Bachelor thesis
Computing Science

Radboud University

A Comparative Study of Group Fairness Metrics for Ranked Outputs

Author:
Muhammad Salibi
muhammad.salibi@ru.nl
s1014926

First supervisor/assessor:
prof. dr. ir. D. Hiemstra
hiemstra@cs.ru.nl

Second assessor:
dr. ir. E. Herder
eelcoherder@cs.ru.nl

March 23, 2022

Abstract

The increasing importance of ranking algorithms as they become an essential part of our daily lives is undeniable. This increase comes together with concerns about the fairness of these algorithms. Novel fairness aware algorithms are often suggested together with a novel fairness metric to assess the algorithm fairness. When considering the various fairness goals and definitions it becomes apparent the difficulty to compare one metric to the other and decide when a metric is suitable for each use case and thus the need for a comparative study of fairness metrics. In this paper, I compare different definitions of fairness and use synthetically generated rankings to test group-fairness metrics that aim to measure statistical parity in the ranking. I find that although these metrics are defined with the same fairness goal in mind, they behaved differently because of different interpretations of statistical parity in ranked outputs. The results of the tests are also used to show limitations of the metrics in cases where it is desired to detect bias *against*, as well as *towards*, a protected group and where the difference between the proportions of the protected and unprotected group is expected to be large.

Contents

Chapter 1

Introduction

Ranking algorithms are found in abundance in our daily lives. They are used for various applications including search engines to show the top results matching a search query, recommendations for movies or videos a user might be interested to watch next, deciding which applicants to invite to an interview, and even governments use these system for applications such as fraud risk assessment. Many examples exist that raise questions about the fairness considerations in such algorithms; Google Images search with keywords of professions returning gender biased results [Kay et al., 2015], the current popularity of a video on YouTube being the most important factor of predicting its future popularity leading to a *rich-gets-richer* dynamics [Borghol et al., 2012]. Or Amazon's automated hiring tool discriminating against female applicants [Dastin, 2018] and the concerns about the System Risk Indicator (SyRI) algorithm used by the Dutch government to detect fraud in social welfare possibly discriminating against neighborhoods with low average income and residents of immigrant background [van Bekkum and Borgesius, 2021] among many other examples. Therefore, the issue of fairness in ranking algorithms has recently gained increasing attention, introducing Fairness-aware Ranking algorithms that attempt to increase fairness in the outputs with minimal negative effect to utility.

To create fairness-aware algorithms, it is essential to first clearly define fairness in the given context and the fairness goal that needs to be met. Secondly, based on the fairness goal defined, create a method or a metric to assess the output of the algorithm to be able to quantify to what extent that output meets the defined fairness goal. It is often the case that a novel fair algorithm is suggested together with a new metric and is optimized for that particular metric but not for others [Chen et al., 2020] making it challenging to compare the performance of different algorithms and creating a need for a comparative analysis of the different fairness metrics. Note that in this document I use the term *fairness metrics* to refer to methods of assessing and quantifying fairness and refrain from using the term *fairness*

measures when possible which is used in literature to refer to both metrics and methods for mitigating unfairness.

Work has been done in comparing different fairness-aware classification models [Friedler et al., 2019, Mehrabi et al., 2021] and Learn-To-Rank systems [Chen et al., 2020], but little has been done in comparing the different fairness metrics used to assess fairness of the results [Garg et al., 2020] and even less comparisons that focus on fairness metrics for ranked outputs. Hence, this research. Since the topic is relatively young and consensus over definitions of fairness and the implications thereof are difficult to find and to match the weight of a bachelor's thesis I limit the scope of this research to definitions of group fairness that are based on group proportions as defined later in this chapter, attempting to explore some of the properties of such metrics when used to assess synthetically generated ranked outputs and hereby answering the research question: **What lessons can be learned from comparing different proportion-based group fairness metrics used in literature against a common ranking?**

Definitions of Fairness

Defining fairness proves to be a challenging task for philosophy, psychology and computer science alike. Although no universal definition exists that can capture all aspects of fair treatment towards a certain population there exist multiple notions of fairness, each of which attempt to satisfy a certain condition that is relevant for the task at hand. Most of the fairness conditions found in literature are suited for classification problems and are not directly applicable to ranking problems where only one item can be assigned a specific rank and a higher position in the ranking is more beneficial. The different definitions mentioned here are also summarized in Table 1.1

Notions of **Group Fairness**, also referred to as *Provider Fairness*, generally aim at treating different groups similarly. These definitions either use an arbitrary number of group labels with the goal of ensuring similar fairness conditions across all groups, or limit the group labels to two labels distinguishing between a protected group and an unprotected group. Fairness conditions that fall under group fairness include *Equalized Odds* where members of the protected and unprotected groups have equal true positive rate and (separately) equal false positive rate. As well as *Demographic Parity*, also referred to as statistical parity, which requires that the probability of receiving a positive outcome or treatment be independent of group membership. [Garg et al., 2020, Mehrabi et al., 2021].

It is often believed that enforcing fairness among groups may introduce injustice at the individual level [Dwork et al., 2012], e.g. a member of the protected group receiving a positive outcome to ensure group fairness while there exists a member of the unprotected group that scores higher for some

Table 1.1: Summery of Fairness Definitions

Fairness for Predictive Classifiers	Fairness in Ranking Algorithms
• **Group Fairness** Treat groups similarly. Fairness conditions include: – *Equalized Odds* Equal true positive rate and equal false positive rate – *Demographic Parity* Probability of receiving a positive treatment is independent of group membership • **Individual Fairness** Similar items receive similar treatment	• **Proportion-Based Fairness** Proportion of members of each group equals a target proportion • **Exposure-Based Fairness** Defined in terms of *exposure* provided by each position in the ranking. Allows to be: – Defined for individual fairness when calculated over individual items. – Defined for group fairness when accumulated over group labels

notion of merit or relevance but does not receive a positive treatment. This has led to the introduction of notions of **Individual fairness** where the goal is to ensure that items with similar merit receive similar treatment or outcome. While it is generally agreed upon that notions of group and individual fairness are incompatible with each other, I tend to agree with the opinions that argue that these different notions do not fundamentally conflict [Binns, 2020], but any apparent conflict between them is a result of not carefully applying the suitable measures for the given situation. In other words, setting a fairness goal that is different than the moral motivation behind introducing fairness in the system. Which emphasizes the importance of understanding the implications of various fairness definitions before applying them to a certain system.

The Demographic parity fairness condition can be extended to ranked outputs by ensuring that the proportion of members of each group equals a target proportion [Yang and Stoyanovich, 2017, Zehlike et al., 2017]. Methods that use this definition of fairness are categorized in [Kırnap et al., 2021] as **Proportion-Based Fairness**. The target proportion can either be an input variable which is decided to be a fair target proportion by experts or legal entities, or derived from the ranking itself and the goal is to ensure that distributions of the groups at certain top positions in the ranking match that of the whole ranking. In contrast, **Exposure-Based Fairness** defines fairness in terms of *exposure* or *attention* provided by each position in the ranking. Exposure can be calculated either for individual items in the ranking to achieve goals of individual fairness or accumulated over groups to achieve group fairness.

With the various definitions of fairness it is difficult to compare metrics that are based on different fairness conditions or aim to achieve different fairness goals. Therefore, in this research I focus on proportion-based group fairness definitions. Metrics based on this definition are commonly adopted in literature and the conditions of demographic parity they are based on is found to most closely match people's intuitive ideas about fairness [Srivastava et al., 2019]. In the reminder of this document I discuss related work, their conclusions and how it relates to this research in Chapter 2. The formal definitions of the metrics chosen for the comparison are described in Chapter 3. The method I follow and the results of the tests are described in Chapters 4 and 5, respectively. In Chapter 6 I discuss the conclusions drawn for the tests and future work.

The code used to assess the different fairness metrics is available at `https://github.com/muhamadsalibi/comparing_fairness_measures`

Chapter 2

Related Work

Categorizing Fairness Definitions

Fairness definitions are categorized in different ways in literature depending on the goal of the research. In this paper I follow a categorization that can capture the general fairness goal of the metrics without being limited to specific metrics, similar to the categorization used in [Garg et al., 2020, Kırnap et al., 2021]. However, other categorizations exist. For example, [Raj et al., 2020] differentiates between two main categories:

- **Single-List Metrics** metrics that operate over a single ranked output. Including a *prefix fairness* family [Yang and Stoyanovich, 2017] which I evaluate as well in this research. [Raj et al., 2020] point out that these metrics only take into consideration a fair distribution of opportunity (i.e. demographic parity) and do not account for relevance. Therefore, these metrics should be integrated with metrics that do account for relevance.
- **Distribution and Sequence Metrics** metrics that assess the fairness of a sequence of rankings. Including an *exposure family* of metrics (Exposure-based fairness) which take into account relevance of the items in the rankings.

Evaluating Fairness Metrics

The authors of [Draws et al., 2021] adapt previously defined fairness metrics for ranked outputs to use for assessing viewpoint diversity in search results. Their work is based on metrics defined in [Yang and Stoyanovich, 2017] and test how the metrics behave in response to synthetically generated ranked results. They generate three sets representing documents with opinions that agree or disagree with the search term in different degrees (three degrees of agreement and three degrees of disagreement in addition to a neutral

degree). One of the sets contains balanced results in the sense that it has similar proportions for each of the different degrees of agreement while the other two sets are skewed towards the agreeing viewpoints. Rankings are then generated with varying degrees of bias (α) for 1000 runs and use the mean to aggregate the results.

[Draws et al., 2021] find that the metrics behaved similarly and returned a score corresponding to low unfairness for low bias value (α) and resulted in steeper curves when the protected and non-protected groups were similar in proportions suggesting it is easier to detect biases in that case. They also conclude that bias is harder to detect for populations where the protected group size is small and that the steep curve shape of Normalized Discounted Difference (Equation 5) makes it more suitable for cases where lower unfairness values are expected. Conversely, the parabolic shape of Normalized Discounted Ratio (Equation 7) makes it unsuitable for detecting unfairness at low values of bias. The results I present in the Chapter 5 show similar findings.

When providing the formal definitions of the metrics the authors note that these metrics are agnostic as to which of the two groups is advantaged in the ranking and that the metrics detect not only when the protected group is disadvantaged but also when the ranking is biased towards the protected group (symmetry of results). Which is not in line with the results of the empirical tests done in my research. Chapter 5 includes a concrete example that shows a lack of symmetry of results for two metrics defined in [Yang and Stoyanovich, 2017] and used in [Draws et al., 2021].

Although the authors of [Raj et al., 2020] discuss the metrics defined in [Yang and Stoyanovich, 2017] and provide their definitions, they decided to exclude them from the empirical analysis for what they describe as numerous edge case breakdowns. They only mention that these breakdowns relate to the proportions of the groups but provide no further explanation on their behaviour other than that none of them work when the unprotected group is empty and that Normalized Discounted Ratio (Equation 7) does not work when the unprotected group has too few members.

Chapter 3

Metrics

In this chapter I list the formulations of the metrics to be tested. At the end of the chapter in Table 3.1 an overview of these metrics and their symbols is added for easier reference.

Divergence-Based Metrics

These measures are defined in [Kırnap et al., 2021]. Their work focuses on metrics for group fairness using proportion and exposure based fairness definitions. The proportion based fairness metrics are defined as follows:

For $\mathcal{G}$ a set of group labels, P_g a target proportion for group g specified as input to the metric and $\tilde{\mathrm{P}}_g$ the actual proportion of group g in the given ranking the following fairness metrics are defined:

Difference:

$$\Delta_{\text{diff}} = \sum_{g \in \mathcal{G}} (\mathrm{P}_g - \tilde{\mathrm{P}}_g) \tag{1}$$

Absolute Difference:

$$\Delta_{\text{abs}} = \sum_{g \in \mathcal{G}} \Big| \mathrm{P}_g - \tilde{\mathrm{P}}_g \Big| \tag{2}$$

Squared Difference:

$$\Delta_{\text{sq}} = \sum_{g \in \mathcal{G}} (\mathrm{P}_g - \tilde{\mathrm{P}}_g)^2 \tag{3}$$

KL Divergence:

$$\Delta_{\text{KL}} = \sum_{g \in \mathcal{G}} \mathrm{P}_g \log\Big(\frac{\mathrm{P}_g}{\tilde{\mathrm{P}}_g}\Big) \tag{4}$$

Prefix Metrics

The metrics defined in [Yang and Stoyanovich, 2017] test statistical parity which is calculated at a series of cut-off points $[10, 20, ...]$ in the ranking and discounted such that statistical parity at the top positions has more weight than lower in the ranking. The discounting method used is similar to that of the Normalized Discounted Cumulative Gain (nDCG) framework [Järvelin and Kekäläinen, 2002]. This is done to express that it is for example more important to be fair at the top-10 positions than at the top-100.

Let τ be a ranking of length N, S^+ the set of items that are in the protected group, S^- the set of the items that are in the unprotected group such that $S^+ \cap S^- = \emptyset$. And $S^+_{1...i}$, $S^-_{1...i}$ the items in the ranking up to position i that belong to the protected and unprotected groups, respectively. The normalizer Z computed as the highest possible value of the metric, then:

Normalized Discounted Difference (rND)

$$\text{rND}(\tau) = \frac{1}{Z} \sum_{i=10,20,...}^{N} \frac{1}{log_2\, i} \left| \frac{|\, S^+_{1...i}\,|}{i} - \frac{|\, S^+\,|}{N} \right| \tag{5}$$

Normalized discounted KL-divergence (rKL)

$$\text{rKL}(\tau) = \frac{1}{Z} \sum_{i=10,20,...}^{N} \frac{D_{KL}(P||Q)}{log_2\, i} \tag{6}$$

Where

$$P = \left(\frac{|\, S^+_{1...i}\,|}{i}, \frac{|\, S^-_{1...i}\,|}{i} \right)$$
$$Q = \left(\frac{|\, S^+\,|}{N}, \frac{|\, S^-\,|}{N} \right)$$
$$D_{KL}(P||Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)} \right)$$

Normalized Discounted Ratio (rRD)

$$\text{rRD}(\tau) = \frac{1}{Z} \sum_{i=10,20,...}^{N} \frac{1}{log_2 i} \left| \frac{|\, S^+_{1...i}\,|}{|\, S^-_{1...i}\,|} - \frac{|\, S^+\,|}{|\, S^-\,|} \right| \tag{7}$$

Similar to [Yang and Stoyanovich, 2017], the work of [Geyik et al., 2019] defines a fairness metric based on proportions of groups in the top i positions of the ranking.

For ranking τ, normalizer $Z = \sum_{i=1}^{|\,\tau\,|} \frac{1}{\log_2(i+1)}$, D_{τ^i} a discrete distribution assigning to each group label $g \in \mathcal{G}$ the proportion of items belonging to that

group over τ^i and D the target distribution. Where τ^i are the top elements in ranking τ up to position i.

Normalized Discounted Cumulative KL-Divergence (NDKL)

$$NDKL(\tau) = \frac{1}{Z} \sum_{i=1}^{|\tau|} \frac{1}{\log_2(i+1)} d_{KL}(D_{\tau^i} || D) \tag{8}$$

Notice that although this metric is very similar to rKL, it is defined for an arbitrary number of group labels while rKL is defined for exactly two groups. They also differ in the normalizer and discount value.

Table 3.1: Overview of the different group proportion based metrics considered

Reference	Measure Name	Symbol	Number of groups	Target Proportion	Definition
[Kırnap et al., 2021]	Difference	Δ_{diff}	≥ 2	Input	(1)
	Absolute Difference	Δ_{abs}	≥ 2	Input	(2)
	Squared Difference	Δ_{sq}	≥ 2	Input	(3)
	KL-Divergence	Δ_{KL}	≥ 2	Input	(4)
[Yang and Stoyanovich, 2017]	Normalized Discounted Difference	rND	$= 2$	Derived	(5)
	Normalized Discounted KL-divergence	rKL	$= 2$	Derived	(6)
	Normalized Discounted Ratio	rRD	$= 2$	Derived	(7)
[Geyik et al., 2019]	Normalized Discounted Cumulative KL-Divergence	NDKL(τ)	≥ 2	Input	(8)

Chapter 4

Method

When testing a fairness metric, we could follow an intuitive approach where we assume that, when aggregated over enough iterations, randomness produces fairness. It is however not enough to only test against a random ranking to understand how metrics would behave in a real world scenario. We still need to know how it transitions between scores of maximum fairness and scores of maximum unfairness. We also expect that changes in the ranking will result in changes of similar magnitude to the fairness score in order for the fairness metric to be considered reliable. In addition to the importance of being aware of cases when a fairness metric fails to detect unfairness either because of the size of either group or other possibly problematic and less common situations.

The work of [Yang and Stoyanovich, 2017] and [Zehlike et al., 2017] is based on a fair ranking generated using the *Ranking Generator Algorithm* that takes two inputs, an initial ranking and a fairness probability $f \in [0, 1]$ indicating preference towards the protected group. The algorithm starts by separating the protected and unprotected members into two lists while maintaining the original relative ranking between members of the same group. The algorithm then proceeds to iteratively create a new ranking by placing top member from the protected group with probability f (or top member from the unprotected group with probability $1 - f$) until one of the lists is empty in which case the remaining members from the other list are added to the ranking. Examples of different rankings of size 10 that can be generated using this algorithm for different proportions and fairness probabilities are given in Figure 4.1. Notice that for fairness probability f = 0, all members of the unprotected group are placed at the top of the ranking followed by all the protected elements. Conversely, when the fairness probability f = 1 then all members of the protected group are placed at the top of the ranking followed by the unprotected members. For a fairness probability f = 0.5 and proportion of protected elements p = 0.7 the algorithm places members from each group at the top of the ranking with probability 0.5, in the given

example 3 members of each group are placed in alternating fashion at the top of the ranking, since no more members remain in the protected group, all remaining members of the unprotected group are added to the end of the ranking.

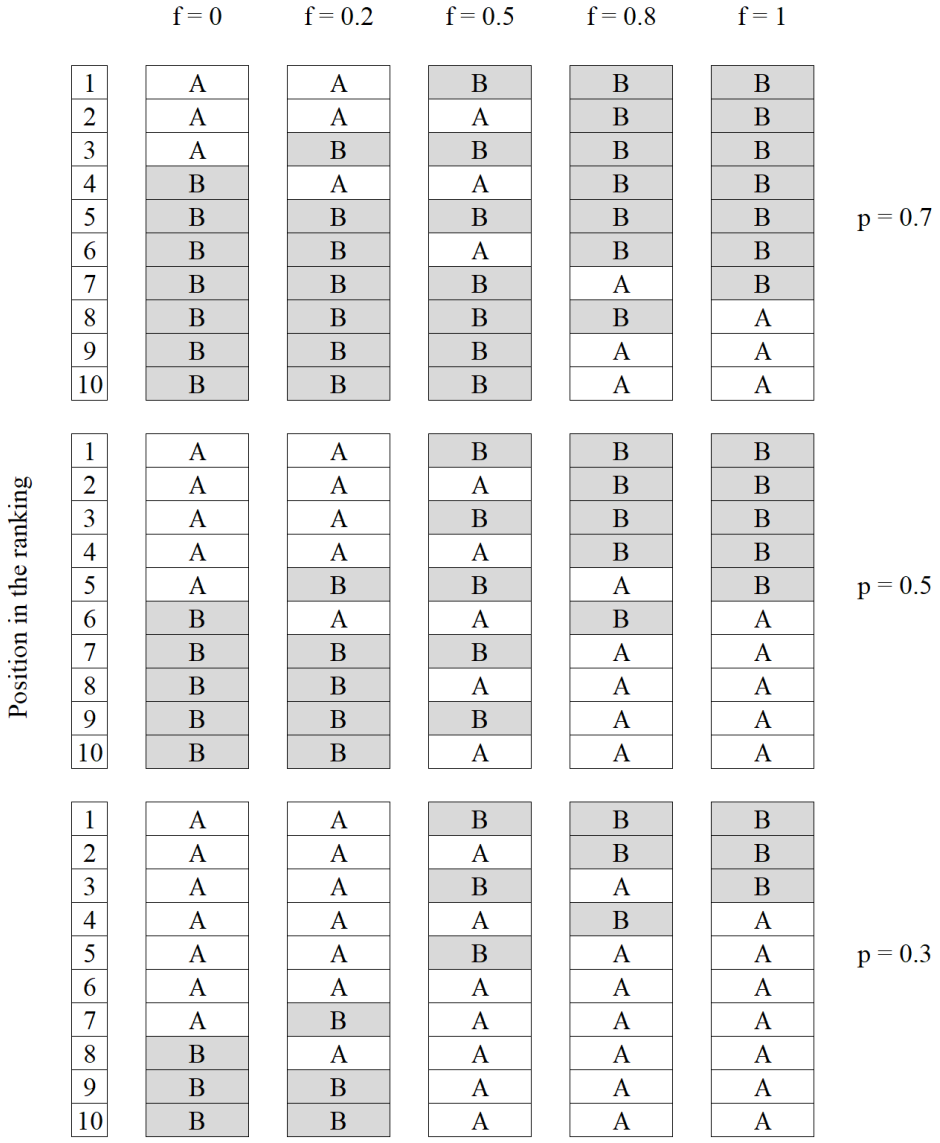

Figure 4.1: Examples of rankings of length 10 generated with the Ranking Generator Algorithm with group label A representing the unprotected group and label B representing the protected group for different fairness probability values f and different proportions of protected group in the ranking p

With this algorithm in mind, we create the tests defined below while limiting the length of the ranking to 100 positions:

- Test 1: The initial ranking constitutes of only members of the unprotected group. Compute the fairness score. Next, for $i \in [1, 100]$ create a ranking where the item at position i is a member of the protected group and all other items are from the unprotected group. Compute fairness at each step.

- Test 2: The initial ranking constitutes of only members of the unprotected group. Compute the fairness score. Next, for $i \in [1, 100]$ create a ranking using the Ranking Generator Algorithm with i members of the protected group and $100 - i$ members of the unprotected group and fairness probability $f = 1$. Compute fairness at each step.

- Test 3: The initial ranking constitutes of only members of the unprotected group. Compute the fairness score. Next, for $i \in [1, 100]$ create a ranking using the Ranking Generator Algorithm with i members of the protected group and $100 - i$ members of the unprotected group and fairness probability $f = 0$. Compute fairness at each step.

- Test 4: The initial ranking constitutes of only members of the unprotected group. Compute the fairness score. Next, for $i \in [1, 100]$ create a ranking using the Ranking Generator Algorithm with i members of the protected group and $100 - i$ members of the unprotected group and fairness probability $f = 0.5$. Compute fairness at each step.

- Test 5: The initial ranking constitutes of only members of the unprotected group. Compute the fairness score. Next, for $i \in [1, 100]$ create a ranking with i members of the protected group and $100 - i$ members of the unprotected group in random positions. Compute fairness at each step. Repeat for 100 runs and aggregate the results.

Chapter 5

Results

Divergence-Based Metrics

The metrics defined in [Kırnap et al., 2021] seem to have the most limitations in use among the metrics tested. Figure 5.1 shows the results of Test 1 for these metrics. The fairness score does not change when the position of the protected item is changed as the metrics only consider the proportions of the group and not the positions of the items. The results for Test 1 only depend on the difference between proportion of protected group in the ranking (this is always 0.01 for this test) and the target proportion of protected group (0.3).

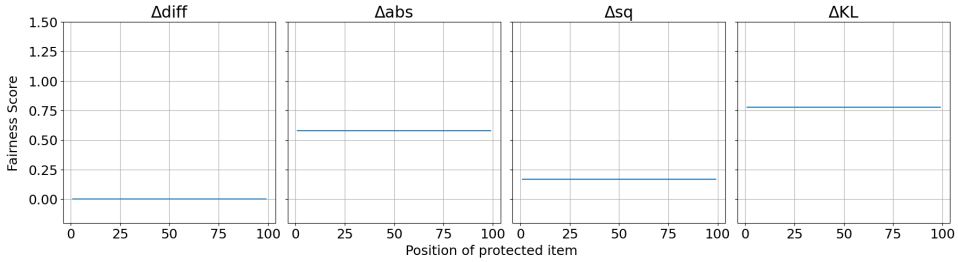

Figure 5.1: Results of Test 1 for divergence-based metrics defined in [Kırnap et al., 2021] for target proportion of the protected group 0.3 where a single item in the protected group is assigned a different position in the ranking at each step.

In Figure 5.2 the results for Test 2 are shown. The results of tests 3 through 5 are identical to this test since they only differ in the positions of protected and unprotected items at each step but not the proportions. Changing the positions in the ranking while maintaining the same proportions has no effect on the fairness score as seen in the results of Test 1. Although Δ_{diff} was defined in terms of proportion-based representations it always evaluated to 0 in all tests. This is because of the way it is calculated based on a simple summation of the differences between target and actual proportions. For the sake of completeness, I include a simplified proof of the

metric always evaluating to 0 for any given group labels, target and actual proportions. The authors mention that the definitions for exposure-based representations follow analogously from these definitions and only point out the limitation of proportion-based divergence metrics in the case of using the TREC Fair Ranking Dataset, where the proportions of different groups in the dataset are identical and therefore do not report test results on that dataset. We only expect this measure to evaluate to a value other than 0 if the proportions of the groups do not add up to 1.

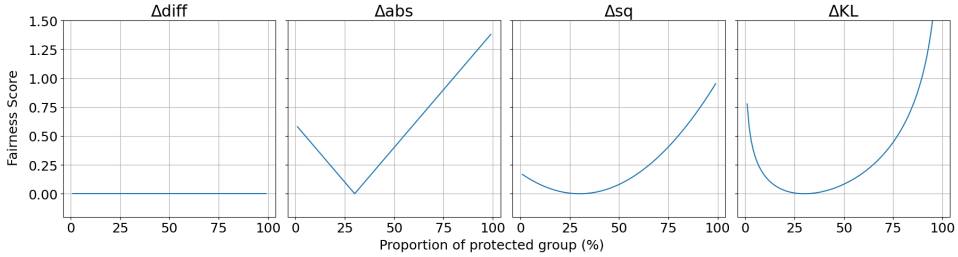

Figure 5.2: Test 2 results of divergence-based metrics defined in [Kırnap et al., 2021] for target proportion of the protected group p=0.3 These metrics do not account for different distributions of group members in the ranking. Instead, they compare the proportion in the ranking to the given target proportion resulting in identical results for tests 2 through 5

Proof Difference metric always evaluates to 0:
Let

- τ be an arbitrary ranking
- $\mathcal{G} = \{0, 1, 2, ..., n\}$ be set of group labels
- $\mathbf{P} = \{x_0, x_1, ..., x_n\}$ is the target proportion distribution where x_i is the target proportion of group i for $i \in \mathcal{G}$ and $x_0 + x_1 + ... + x_n = 1$
- $\tilde{\mathbf{P}} = \{y_0, y_1, ..., y_n\}$ is the actual proportion distribution where y_i is the actual proportion of group i in τ for $i \in \mathcal{G}$ and $y_0 + y_1 + ... + y_n = 1$

Then

$$\begin{aligned}\Delta_{\text{diff}} &= \sum_{i \in \mathcal{G}} (x_i - y_i) \\ &= (x_0 - y_0) \\ &+ (x_1 - y_1) \\ &+ (x_2 - y_2) \\ &\vdots \\ &+ (x_n - y_n)\end{aligned}$$

which by reordering the terms (addition is commutative) can be rewritten as:

$$\begin{aligned}\Delta_{\text{diff}} &= x_0 + x_1 + ... + x_n - y_1 - y_2 - .... - y_n \\ &= (x_0 + x_1 + ... + x_n) - (y_1 + y_2 + ... + y_n)\end{aligned}$$

But since we know that $x_0 + x_1 + ... + x_n = 1$ and $y_0 + y_1 + ... + y_n = 1$ Then $\Delta_{\text{diff}} = 1 - 1 = 0$ for any $\tau, \mathcal{G}, \mathbf{P}, \tilde{\mathbf{P}}$,

Prefix Metrics

rKL and rND had similar behaviour for Test 1 (Figure 5.3), but were surprisingly very different than the result of rRD where placing a protected element among the top 10 positions was considered maximally unfair by rKL and rND but maximally fair for rRD.

In the case of NDKL, the results were similar in trend to those in rKL and rND. Namely, starting at higher unfairness scores for positions at the top of the ranking and quickly dropping to lower unfairness scores. The scores for NDKL are generally lower that those of the other metrics in this section. All the other metrics state that the returned value is in the range [0, 1], while NDKL is defined only to be a non-negative value where a larger value denotes a higher degree of unfairness. It also has smoother transitions between the scores, which is a result of the smaller cut-off points in NDKL.

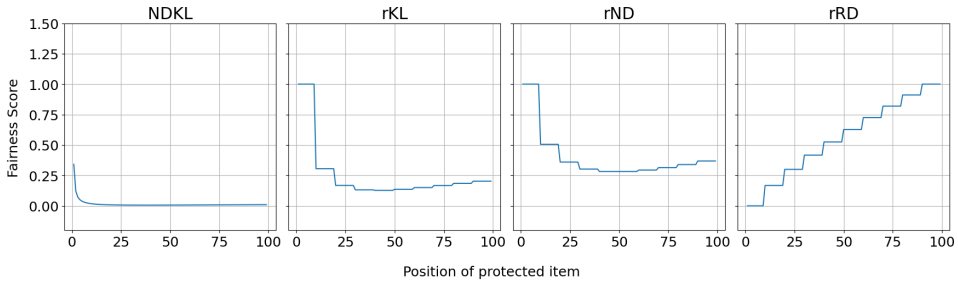

Figure 5.3: Results of Test 1 for prefix metrics where a single item in the protected group is assigned a different position in the ranking at each step.

We can use results from Test 2 and Test 3 to show that the note in [Draws et al., 2021] about metrics defined in [Yang and Stoyanovich, 2017] being agnostic of the semantics of protected and unprotected group label does not hold for all metrics and that they do not treat bias *towards* the protected group similar to bias *against* it. Since that would suggest that for a given distribution of two groups in a ranking, switching between the labels protected and unprotected would result in the same fairness score.

Therefore, we are looking for are two rankings that are similar in proportions and positions but are opposite in group labels. For example, in

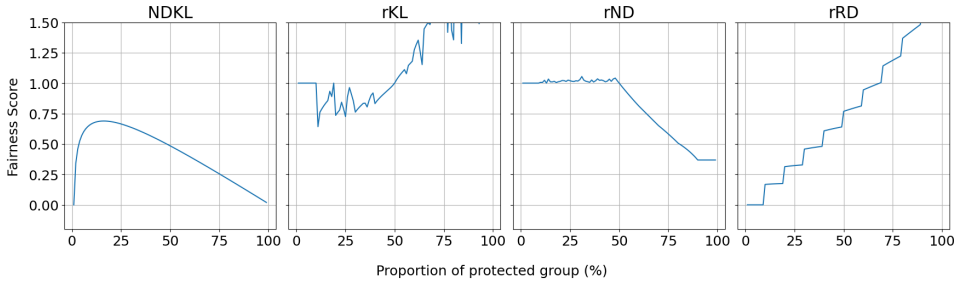

Figure 5.4: Test 2 results of Prefix Metrics where rankings with increasing proportion of protected group are generated with fairness probability f = 1

Figure 4.1 we look at the ranking corresponding to f = 0, p = 0.7 and the ranking for f = 1, p = 0.3

We compare the fairness score for rRD in Test 2 (Figure 5.4) when proportion of the protected group is 5% when all members of the protected group are placed at the top of the ranking (therefore, unprotected group has 95% proportion and is placed at the end of the ranking), we compare this value to rRD in Test 3 (Figure 5.5) when the proportion of the protected group is 95% all members of which are placed at the end of the ranking (therefore, the unprotected group has 5% proportion and is placed at the top of the ranking). The only difference between the two cases is switching the group label from protected to unprotected and vice-versa but it made the difference between maximally fair and maximally unfair scores of the same metric. Similarly, rKL does not show symmetry of results.

However, the results for rND do go in line with that note, this is reflected in the line symmetry in results of Test 4 5.6 around the line $x = 50\%$ and the symmetry between results of Test 2 and Test 3 of that metric.

Although the results of NDKL have not shown exact symmetry, they do show similarities between biases towards and against the protected group.

An issue that arose when interpreting the results of rKL was that it differed greatly on each run of the code. When analysing the source code that is originally provided by the authors of [Yang and Stoyanovich, 2017], the issue seems to be a result of calculating the normalizer for rKL based on a maximum value over a random sample of fixed size. The same issue in this metric was also mentioned in [Draws et al., 2021]

The results of Test 5 (Figure 5.7) show the difference between the behaviour of NDKL when compared to rKL, rND and rRD when there is a big difference in proportions between the protected group and the unprotected group. NDKL returns a lower unfairness score for these cases while the other three metrics return higher unfairness scores.

All three metrics returned a low unfairness score for the random rankings (Test 5) except when the proportions of protected and unprotected groups

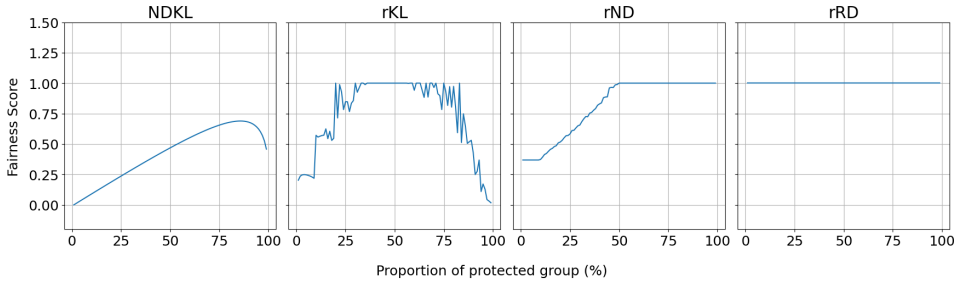

Figure 5.5: Test 3 results of Prefix Metrics where rankings with increasing proportion of protected group are generated with fairness probability f = 0

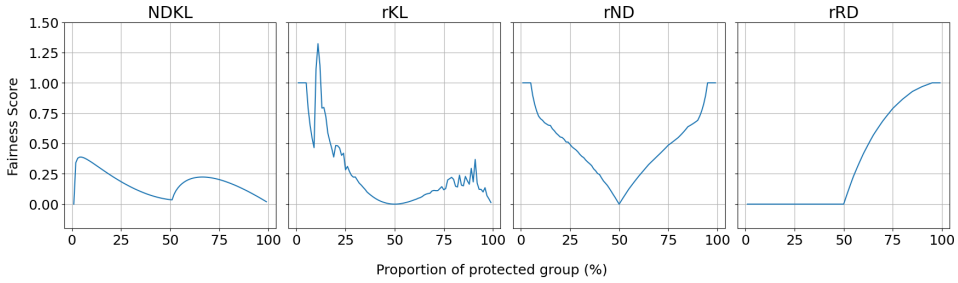

Figure 5.6: Test 4 results of Prefix Metrics where rankings with increasing proportion of protected group are generated with fairness probability f = 0.5

are very different which is generally inline with our intuition of randomness leading to fairness. However, both rND and rRD had indicated slightly higher unfairness than that indicated by rKL and we see greater local maxima around the cut-points 10 and 20 for rKL than the other metrics.

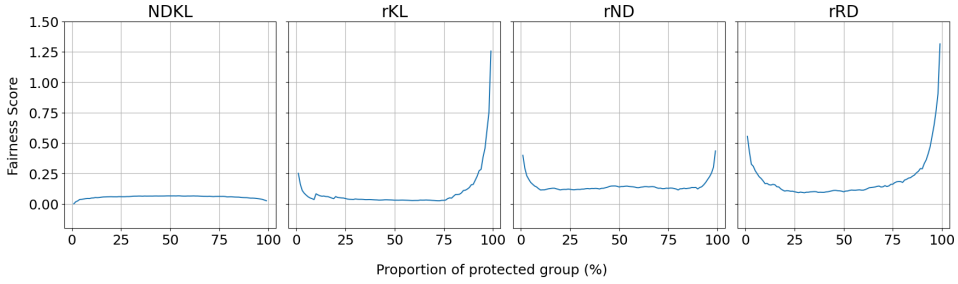

Figure 5.7: Test 5 results of Prefix Metrics average fairness scores over 100 randomly generated rankings with increasing proportion of protected group

The findings of the previous results can be summarized as follows:

- The tests results show that Divergence-Based Metrics do not express higher positions in the ranking being more beneficial and their use is limited to cases where fairness is only defined based on the proportions of different groups in the ranking and not their positions.
- Difference metric always evaluates to 0 for any given ranking and proportion distributions under the assumption that proportion distributions add up to 1.
- rND shows near perfect symmetry of results allowing for detecting bias against, as well as towards, the protected group. rRD and rKL showed a clear lack of symmetry making their use only preferred when ensuring fairness towards the protected group is more important and cases where unfairness towards the unprotected group is unlikely to occur in the output.
- All Prefix Metrics returned on average a low unfairness score for randomly generated rankings except for cases when the difference in proportions between protected and unprotected groups was high. In that case, NDKL is preferred as it was more stable around these edge cases.
- The results of rKL were often unreliable due to the way normalizer Z is calculated for that metric by the original authors. As kullback-Leibler Divergence value is in the range $[0, \infty)$. The normalizer is estimated based on random samples that were different every run of the code resulting in different unfairness values for the same rankings.

Chapter 6

Conclusion

Concerns about the fairness in the results of ranking systems led to the introduction of various novel fairness-aware algorithms. The fairness of these algorithms is often measured and optimized against a novel metric that is introduced together with the algorithm, making it difficult to compare the performance of different metrics. In this research I compared group fairness metrics that aim to measure statistical parity in the ranking. The aim of this research is to study the behaviour of these metrics when the input ranking is changed in a controlled manner and identify cases when the metrics have a similar or dissimilar behaviour over the same ranking.

I find that although the metrics aim to satisfy a similar fairness goal, they still behaved differently when tested against a common ranking. These differences seem to result from different interpretations of statistical parity in ranked outputs. The main factors in introducing differences in behaviour include: 1) Whether or not the metric differentiates between a protected and an unprotected group. This has an effect on whether the metric is able to detect bias against either group the same way or is better suited for cases when the goal is ensuring fairness towards a specific group. This can be tested by comparing fairness scores of two rankings that are identical in item positions but opposite in group labels. 2) The source of the target proportion distribution. In the case of divergence based fairness metrics the target proportion was given as input to the metric which is assumed to be fair. On the other hand, prefix metrics consider the proportion distribution in the whole ranking to be fair and test whether prefixes of the ranking have the same proportion distribution.

I note that the Difference metric always evaluates to 0 under the assumption that the proportions of the groups add up to 1 which greatly limits its usefulness. I also find that rND should be preferred when measuring fairness towards, as well as against a protected group, while NDKL is more suitable for cases when the difference in proportion between the protected and unprotected group is high. In the case of rKL, the issue with how the

normalizer is computed must be addressed before it can be reliably used to measure fairness.

The results of this research provide more insight into the behaviour of the chosen fairness metrics and can help decide whether or not a fairness metric is suitable in assessing fairness in a given ranking. Future work can compare fairness metrics that share similar fairness assumptions including protected group fairness and the target proportion distribution. It can also address the issues that result in using KL-divergence as a method to assess statistical parity of proportion distributions including that it has an unbounded maximum value, making it challenging to normalize the results into a fixed range.

References

[Binns, 2020] Binns, R. (2020). On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 514–524.

[Borghol et al., 2012] Borghol, Y., Ardon, S., Carlsson, N., Eager, D., and Mahanti, A. (2012). The untold story of the clones: Content-agnostic factors that impact youtube video popularity. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1186–1194.

[Chen et al., 2020] Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., and He, X. (2020). Bias and debias in recommender system: A survey and future directions. *arXiv preprint arXiv:2010.03240*.

[Dastin, 2018] Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women. `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G`. Date accessed: February 2022.

[Draws et al., 2021] Draws, T., Tintarev, N., and Gadiraju, U. (2021). Assessing viewpoint diversity in search results using ranking fairness metrics. *ACM SIGKDD Explorations Newsletter*, 23(1):50–58.

[Dwork et al., 2012] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

[Friedler et al., 2019] Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338.

[Garg et al., 2020] Garg, P., Villasenor, J., and Foggo, V. (2020). Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3662–3666. IEEE.

[Geyik et al., 2019] Geyik, S. C., Ambler, S., and Kenthapadi, K. (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*, pages 2221–2231.

[Järvelin and Kekäläinen, 2002] Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

[Kay et al., 2015] Kay, M., Matuszek, C., and Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, pages 3819–3828.

[Kırnap et al., 2021] Kırnap, Ö., Diaz, F., Biega, A., Ekstrand, M., Carterette, B., and Yilmaz, E. (2021). Estimation of fair ranking metrics with incomplete judgments. In *Proceedings of the Web Conference 2021*, pages 1065–1075.

[Mehrabi et al., 2021] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

[Raj et al., 2020] Raj, A., Wood, C., Montoly, A., and Ekstrand, M. D. (2020). Comparing fair ranking metrics. *arXiv preprint arXiv:2009.01311*.

[Srivastava et al., 2019] Srivastava, M., Heidari, H., and Krause, A. (2019). Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, pages 2459–2468.

[van Bekkum and Borgesius, 2021] van Bekkum, M. and Borgesius, F. Z. (2021). Digital welfare fraud detection and the dutch syri judgment. *European Journal of Social Security*, 23(4):323–340.

[Yang and Stoyanovich, 2017] Yang, K. and Stoyanovich, J. (2017). Measuring fairness in ranked outputs. In *Proceedings of the 29th international conference on scientific and statistical database management*, pages 1–6.

[Zehlike et al., 2017] Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., and Baeza-Yates, R. (2017). Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578.