work is supported by the NSF FAIR Grant 2007932, entitled "III: Small: Fair Decision Making by Consensus: Interactive Bias Mitigation Technology". The authors also would like to thank DARPA WASH Grant HR00111780032-WASH-FP-031. Lastly, we would like to thank Kathleen Cachel of Worcester Polytechnic Institute's DAISY lab for her valuable feedback on the manuscript.

10 APPENDIX

10.1 Proof of Theorem 1

Proof. By the definition of *rKL*, *rRD*, *rND*, *expRR*, *expDCG*, and *pair*, each metric equals 0 when $P = Q$. Let a be a value between 0 and 1 such that when $\alpha = a$, $P = Pr(x \in G_p \mid \rho(x) \leq k) = \frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(a) = Q$. Thus each of the aforementioned metrics equals 0 when $\alpha = a$.

Furthermore, if $\alpha > a$ then by Assumption 1, $\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha) \geq Q \,\forall\, k$ and is strictly greater than Q for some k. Conversely, if $\alpha < Q$ then $\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha) \leq Q \,\forall\, k$ (by Assumption 1) and is strictly less than Q for some k. As each of the metrics is > 0 when $P \neq Q$, then when considering a sum over all k the minimum will occur only at $\alpha = a$. □

10.2 Proof of Theorem 2

Proof. We show that the slopes of the *rND*, *rKL*, *rRD*, and *expRR* metrics are the same everywhere other than the critical points (i.e. at the minimum value and at the limits of the domain of α, where the derivative is undefined) when the metrics are expressed as functions of α. We begin by showing the slope for *rKL*:

$$\frac{d}{d\alpha}rKL = (ln(\frac{P}{Q}) + 1)\cdot\frac{d}{d\alpha}\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha)$$

$$+(-ln(\frac{1-P}{1-Q}) - 1)\cdot\frac{d}{d\alpha}\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha)$$

$$=(ln(\frac{P}{Q}) - ln(\frac{1-P}{1-Q}))\cdot\frac{d}{d\alpha}\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha)$$

If $\alpha > Q$ then $P > Q$ and thus $ln(\frac{P}{Q}) - ln(\frac{1-P}{1-Q}) > 0$. Conversely, if $\alpha < 0$ then $ln(\frac{P}{Q}) - ln(\frac{1-P}{1-Q}) < 0$. Thus if $\alpha > Q$:

$$sign(\frac{d}{d\alpha}rKL) = sign(\frac{d}{d\alpha}\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha))$$

and if $\alpha < Q$:

$$sign(\frac{d}{d\alpha}rKL) = -sign(\frac{d}{d\alpha}\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha))$$

Moving on to *rND*:

$$\frac{d}{d\alpha}rND = \frac{P-Q}{|P-Q|}\cdot\frac{d}{d\alpha}\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha)$$

As $\frac{P-Q}{|P-Q|} > 1$ when $\alpha > Q$ and $v\frac{P-Q}{|P-Q|} < 1$ when $\alpha < Q$, then if $\alpha > Q$:

$$sign(\frac{d}{d\alpha}rND) = sign(\frac{d}{d\alpha}\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha))$$

and if $\alpha < Q$:

$$sign(\frac{d}{d\alpha}rND) = -sign(\frac{d}{d\alpha}\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha))$$

Next, we show the slope of *rRD*:

$$\frac{d}{d\alpha}rRD = \frac{\left(\frac{x}{(1-x)^2} + \frac{1}{1-x}\right)\left(\frac{x}{1-x} - \frac{Q}{1-Q}\right)}{\left|\frac{x}{1-x} - \frac{Q}{1-Q}\right|}$$

$$\cdot\frac{d}{d\alpha}\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha)$$

When $\alpha > Q$ then $\left(\frac{x}{1-x} - \frac{Q}{1-Q}\right) > 0$, which implies:

$$\frac{\left(\frac{x}{(1-x)^2} + \frac{1}{1-x}\right)\left(\frac{x}{1-x} - \frac{Q}{1-Q}\right)}{\left|\frac{x}{1-x} - \frac{Q}{1-Q}\right|} > 0$$

Conversely, if $\alpha < Q$ then:

$$\frac{\left(\frac{x}{(1-x)^2} + \frac{1}{1-x}\right)\left(\frac{x}{1-x} - \frac{Q}{1-Q}\right)}{\left|\frac{x}{1-x} - \frac{Q}{1-Q}\right|} < 0$$

$$sign(\frac{d}{d\alpha}rRD) = sign(\frac{d}{d\alpha}\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha))$$
$$\text{if } \alpha > Q$$
$$sign(\frac{d}{d\alpha}rRD) = -sign(\frac{d}{d\alpha}\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha))$$
$$\text{if } \alpha < Q$$

Moving on to the exposure rankings:

$$\begin{aligned}\frac{d}{d\alpha}exp_{G_p}(\rho) &= \frac{d}{d\alpha}|\frac{1}{|G_p|}\sum_{i=1}^{k}\sum_{j\in G_p} v_i\cdot i\cdot f_{i,j}(\alpha) \\ &\quad -(1-\frac{1}{|G_p|}\sum_{i=1}^{k}\sum_{j\in G_p} v_i\cdot i\cdot f_{i,j}(\alpha))| \\ &= \frac{d}{d\alpha}|\frac{2}{G_p}\sum_{i=1}^{k}\sum_{j\in G_p} v_i\cdot i\cdot f_{i,j}(\alpha)-1| \\ &= \frac{\frac{2}{G_p}\sum_{i=1}^{k}\sum_{j\in G_p} v_i\cdot i\cdot f_{i,j}(\alpha)-1}{|\frac{2}{G_p}\sum_{i=1}^{k}\sum_{j\in G_p} v_i\cdot i\cdot f_{i,j}(\alpha)-1|} \\ &\quad \cdot\frac{d}{d\alpha}\frac{2}{G_p}\sum_{i=1}^{k}\sum_{j\in G_p} v_i\cdot i\cdot f_{i,j}(\alpha)\end{aligned}$$

If $\alpha < q$, then the term that the above derivative is multiplied by is less than 0, and if $\alpha > Q$ then the term is positive. Additionally, for v_i as defined for exposure the following holds:

$$sign(\frac{d}{d\alpha}\frac{2}{G_p}\sum_{i=1}^{k}\sum_{j\in G_p} v_i\cdot i\cdot f_{i,j}(\alpha)) = sign(\frac{d}{d\alpha}\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha))$$

Thus, if $\alpha > Q$:

$$sign(\frac{d}{d\alpha}exp_{G_p}) = sign(\frac{d}{d\alpha}\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha))$$

and if $\alpha < Q$:

$$sign(\frac{d}{d\alpha}exp_{G_p}) = -sign(\frac{d}{d\alpha}\frac{1}{k}\sum_{i=1}^{k}\sum_{j\in G_p} f_{i,j}(\alpha))$$

□

REFERENCES

[1] Enrique Amigó, Damiano Spina, and Jorge Carrillo-de Albornoz. 2018. An axiomatic analysis of diversity evaluation metrics: Introducing the rank-biased utility metric. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 625–634.

[2] Abolfazl Asudeh, HV Jagadish, Julia Stoyanovich, and Gautam Das. 2019. Designing fair ranking schemes. In *Proceedings of the 2019 International Conference on Management of Data*. ACM, 1259–1276.

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

[4] 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*. ACM, 405–414.

[5] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*. 4349–4357.

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[7] Luca Busin and Stefano Mizzaro. 2013. Axiometrics: An axiomatic approach to information retrieval effectiveness metrics. In *Proceedings of the 2013 Conference on the Theory of Information Retrieval*. 22–29.

[8] L Elisa Celis, Anay Mehrotra, and Nisheeth K Vishnoi. 2020. Interventions for ranking in the presence of implicit bias. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 369–380.

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

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

[11] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. arXiv:2004.13157 [cs.IR]

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 214–226.

[13] Ruoyuan Gao and Chirag Shah. 2020. Toward creating a fairer ranking in search engine results. *Information Processing & Management* 57, 1 (2020), 102138.

[14] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA). ACM, 2221–2231.

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

[16] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum*, Vol. 51. Acm New York, NY, USA, 4–11.

[17] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 781–789.

[18] Caitlin Kuhlman and Elke Rundensteiner. 2020. Rank aggregation algorithms for fair consensus. *Proceedings of the VLDB Endowment* 13, 12 (2020), 2706–2719.

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

[20] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 2243–2251.

[21] Harikrishna Narasimhan, Andy Cotter, Maya Gupta, and Serena Lutong Wang. 2020. Pairwise Fairness for Ranking and Regression. In *33rd AAAI Conference on Artificial Intelligence*.

[22] Karl Pearson. 1900. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50, 302 (1900), 157–175.

[23] Evaggelia Pitoura, Panayiotis Tsaparas, Giorgos Flouris, Irini Fundulaki, Panagiotis Papadakos, Serge Abiteboul, and Gerhard Weikum. 2018. On Measuring Bias in Online Information. *ACM SIGMOD Record* 46, 4 (2018), 16–21.

[24] Ronald E Robertson, David Lazer, and Christo Wilson. 2018. Auditing the personalization and composition of politically-related search engine results pages. In *Proceedings of the 2018 World Wide Web Conference*. 955–965.

[25] Wright F Robertson, T and R Dykstra. 1989. Order restricted statistical inference. *Statistical Papers* (1989).

[26] Tetsuya Sakai. 2012. Evaluation with informational and navigational intents. In *Proceedings of the 21st international conference on World Wide Web*. 499–508.

[27] 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.

[28] Sahil Verma, Ruoyuan Gao, and Chirag Shah. 2020. Facets of Fairness in Search and Recommendation. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 1–11.

[29] Austin Waters and Risto Miikkulainen. 2014. Grade: Machine learning support for graduate admissions. *AI Magazine* 35, 1 (2014), 64–64.

[30] Yongkai Wu, Lu Zhang, and Xintao Wu. 2018. On discrimination discovery and removal in ranked data using causal graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2536–2544.

[31] 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, 22:1–22:6.

[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.

[33] Meike Zehlike and Carlos Castillo. 2018. Reducing Disparate Exposure in Ranking: A Learning To Rank Approach. *arXiv preprint arXiv:1805.08716* (2018).