Causal Intersectionality and Fair Ranking

Ke Yang
New York University, NY, USA

Joshua R. Loftus
London School of Economics, UK

Julia Stoyanovich
New York University, NY, USA

Abstract

In this paper we propose a causal modeling approach to intersectional fairness, and a flexible, task-specific method for computing intersectionally fair rankings. Rankings are used in many contexts, ranging from Web search to college admissions, but causal inference for fair rankings has received limited attention. Additionally, the growing literature on causal fairness has directed little attention to intersectionality. By bringing these issues together in a formal causal framework we make the application of intersectionality in algorithmic fairness explicit, connected to important real world effects and domain knowledge, and transparent about technical limitations. We experimentally evaluate our approach on real and synthetic datasets, exploring its behavior under different structural assumptions.

2012 ACM Subject Classification Computing methodologies → Ranking

Keywords and phrases fairness, intersectionality, ranking, causality

Digital Object Identifier 10.4230/LIPIcs.FORC.2021.7

Related Version *Extended Version*: `https://arxiv.org/abs/2006.08688`

Supplementary Material *Software (Source Code)*: `https://github.com/DataResponsibly/CIFRank` archived at `swh:1:dir:d16d7e476bfacf0d8c562f2f96b0ead488ad7138`

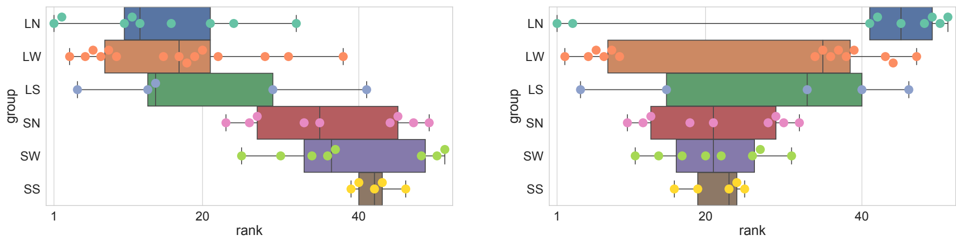
Funding This work was supported in part by National Science Foundation (NSF) Grants No. 1926250, 1934464, and 1916505.

1 Introduction

The machine learning community recognizes several important normative dimensions of information technology including privacy, transparency, and fairness. In this paper we focus on fairness – a broad and inherently interdisciplinary topic of which the social and philosophical foundations are not settled [11]. To connect to these foundations, we take an approach based on *causal modeling*. We assume that a suitable causal generative model is available and specifies relationships between variables including the *sensitive attributes*, which define individual traits or social group memberships relevant for fairness. The model is a statement about how the world works, and we define fairness based on the model itself. In addition to being philosophically well-motivated and explicitly surfacing normative assumptions, the connection to causality gives us access to a growing literature on causal methods in general and causal fairness in particular.

Research on algorithmic fairness has mainly focused on classification and prediction tasks, while we focus on ranking. We consider two types of ranking tasks: score-based and learning to rank (LTR). In score-based ranking, a given set of candidates is sorted on the score attribute (which may itself be computed on the fly) and returned in sorted order. In LTR, supervised learning is used to predict the ranking of unseen items. In both cases, we typically return the highest scoring k items, the top-k. Set selection is a special case of ranking that ignores the relative order among the top-k.

2nd Symposium on Foundations of Responsible Computing (FORC 2021).
Editors: Katrina Ligett and Swati Gupta; Article No. 7; pp. 7:1–7:20
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany



(a) original ranking. **(b)** counterfactually fair.

Figure 1 CSRanking by weighted publication count, showing positions of intersectional groups by department size, large (L) and small (S), and location, North East (N), West (W), South East (S). Observe that the top-20 in Figure 1a is dominated by large departments, particularly those from the West and from the North East. Treating small departments from the South East as the disadvantaged intersectional group, and applying the techniques described in Section 2 of the paper, we derive the ranking in Figure 1b that has more small department at the top-20 and is more geographically balanced.

Further, previous research mostly considered a single sensitive attribute, while we use multiple sensitive attributes for the fairness component. As noted by Crenshaw [14], it is possible to give the appearance of being fair with respect to each sensitive attribute such as race and gender separately, while being unfair with respect to *intersectional* subgroups. For example, if fairness is taken to mean proportional representation among the top-k, it is possible to achieve proportionality for each gender subgroup (e.g., men and women) and for each racial subgroup (e.g., Black and White), while still having inadequate representation for a subgroup defined by the intersection of both attributes (e.g., Black women). The literature on intersectionality includes theoretical and empirical work showing that people adversely impacted by more than one form of structural oppression face additional challenges in ways that are more than additive [12, 16, 37, 43].

1.1 Contribution

We define intersectional fairness for ranking in a similar manner to previous causal definitions of fairness for classification or prediction tasks [10, 26, 30, 36, 55]. The idea is to model the causal effects between sensitive attributes and other variables, and then make algorithms fairer by removing these effects. With a given ranking task, set of sensitive attributes, and causal model, we propose ranking on counterfactual scores as a method to achieve intersectional fairness. From the causal model we compute model-based counterfactuals to answer a motivating question like "What would this person's data look like if they had (or had not) been a Black woman (for example)?" We compute counterfactual scores *treating every individual in the sample as though they had belonged to one specific, baseline intersectional subgroup*. For score-based ranking we then rank these counterfactual scores, but the same approach to causal intersectional fairness can be combined with other machine learning tasks, including prediction (not necessarily specific to ranking).

The choice of a baseline counterfactual subgroup is essentially arbitrary, and there are other possibilities like randomizing or averaging over all subgroups. We focus on using one subgroup now for simplicity, but in principle this choice can depend on problem specifics and future work can investigate dependence on this choice. In fact, our framework allows for numeric sensitive attributes, like age for example, where treating everyone according to one baseline counterfactual is possible even though subgroup terminology breaks down. In this case we can still try to rank every individual based on an answer to a motivating question like "What would this person's data look like if they were a 45-year old Black woman?"

While intersectional concerns are usually raised when data is about people, they also apply for other types of entities. Figure 1 gives a preview of our method on the CSRankings dataset [5] that ranks 51 computer science departments in the US by a weighted publication count score (lower ranks are better). Departments are of two sizes, large (L, with more than 30 faculty members) and small (S), and are located in three geographic areas, North East (N), West (W), and South East (S). The original ranking in Figure 1a prioritizes large departments, particularly those in the North East and in the West. The ranking in Figure 1b was derived using our method, treating small departments from the South East as the disadvantaged intersectional group; it includes small departments at the top-20 and is more geographically balanced.

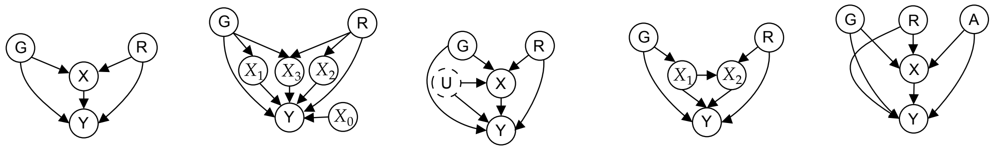
We begin with relatively simple examples to motivate our ideas before considering more complex ones. The framework we propose can, under the right conditions, disentangle multiple interlocked "bundles of sticks," to use the metaphor in Sen and Wasow [42] for causally interpreting sensitive attributes that may be considered immutable. We see this as an important step towards a more nuanced application of causal modeling to fairness.

1.2 Motivating example: Hiring by a moving company

Consider an idealized hiring process of a moving company, inspired by Datta et al. [15], in which a dataset of applicants includes their gender G, race R, weight-lifting ability score X, and overall qualification score Y. A ranking of applicants $\boldsymbol{\tau}$ sorts them in descending order of Y. We assume that the structural causal model shown in Figure 2a describes the data generation process, and our goal is to use this model to produce a ranking that is fair with respect to race, gender, and the intersectional subgroups of these categories. The arrows in the graph pointing from G and R directly to Y represent the effect of "direct" discrimination. Under US labor law, the moving company may be able to make a "business necessity" argument [17] that they are not responsible for any "indirect" discrimination on the basis of the mediating variable X. If discrimination on the basis of X is considered unenforceable, we refer to X as a *resolving mediator*, and denote this case as the resolving case, following the terminology of Kilbertus et al. [26].

A mediator X may be considered resolving or not; this decision can be made separately for different sensitive attributes, and the relative strengths of causal influences of sensitive attributes on both X and Y can vary, creating potential for explanatory nuance even in this simple example. Suppose that X is causally influenced by G but not by R, or that the relative strength of the effect of G on X is larger than that of R. Then, if X is considered resolving, the goal is to remove direct discrimination on the basis of both R and G, but hiring rates might still differ between gender groups if that difference is explained by each individual's value of X. On the other hand, if X is not considered resolving, then the goal also includes removing indirect discrimination through X, which, in addition to removing direct discrimination, might accomplish positive discrimination, in the style of affirmative, action based on the effect of G on X.

Once the goal has been decided, we use the causal model to compute counterfactual scores Y – the scores that would have been assigned to the individuals if they belonged to one particular subgroup defined by fixed values of R and G, while holding the weight-lifting score X fixed in the resolving case – and then rank the candidates based on these scores. The moving company can then interview or hire the highly ranked candidates, and this process would satisfy a causal and intersectional definition of fairness. We analyze a synthetic dataset based on this example in Section 3 with results shown in Figure 3a.



(a) $\mathcal{M}_1$ **(b)** $\mathcal{M}_2$ **(c)** $\mathcal{M}_3$ **(d)** $\mathcal{M}_4$ **(e)** $\mathcal{M}_5$

Figure 2 Causal models that include sensitive attributes G (gender), R (race), and A (age), utility score Y, other covariates $\mathbf{X}$, and a latent (unobserved) variable U.

1.3 Organization of the paper

In Section 2 we introduce notation and describe the particular causal modeling approach we take, using directed acyclic graphs and structural equations, but we also note that our higher level ideas can be applied with other approaches to causal modeling. We present the necessary modeling complexity required for interaction effects in the causal model, the process of computing counterfactuals for both the resolving and non-resolving cases, and the formal fairness definition that our process aims to satisfy. In Section 3 we demonstrate the effectiveness of our method on real and synthetic dataset. We present a non-technical interpretation of our method, and discuss its limitations, in Section 4. We summarize related work in Section 5 and conclude in Section 6. Our code is publicly available at `https://github.com/DataResponsibly/CIFRank`.

2 Causal intersectionality

In this section we describe the problem setting, and present our proposed definition of intersectional fairness within causal models and an approach to computing rankings satisfying the fairness criterion.

2.1 Model and problem setting

2.1.1 Causal model

As an input, our method requires a structural causal model (SCM), which we define briefly here and refer to [23, 33, 39, 44] for more detail. An SCM consists of a directed acyclic graph (DAG) $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, where the vertex set $\mathbf{V}$ represents variables, which may be observed or latent, and the edge set $\mathbf{E}$ indicates causal relationships from source vertices to target vertices. Several example DAGs are shown in Figure 2, where vertices with dashed circles indicate latent variables.

For $V_j \in \mathbf{V}$ let $\text{pa}_j = \text{pa}(V_j) \subseteq \mathbf{V}$ be the "parent" set of all vertices with a directed edge into V_j. If pa_j is empty, we say that V_j is exogenous, and otherwise we assume that there is a function $f_j(\text{pa}_j)$ that approximates the expectation or some other link function, such as the logit, of V_j. Depending on background knowledge or the level of assumptions we are willing to hazard, we assume that functions f_j are either known or can be estimated from the data. We also assume a set of sensitive attributes $\mathbf{A} \subseteq \mathbf{V}$, chosen *a priori*, for which existing legal, ethical, or social norms suggest that the ranking algorithm should be fair.

2.1.2 Problem setting

In most of our examples we consider two sensitive attributes, which we denote G and R, motivated by the example of Crenshaw [14] of gender and race. We let Y denote an outcome variable that is used as a utility score in our ranking task, and $\mathbf{X}$ be *a priori* non-sensitive predictor variables. In examples with pathways from sensitive attributes to Y passing through $\mathbf{X}$ we call the affected variables in $\mathbf{X}$ mediators. Finally, U may denote an unobserved confounder. In some settings a mediator may be considered *a priori* to be a legitimate basis for decisions even if it results in disparities. This is what Foulds et al. [18] call the infra-marginality principle, others [10, 30, 36] refer to as path-specific effects, and Zhang and Bareinboim [55] refer to as indirect effects; Kilbertus et al. [26] call such mediators *resolving variables*. We adopt the latter terminology and will show examples of different cases later. In fact, our method allows mediators to be resolving for one sensitive attribute and not for the other, reflecting nuances that may be necessary in intersectional problems.

For simplicity of presentation, we treat some sensitive attributes as binary indicators of a particular privileged status, rather than using a more fine grained coding of identity, but note that this is not a necessary limitation of the method. Our experiments in Section 3 use models $\mathcal{M}_1$ in Figure 2a and $\mathcal{M}_5$ in Figure 2e, but richer datasets and other complex scenarios such as $\mathcal{M}_2$ also fit into our framework. *Sequential ignorability* [21, 38, 40, 47] is a standard assumption for model identifiability that can be violated by unobserved confounding between a mediator and an outcome, as in $\mathcal{M}_3$ in Figure 2c, or by observed confounding where one mediator is a cause of another, as in $\mathcal{M}_4$ in Figure 2d. We include these as indications of qualitative limitations of this framework.

2.2 Counterfactual intersectional fairness

2.2.1 Intersectionality

It is common in predictive modeling to assume a function class that is linear or additive in the inputs, that is, for a given non-sensitive variable V_j:

$$f_j(\mathrm{pa}_j) = \sum_{V_l \in \mathrm{pa}_j} f_{j,l}(V_l).$$

Such simple models may be less likely to overfit and are more interpretable. However, to model the intersectional effect of multiple sensitive attributes *we must avoid this assumption*. Instead, we generally assume that f_j contains non-additive interactions between sensitive attributes. With rich enough data, such non-linear f_j can be modeled flexibly, but to keep some simplicity in our examples we will consider functions with linear main effects and second order interactions. That is, if the set pa_j of parents of V_j includes q sensitive attributes $A_{j_1}, A_{j_2}, \ldots, A_{j_q}$ and p non-sensitive attributes $X_{j_{q+1}}, X_{j_{q+2}}, \ldots X_{j_{q+p}}$, we assume

$$f_j(\mathrm{pa}_j) = \beta_0^{(j)} + \sum_{l=1}^{p} \beta_l^{(j)} X_{j_{q+l}} + \sum_{l=1}^{q} \eta_l^{(j)} A_{j_l} + \sum_{l=1}^{q-1} \sum_{r=l+1}^{q} \eta_{r,l}^{(j)} A_{j_l} A_{j_r}. \tag{1}$$

The coefficients (or weights) $\eta_l^{(j)}$ model the main causal effect on V_j of disadvantage on the basis of sensitive attribute A_{j_l}, while $\eta_{r,l}^{(j)}$ model the non-additive combination of adversity related to the interactions of A_{j_r} and A_{j_l}. For the example the model $\mathcal{M}_1$ in Figure 2a with sensitive attributes G and R, mediator X, and outcome Y, we can write (1) for Y as

$$f_Y(X, G, R) = \beta_0^{(Y)} + \beta_1^{(Y)} X + \eta_G^{(Y)} G + \eta_R^{(Y)} R + \eta_{R,G}^{(Y)} RG \tag{2}$$

For ease of exposition we mostly focus on categorical sensitive attributes, and in that case (1) can be reparameterized with a single sensitive attribute with categories for each intersectional subgroup. In the simplest cases then it may appear this mathematical approach to intersectional fairness reduces to previously considered fairness problems. However, our framework is not limited to the simplest cases. And even with two binary sensitive attributes it may be necessary to model the separate causal relationships between each of these and one or more mediators, which may also be considered resolving or non-resolving separately with respect to each sensitive attribute. With numeric attributes our framework can include non-linear main effects and higher order interactions, and in Appendix A.2 we present results for an experiment with a numeric sensitive attribute.

Our experiments use simpler examples with one mediator so the results are easier to interpret and compare to non-causal notions of fairness in ranking. Sophisticated models like Figure 2b, with combinations of resolving and non-resolving mediators, would be more difficult to compare to other approaches, but we believe this reflects that real-world intersectionality can pose hard problems that our framework is capable of analyzing. And while identifiability and estimation are simplified in binary examples, the growing literature on causal mediation discussed in Section 5 can be used on harder problems.

2.2.2 Counterfactuals

Letting $\mathbf{A}$ denote the vector of sensitive attributes and $\mathbf{a}'$ any possible value for these, we compute the counterfactual $Y_{\mathbf{A} \leftarrow \mathbf{a}'}$ by replacing the observed value of $\mathbf{A}$ with $\mathbf{a}'$ and then propagating this change through the DAG: any directed descendant V_j of $\mathbf{A}$ has its value changed by computing $f_j(\mathrm{pa}_j)$ with the new value of $\mathbf{a}'$, and this operation is iterated until it reaches all the terminal nodes that are descendants of any of the sensitive attributes $\mathbf{A}$. We interpret these model-based counterfactuals informally as "the value Y would have taken if $\mathbf{A}$ had been equal to $\mathbf{a}'$."

For graphs with resolving mediators we may keep the mediator fixed while computing counterfactuals. We describe this process in detail for model $\mathcal{M}_1$ in Figure 2a, with both the resolving and the non-resolving cases. We focus on this model for clarity, but all that we say in the rest of this section requires only minor changes to hold for other models such as $\mathcal{M}_2$ without loss of generality, provided they satisfy sequential ignorability [21, 38, 40, 47]. Our implementation is similar to what Kusner et al. [30] refer to as "Level 3" assumptions, but we denote exogenous error terms as ϵ instead of U.

We consider the case where Y is numeric and errors are additive

$$X = f_X(G, R) + \epsilon^X, \quad Y = f_Y(X, G, R) + \epsilon^Y.$$

with f_Y given in (2) and f_X defined similarly. The case where Y is not continuous fits in the present framework with minor modifications, where we have instead a probability model with corresponding link function g so that

$$\mathbb{E}[Y|X, G, R] = g^{-1}(f_Y(X, G, R)).$$

Suppose that the observed values for observation i are (y_i, x_i, g_i, r_i), with exogenous errors $\epsilon_i^X, \epsilon_i^Y$. Since we do not model any unobserved confounders in model $\mathcal{M}_1$, we suppress the notation for U and denote counterfactual scores, for some $(g', r') \neq (g, r)$, as:

$$Y_i' := (Y_i)_{\mathbf{A} \leftarrow \mathbf{a}'} = (Y_i)_{(G,R) \leftarrow (g', r')}.$$

If X is **non-resolving**, then we first compute counterfactual X as $x_i' := f_X(g', r') + \epsilon_i^X$, substituting (g', r') in place of the observed (g_i, r_i). Then we do the same substitution while computing:

$$Y_i' = f_Y(x_i', g', r') + \epsilon_i^Y = f_Y(f_X(g', r') + \epsilon_i^X, g', r') + \epsilon_i^Y.$$

If X is **resolving**, then we keep the observed X and compute:

$$Y_i' = f_Y(x_i, g', r') + \epsilon_i^Y.$$

If X is **semi-resolving**, for example resolving for R but not for G, in which case we compute counterfactual X as $x_i' := f_X(g', r_i) + \epsilon_i^X$ and then

$$Y_i' = f_Y(f_X(g', r_i) + \epsilon_i^X, g', r') + \epsilon_i^Y.$$

If the functions f_X, f_Y have been estimated from the data, then we have observed residuals r_i^X, r_i^Y instead of model errors in the above. Finally, in cases where we model unobserved confounders U we may also attempt to integrate over the estimated distribution of U as described in [30].

2.3 Counterfactually fair ranking

2.3.1 Ranking task

We use an outcome or utility score Y to rank a dataset $\mathbf{D}$, assumed to be generated by a model $\mathcal{M}$ from among the example SCMs in Figure 2. If the data contains a mediating predictor variable X, then the task also requires specification of the resolving status of X. Letting $n = |\mathbf{D}|$, a ranking is a permutation $\boldsymbol{\tau} = \hat{\boldsymbol{\tau}}(\mathbf{D})$ of the n individuals or items, usually satisfying:

$$Y_{\boldsymbol{\tau}(1)} \geq Y_{\boldsymbol{\tau}(2)} \geq \cdots \geq Y_{\boldsymbol{\tau}(n)}. \tag{3}$$

To satisfy other objectives, like fairness, we generally output a ranking $\hat{\boldsymbol{\tau}}$ that is not simply sorting on the observed values of Y. Specifically, we aim to compute counterfactually fair rankings.

▶ **Definition 1** (Counterfactually fair ranking). *A ranking $\hat{\boldsymbol{\tau}}$ is counterfactually fair if, for all possible x and pairs of vectors of actual and counterfactual sensitive attributes $a \neq a'$, respectively, we have:*

$$\begin{aligned} &\mathbb{P}(\hat{\boldsymbol{\tau}}(Y_{\mathbf{A}\leftarrow\mathbf{a}}(U)) = k \mid \mathbf{X} = \mathbf{x}, \mathbf{A} = \mathbf{a}) \\ &= \mathbb{P}(\hat{\boldsymbol{\tau}}(Y_{\mathbf{A}\leftarrow\mathbf{a}'}(U)) = k \mid \mathbf{X} = \mathbf{x}, \mathbf{A} = \mathbf{a}) \end{aligned} \tag{4}$$

for any rank k, and with suitably randomized tie-breaking. If any mediators are considered resolving then the counterfactual $Y_{\mathbf{A}\leftarrow\mathbf{a}'}(U)$ in this definition is computed accordingly, holding such mediators fixed.

This definition is one natural adaptation of causal definitions in the recent literature on fairness in classification and prediction tasks [10, 26, 30, 36, 55] to the ranking setting. To satisfy Equation 4, we rank using counterfactuals that treat all individuals or items in the dataset according to one fixed baseline value $\mathbf{a}'$.

There are other possible definitions relaxing (4), for example using expected rank or enforcing equality for some but not all values of k. We leave the problems of deriving algorithms satisfying these and comparing performance to future work.

2.3.2 Implementation

We use the following procedure to compute counterfactually fair rankings, keeping our focus on model $\mathcal{M}_1$ in Figure 2a for clarity and readability.

1. For a (training) dataset $\mathbf{D}$, we estimate the parameters of the assumed causal model $\mathcal{M}$. A variety of frequentist or Bayesian approaches for estimation can be used. Our experiments use the R package `mediation` [46] on model $\mathcal{M}_1$ in Figure 2a.
2. From the estimated causal model we compute counterfactual records on the (training) data, transforming each observation to one reference subgroup $\mathbf{A} \leftarrow \mathbf{a}'$, we set $\mathbf{a}'$ to be the disadvantaged intersectional group. This yields counterfactual training data $\mathbf{D}_{\mathbf{A}\leftarrow\mathbf{a}'}$.
3. For score-based ranking, we sort $Y_{\mathbf{A}\leftarrow\mathbf{a}'}$ in descending order to produce the counterfactually fair ranking $\hat{\boldsymbol{\tau}}(Y_{\mathbf{A}\leftarrow\mathbf{a}'})$. For learning to rank (LTR), we apply a learning algorithm on $\mathbf{D}_{\mathbf{A}\leftarrow\mathbf{a}'}$ and consider two options, depending on whether the problem structure allows the use of the causal model at test time: if it does, then we in-process the test data from the learned causal model before ranking counterfactual test scores, and if it does not, then we rank the unmodified test data. We refer to the first case as cf-LTR and emphasize that in the second case *counterfactually fairness may not hold, or hold only approximately, on test data.*

Proposition 2 below says that this implementation, under common causal modeling assumptions, satisfies our fair ranking criteria. The proof is in Appendix A.1.

▶ **Proposition 2** (Implementing counterfactually fair ranking). *If the assumed causal model $\mathcal{M}$ is identifiable and correctly specified, implementations described above produce counterfactually fair rankings in the score-based ranking and cf-LTR tasks.*

3 Experimental Evaluation

In this section we investigate the behavior of our framework under different structural assumptions of the underlying causal model on real and synthetic datasets. We quantify performance with respect to several fairness and utility measures, for both score-based rankers and for learning to rank.

3.1 Datasets and evaluation measures

Datasets

We present experimental results on the real dataset COMPAS [1] and on a synthetic benchmark that simulates hiring by a moving company, inspired by Datta et al. [15]. We also present results on another synthetic benchmark that is a variant of the moving company dataset, but with an additional numerical sensitive attribute, in Appendix A.2.

COMPAS contains arrest records with sensitive attributes gender and race. We use a subset of *COMPAS* that includes Black and White individuals of either gender with at least 1 prior arrest. The resulting dataset has 4,162 records with about 25% White males, 59% Black males, 6% White females, and 10% Black females. We fit the causal model $\mathcal{M}_1$ in Figure 2a with gender G, race R, number of prior arrests X, and COMPAS decile score Y, with larger Y predicting higher likelihood of recidivism. In our use of this dataset, we will rank defendants on Y from lower to higher, prioritizing them for release or for access to supportive services as part of a comprehensive reform of the criminal justice system.

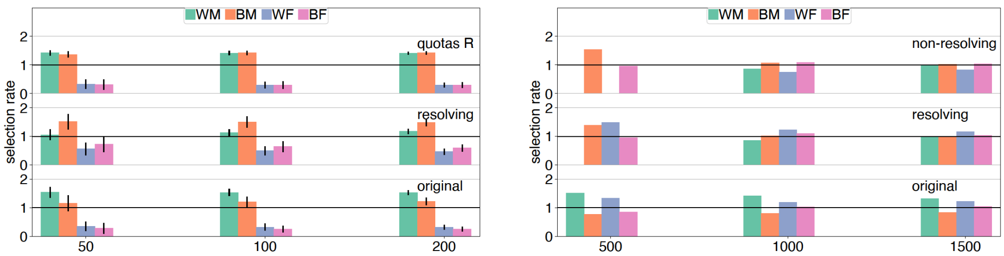
Moving company is a synthetic dataset drawn from the causal model $\mathcal{M}_1$ in Figure 2a, with edge weights: $w(G \rightarrow X) = 1$, $w(R \rightarrow X) = 0$, $w(G \rightarrow Y) = 0.12$, $w(R \rightarrow Y) = 0.08$, and $w(X \rightarrow Y) = 0.8$. This dataset is used in the scenario we discussed in our motivating example in Section 1.2: Job applicants are hired by the moving company based on their qualification score Y, computed from weight-lifting ability score X, and affected by gender G and race R, either directly or through X. Specifically, weight-lifting ability X is lower for female applicants than for male applicants; qualification score Y is lower for female applicants and for Blacks. Thus, the intersectional group Black females faces greater discrimination than either the Black or the female group. In our experiments in this section, we assume that women and Blacks each constitute a 37% minority of the applicants, and that gender and race are assigned independently. As a result, there are about 40% White males, 14% Black females, and 23% of both Black males and White females in the input with 2,000 records.

Fairness measures

We investigate whether the counterfactual ranking derived using our method is fair with respect to intersectional groups of interest, under the given structural assumptions of the underlying causal model. We consider two types of fairness measures: those that compare ranked outcomes across groups, and those that compare ranked outcomes within a group. To quantify fairness across groups, we use two common measures of fairness in classification that also have a natural interpretation for rankings: *demographic parity (DP) at top-k* and *equal opportunity (EO) at top-k*, for varying values of k. To quantify fairness within a group, we use a rank-aware measure called *in-group fairness ratio (IGF-Ratio)*, proposed by Yang et al. [49] to surface intersectional fairness concerns in ranking. We report our IGF-Ratio results in Appendix A.3, and refer the reader to an extended version of this paper [50] for experiments with other rank-aware fairness measures.

Demographic parity (DP) is achieved if the proportion of the individuals belonging to a particular group corresponds to their proportion in the input. We will represent DP by showing selection rates for each intersectional group at the top-k, with a value of 1 for all groups corresponding to perfect DP.

Equal opportunity (EO) in binary classification is achieved when the likelihood of receiving the positive prediction for items whose true label is positive does not depend on the values of their sensitive attributes [19]. To measure EO for LTR, we will take the set of items placed at the top-k in the ground-truth ranking to correspond to the positive class *for that value of k*. We will then present sensitivity (true positives / true positives + false negatives) per intersectional group at the top-k. If sensitivity is equal for all groups, then the method achieves EO.



(a) *moving company.* **(b)** *COMPAS.*

Figure 3 Demographic parity on the *moving company* and *COMPAS* datasets. X-axis shows the top-k values of the rankings and Y-axis shows the selection rate while each span of Y-axis represents different rankings and each color represents an intersectional group. The assumed causal model for both *moving company* and *COMPAS* is $\mathcal{M}_1$ in Figure 2a.

Utility measures

When the distribution of scores Y differs across groups, then we may need to sacrifice score-utility to produce a fair ranking. We evaluate the score-utility of the counterfactual rankings using two measures, *Y-utility loss at top-k*, applicable for both score-based ranking and LTR, and *average precision (AP)*, applicable only for LTR. Both compare a "ground truth" ranking $\boldsymbol{\tau}$ induced by the order of the observed scores Y to a proposed fair ranking $\boldsymbol{\sigma}$ (we use $\boldsymbol{\sigma}$ rather than $\hat{\boldsymbol{\tau}}$ here to make notation more readable).

We define Y-utility loss at top-k as $L_k(\boldsymbol{\sigma}) = 1 - \sum_{i=1}^{k} Y_{\boldsymbol{\sigma}(i)} / \sum_{i=1}^{k} Y_{\boldsymbol{\tau}(i)}$. $Y_{\boldsymbol{\sigma}(i)}$ is the observed score of the item that appears at position i in $\boldsymbol{\sigma}$, while $Y_{\boldsymbol{\tau}(i)}$ is the observed score of the item at position i in the original ranking $\boldsymbol{\tau}$. L_k ranges between 0 (best) and 1 (worst).

Average precision (AP) quantifies, in a rank-compounded manner, how many of the items that should be returned among the top-k are indeed returned. Recall that $\boldsymbol{\tau}_{1\ldots k}$ denotes the *set* of the top-k items in a ranking $\boldsymbol{\tau}$. We define precision at top-k as $P_k = |\boldsymbol{\tau}_{1\ldots k} \cap \boldsymbol{\sigma}_{1\ldots k}|/k$, where $\boldsymbol{\tau}$ is the "ground truth" ranking and $\boldsymbol{\sigma}$ is the predicted ranking. Then, $AP_k(\boldsymbol{\sigma}) = \sum_{i=1}^{k} P_i \times \mathbb{1}[\boldsymbol{\sigma}(i) \in \boldsymbol{\tau}_{1\ldots k}]/k$, where $\mathbb{1}$ is an indicator function that returns 1 if the condition is met and 0 otherwise. AP_k ranges between 0 (worst) and 1 (best).

3.2 Score-based ranking

In the first set of experiments, we focus on score-based rankers, and quantify performance of our method in terms of demographic parity (Figure 3 and 5) and score-based utility, on *moving company* (over 100 executions) and *COMPAS*.

Synthetic datasets

Recall that, in the *moving company* example, the goal is to compute a ranking of the applicants on their qualification score Y that is free of racial discrimination, while allowing for a difference in weight-lifting ability X between gender groups, thus treating X as a resolving variable. Figure 3a compares DP of three rankings for the moving company example: *original*, *resolving*, and *quotas on R*, described below.

Recall that perfect DP is achieved when selection rate equals to 1 for all groups. We observe that the *original* ranking, the bottom set of bars in Figure 3a, under-represents women (WF and BF) compared to their proportion in the input, and that White men (WM) enjoy a higher selection rate than do Black men (BM). Specifically, there are between 62-64% White men (40% in the input), 27-28% Black men (23% in the input), 6% White women (23% in the input), and 3-9% Black women (14% in the input) for $k = 50, 100, 200$.

In comparison, in the counterfactually fair ranking in which X is treated as *resolving*, shown as the middle set of bars in Figure 3a, selection rates are higher for the Blacks of both genders than for the Whites. For example, selection rate for White men is just over 1, while for Black men it's 1.5. Selection rates also differ by gender, because weight-lifting ability X is a mediator, and it encodes gender differences.

Finally, the ranking *quotas R*, the top set of bars in Figure 3a, shows demographic party for racial groups when the ranking is computed using representation constraints (quotas) on race R. This ranking is computed by independently sorting Black and White applicants on Y and selecting the top individuals from each list in proportion to that group's representation in the input. Opting for quotas on race rather than on gender, or on a combination of gender and race, is reasonable here, and it implicitly encodes a normative judgement that is explicit in our causal model $\mathcal{M}_1$ in Figure 2a – that race should not impact the outcome, while gender may.

Appendix A.2 describes another synthetic dataset, *moving company+age*, with three sensitive attributes: categorical gender G and race R, and numerical age A, with records drawn from the causal model $\mathcal{M}_5$ in Figure 2e. Our results on this dataset further showcase the flexibility of our framework.

Real datasets

We now present results of an evaluation of our method on a real dataset, *COMPAS*. Figure 3b shows demographic parity (DP) of three different rankings: *original*, *resolving*, and *non-resolving*, discussed below. Recall that in our use of *COMPAS* defendants are ranked on their decile score Y from lower to higher, prioritizing them for release or for access to supportive services. Our goal is to produce a ranking that is free of racial and gender discrimination. There is some debate about whether the number of prior arrests, X, should be treated as a resolving variable. By treating X as non-resolving, we are stating that the number of prior arrests is itself subject to racial discrimination.

We observe that, in the *original* ranking, shown as the bottom set of bars in Figure 3b, Whites of both genders are selected at much higher rates than Blacks. Gender has different effect by race: men are selected at higher rates for Whites, and at lower rates for Blacks. There are 33-38% White men (25% in the input), 46-49% Black men (59% in the input), 7-8% White women (6% in the input), and 8-10% Black women (10% in the input), for $k = 500, 1000, 1500$.

Comparing the original ranking to the counterfactually fair ranking that treats the number of prior arrests X as a *resolving* mediator, shown as the middle set of bars in Figure 3b, we observe an increase in selection rates for Black males and Black females, and a significant reduction in selection rates for White males. Further, comparing with the counterfactually fair ranking that treats X as *non-resolving*, the top set of bars in Figure 3b, we observe that only Black individuals are represented at the top-500, and that selection rates for all intersectional groups for larger values of k are close to 1, achieving demographic parity.

We also computed utility loss at top-k, based on the original Y scores (see Section 3.1 for details). For *moving company*, we found that counterfactually fair ranking *resolving* suffers at most 1% loss across the values of k, slightly higher than the loss of the *quotas R* ranking, which is close to 0. For *COMPAS*, we found that overall utility loss is low in most cases, ranging between 3% and 8% in the fair ranking *resolving*, and between 3% and 10% in the fair ranking *non-resolving*. The slightly higher loss for the latter case is expected, because we are allowing the model to correct for historical discrimination in the data more strongly in this case, thus departing from the original ranking further.

3.3 Learning to rank

We now investigate the usefulness of our method for supervised learning of counterfactually fair ranking models. We use ListNet, a popular Learning to Rank algorithm, as implemented by Ranklib[1]. ListNet is a listwise method – it takes ranked lists as input and generates predictions in the form of ranked lists. We choose ListNet because of its popularity and effectiveness (see additional information about ListNet and other predictive ranking models in [32] and [34], respectively).

We conduct experiments in two regimes that differ in whether to apply our method as a preprocessing fairness intervention on the test set (see Implementation in Section 2). In both regimes, we make the training datasets counterfactually fair. Specifically, we first fit a causal

[1] `https://sourceforge.net/p/lemur/wiki/RankLib/`

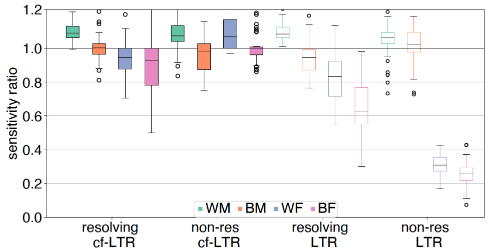


Figure 4 Equal opportunity on *moving company* with $k = 200$. X-axis shows the treatments: training & test on fair rankings with X as resolving (resolving cf-LTR) and non-resolving (non-resolving cf-LTR); training on fair rankings & test on unmodified rankings with X as resolving (resolving LTR) and non-resolving (non-resolving LTR). Y-axis shows the ratio of sensitivity between each counterfactually fair treatment and the original ranking. Intersectional groups are denoted by different colors. Solid boxes correspond to cf-LTR variants. All results are over 50 training/test pairs.

model $\mathcal{M}$ on the training data, then update the training data to include counterfactually fair values of the score Y and of any non-resolving mediators X, and finally train the ranking model $\mathcal{R}$ (e.g., ListNet) on the fair training data. We now have two options: (1) to run $\mathcal{R}$ on the *unmodified (biased) test data*, called *LTR* in our experiments, or; (2) to *preprocess test data* using $\mathcal{M}$, updating test with counterfactually fair values for the score Y and for any non-resolving mediators X, before passing it on to $\mathcal{R}$, called *cf-LTR*.

Note that the cf-LTR setting shows the effectiveness of our method for the disadvantaged intersectional groups, in that the performance of the model is compareble across groups, while LTR setting shows the performance of a ranking model on biased test data. Similar to score-based ranking, we also consider two structural assumptions of the underlying causal model: resolving and non-resolving for each setting above.

We quantify performance of our method in terms of equal opportunity (EO) and average precision (AP) (see Section 3.1), on *moving company* over 50 training/test pairs. Figure 4 shows performance of the ranking model (e.g., ListNet) in terms of equal opportunity on *moving company*, comparing four settings produced from above options: resolving cf-LTR, non-resolving cf-LTR, resolving LTR, and non-resolving LTR. Recall that a method achieves equal opportunity (EO) if sensitivity is equal across groups. Note that sensitivity is affected by groups' representation in the data, meaning that higher sensitivity for a group might be due to its limited representation in the top-k rankings (lower positives) rather than the better treatment in the model (higher true positives). Thus, to reduce the effect of imbalanced representation across groups, we present *sensitivity ratio*: the ratio of the sensitivity at each setting above (with the fairness treatment on training, or on both training and test data) to the sensitivity of the original ranking model (without any fairness intervention) in Figure 4.

Note that the original ranking model achieves high sensitivity for all intersectional groups (0.9, 0.9, 0.95, and 1 for White men, Black men, White women, and Black women, respectively) and so can be seen as achieving EO within gender groups, because their representation at the top-k is similar. As shown in Figure 4, performance of the fair ranking models (e.g., the cf-LTR variants in the left two columns for resolving and non-resolving X respectively), in which both the training and the test data are counterfactually fair, is comparable to the original ranking model in terms of sensitivity, with the medians of all boxes close to the sensitivity ratio of 1.

The resolving variants (e.g., resolving cf-LTR and LTR columns in Figure 4) show lower sensitivity for women, likely because women are selected at lower rates since X is treated as resolving for gender). The LTR variants (e.g., resolving and non-res LTR columns in Figure 4) show lower sensitivity for women because the test dataset is unmodified in this set of experiments. Finally, when the fairness intervention is applied on both training and test datasets (e.g., resolving and non-res cf-LTR columns in Figure 4), it leads to better sensitivity for women.

We also quantified utility as average precision (AP) in evaluating supervised learning of counterfactually fair ranking models. For *moving company*, AP is 77% for the original ranking model when unmodified ranking are used for training and test. For counterfactually fair training data with non-resolving X (weight-lifting), AP on unmodified test (non-res LTR) is 27% but it increases to 91% when test data is preprocessed (non-res cf-LTR). For counterfactually fair training data with resolving X, AP is 68% for unmodified test (resolving LTR) and 83% when test is preprocessed (resolving cf-LTR).

4 Discussion

This work aims to mitigate the negative impacts of ranking systems on people due to attributes that are out their control. In this section we anticipate and discuss concerns that may arise in the application of our method.

There are objections to modeling sensitive attributes as causes rather than considering them to be immutable, defining traits. Some of these objections and responses to them are discussed in [33]. In the present work we proceed with an understanding that the model is a simplified and reductive approximation, and support for deploying an algorithm and claiming it is fair should require an inclusive vetting process where formal models such as these are *tools for inclusively achieving consensus* and not for rubber stamping or obfuscation.

There are many issues outside the scope of the present work but which are important in any real application. Choices of which attributes are sensitive, which mediators are resolving (and for which sensitive attributes), the social construction and definitions of sensitive attributes, choices of outcome/utility or proxies thereof, technical limitations in causal modeling, the potential for (adversarial) misuse are all issues that may have adverse impacts when using our method. We do stress that these are not limitations inherent to our approach in particular, rather, these concerns arise for virtually any approach in a sensitive application. For an introductions to these issues, including a causal approach to them, see [4, 29].

Further, like any approach based on causality, our method relies on strong assumptions that are untestable in general, though they may be falsified in specific cases. Sequential ignorability in particular is a stronger assumption in cases with more mediating variables, or with a mediator that is causally influenced by many other variables (observed or unobserved). Such cases increase the number of opportunities for sequential ignorability to be violated for one of the mediators or by one of the many causes of a heavily influenced mediator.

Finally, intersectional fairness is not a purely statistical or algorithmic issue. As such, any technical method will require assumptions at least as strong as the causal assumptions we make. In particular, there are normative and subtle empirical issues embedded in any approach to fairness, such as the social construction of sensitive attributes, or the choice of which mediators may be considered resolving in our framework. For these reasons *we believe the burden of proof should fall on any approaches assuming the world (causal model) is already less unfair or that fairness interventions should be minimized*, for example by the use of resolving variables.

5 Related Work

Intersectionality. From the seeds of earlier work [13], including examples that motivated our experiments [14], intersectional feminism has developed into a rich interdisciplinary framework to analyze power and oppression in social relations [12, 43]. We refer especially to the work of Noble [37], and D'Ignazio and Klein [16], in the context of data and information technology. Other recent technical work in this area focuses on achieving guarantees across intersectional subgroups [20, 24, 27], including on computer vision tasks [7], or makes connections to privacy [18]. These do not take a causal approach or deal with ranking tasks. In our framework, intersectionality does not simply refer to a redefinition of multiple categorical sensitive attributes into a single product category or inclusion of interaction terms, as was done in recent work [20, 24, 27]. Specific problems may imply different constraints or interpretations for different sensitives attributes, as shown in the *moving company* example, where a mediator (e.g., weight-lifting ability) may be considered resolving for one sensitive attribute but not for another.

Causality and fairness. A growing literature on causal models for fair machine learning [10, 26, 30, 36, 55] emphasizes that fairness is a normative goal that relates to real world (causal) relationships. One contribution of the present work is to connect intersectionality and fair ranking tasks to this literature, and therefore to the rich literature on causal modeling. Some recent work in causal fairness focuses on the impact of learning optimal, fair policies, potentially under relaxations of standard causal assumptions that allow interference [28, 35]. Some of the most closely related work uses causal modeling to analyze intersectional fairness from a philosophical standpoint [6] or in a public health setting [22], but these are focused on foundations and interpretation, rather than on implementation or machine learning tasks.

Ranking and fairness. While the majority of the work on fairness in machine learning focuses on classification or risk prediction, there is also a growing body of work on fairness and diversity in ranking [2, 8, 9, 31, 45, 48, 49, 51, 52, 53], including a recent survey [54]. Yang et al. [49] consider intersectional concerns, although not in a causal framework. The authors observe that when representation constraints are stated on individual attributes, like race and gender, and when the goal is to maximize score-based utility subject to these constraints, then a particular kind of unfairness can arise, namely, utility loss may be imbalanced across intersectional groups. Barnabò et al. [3] study similar problem through explicitly modeling the trade-off between utility and fairness constraints. In our experiments we observed a small imbalance in utility loss across intersectional groups (1-5%) and will investigate the conditions under which this happens in future work. Finally, Wu et al. [48] apply causal modeling to fair ranking but estimates scores from observed ranks, uses causal discovery algorithms to learn an SCM, and does not consider intersectionality, while the present work considers the case when scores are observed and the SCM chosen *a priori*.

6 Conclusion

Our work builds on a growing literature for causal fairness to introduce a modeling framework for intersectionality and apply it to ranking. Experiments show that this approach can be flexibly applied to different scenarios, including ones with mediating variables, and the results compare reasonably to intuitive expectations we may have about intersectional fairness for those examples. The flexibility of our approach and its connection to causal methodology makes possible a great deal of future work including exploring robustness of rankings to unmeasured confounding [25] or uncertainty about the underlying causal model [41].

Future technical work can relax some assumptions under specific combinations of model structures, estimation methods, and learning task algorithms. For example, we have shown in experiments that the LTR task (without in-processing) with ListNet works reasonably well, but future work could identify the conditions when this insensitivity of a learned ranker to counterfactual transformations on the training data guarantees that counterfactual fairness will hold at test time, perhaps with explicit bounds on discrepancies due to issues like covariate shift. We proposed ranking on counterfactual scores, treating everyone as a member of the disadvantaged intersectional group, but there are other possible fair strategies. For any fixed baseline intersectional group, for example the most advantaged one, if we compute counterfactuals and treat everyone as though they belong to that fixed group, we would also achieve intersectional counterfactual fairness. The same is true if we treat everyone based on the average of their counterfactual values for all intersectional subgroups. Future work may explore whether any of these choices have formal or computational advantages, making them preferable in specific settings.

References

1 Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica*, 23, 2016.

2 Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das. Designing fair ranking schemes. In *ACM SIGMOD*, pages 1259–1276, 2019. `doi:10.1145/3299869.3300079`.

3 Giorgio Barnabò, Carlos Castillo, Michael Mathioudakis, and Sergio Celis. Intersectional affirmative action policies for top-k candidates selection. *CoRR*, abs/2007.14775, 2020. `arXiv:2007.14775`.

4 Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NIPS Tutorial*, 2017.

5 Emery Berger. CSRankings: Computer Science Rankings, 2017–2020. Online, retrieved June 2, 2020. URL: `http://csrankings.org/`.

6 Liam Kofi Bright, Daniel Malinsky, and Morgan Thompson. Causally interpreting intersectionality theory. *Philosophy of Science*, 83(1):60–81, 2016.

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

8 L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. Interventions for ranking in the presence of implicit bias. In *FAT**, pages 369–380. ACM, 2020. `doi:10.1145/3351095.3372858`.

9 L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. Ranking with fairness constraints. In *ICALP*, volume 107 of *LIPIcs*, pages 28:1–28:15, 2018. `doi:10.4230/LIPIcs.ICALP.2018.28`.

10 Silvia Chiappa. Path-specific counterfactual fairness. In *AAAI*, volume 33, pages 7801–7808, 2019.

11 Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM*, 63(5):82–89, 2020. `doi:10.1145/3376898`.

12 Patricia Hill Collins. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. routledge, 2002.

13 Combahee River Collective. The Combahee river collective statement. *Home girls: A Black feminist anthology*, pages 264–74, 1983.

14 Kimberle Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, page 139, 1989.

15 Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE SP*, pages 598–617, 2016. doi:10.1109/SP.2016.42.

16 Catherine D'Ignazio and Lauren F Klein. *Data feminism.* MIT Press, 2020.

17 Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

18 James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *IEEE ICDE*, pages 1918–1921, 2020.

19 Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NIPS*, pages 3315–3323, 2016. URL: https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html.

20 Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *ICML*, pages 1939–1948, 2018.

21 Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical science*, pages 51–71, 2010.

22 John W Jackson and Tyler J VanderWeele. Intersectional decomposition analysis with differential exposure, effects, and construct. *Social Science & Medicine*, 226:254–259, 2019.

23 Pearl Judea. Causality: models, reasoning, and inference. *Cambridge University Press*, 2000.

24 Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *ICML*, pages 2569–2577, 2018. URL: http://proceedings.mlr.press/v80/kearns18a.html.

25 Niki Kilbertus, Philip J. Ball, Matt J. Kusner, Adrian Weller, and Ricardo Silva. The sensitivity of counterfactual fairness to unmeasured confounding. In *UAI*, page 213, 2019. URL: http://proceedings.mlr.press/v115/kilbertus20a.html.

26 Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *NIPS*, pages 656–666, 2017.

27 Michael P. Kim, Amirata Ghorbani, and James Y. Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *AIES*, pages 247–254, 2019. doi:10.1145/3306618.3314287.

28 Matt Kusner, Chris Russell, Joshua Loftus, and Ricardo Silva. Making decisions that reduce discriminatory impacts. In *ICML*, pages 3591–3600, 2019.

29 Matt J Kusner and Joshua R Loftus. The long road to fairer algorithms. *Nature*, 2020.

30 Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *NIPS*, pages 4066–4076, 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

31 Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*, pages 1334–1345. IEEE, 2019. doi:10.1109/ICDE.2019.00121.

32 Tie-Yan Liu. *Learning to Rank for Information Retrieval.* Springer, 2011. doi:10.1007/978-3-642-14267-3.

33 Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint*, 2018. arXiv:1805.05859.

34 Bhaskar Mitra, Nick Craswell, et al. *An introduction to neural information retrieval.* Now Foundations and Trends, 2018.

35 Razieh Nabi, Daniel Malinsky, and Ilya Shpitser. Learning optimal fair policies. In *International Conference on Machine Learning*, pages 4674–4682, 2019.

36 Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

37 Safiya Umoja Noble. *Algorithms of oppression: How search engines reinforce racism*. nyu Press, 2018.

38 Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 411–420, 2001.

39 James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

40 James M Robins. Semantics of causal dag models and the identification of direct and indirect effects. *Oxford Statistical Science Series*, pages 70–82, 2003.

41 Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *NIPS*, pages 6414–6423, 2017.

42 Maya Sen and Omar Wasow. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19, 2016.

43 Stephanie A Shields. Gender: An intersectionality perspective. *Sex roles*, 59(5-6):301–311, 2008.

44 Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

45 Julia Stoyanovich, Ke Yang, and H. V. Jagadish. Online set selection with fairness and diversity constraints. In *EDBT*, pages 241–252. OpenProceedings.org, 2018. `doi:10.5441/002/edbt.2018.22`.

46 Dustin Tingley, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 2014.

47 Tyler VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

48 Yongkai Wu, Lu Zhang, and Xintao Wu. 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, KDD 2018, London, UK, August 19-23, 2018*, pages 2536–2544, 2018. `doi:10.1145/3219819.3220087`.

49 Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. Balanced ranking with diversity constraints. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 6035–6042, 2019. `doi:10.24963/ijcai.2019/836`.

50 Ke Yang, Joshua R. Loftus, and Julia Stoyanovich. Causal intersectionality for fair ranking, 2020. `arXiv:2006.08688`.

51 Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *ACM SSDBM*, pages 22:1–22:6, 2017. `doi:10.1145/3085504.3085526`.

52 Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In Ee-Peng Lim, Marianne Winslett, Mark Sanderson, Ada Wai-Chee Fu, Jimeng Sun, J. Shane Culpepper, Eric Lo, Joyce C. Ho, Debora Donato, Rakesh Agrawal, Yu Zheng, Carlos Castillo, Aixin Sun, Vincent S. Tseng, and Chenliang Li, editors, *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 1569–1578. ACM, 2017. `doi:10.1145/3132847.3132938`.

53 Meike Zehlike and Carlos Castillo. Reducing disparate exposure in ranking: A learning to rank approach. In Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen, editors, *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 2849–2855. ACM / IW3C2, 2020. `doi:10.1145/3366424.3380048`.

54 Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking: A survey, 2021. `arXiv:2103.14000`.

55 Junzhe Zhang and Elias Bareinboim. Fairness in decision-making – the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

A Appendix

A.1 Proof of Proposition 2

Proposition 2 (Implementing counterfactually fair ranking) If the assumed causal model $\mathcal{M}$ is identifiable and correctly specified, implementations described above produce counterfactually fair rankings in the score based ranking and cf-LTR tasks.

Proof of Proposition 2. The proof is essentially by construction, but we provide more detail now for model $\mathcal{M}_1$. Fixing a baseline intersectional subgroup (g_0, r_0), the counterfactual training data in our implementation will use $Y_{(G,R)\leftarrow(g_0,r_0)}$, either by ranking these for score based ranking or training a predictive model for LTR. We wish to show that

$$\mathbb{P}(\hat{\boldsymbol{\tau}}(Y_{(G,R)\leftarrow(g,r)}) = k \mid X = x, (G, R) = (g, r)) \tag{5}$$

is unchanged under all counterfactual transformations, denoted by $Y_{(G,R)\leftarrow(g',r')}$, if the causal model has been correctly specified. First, we consider the case where the functions f_X, f_Y are known. If X is resolving, then

$$(Y_i)_{(G,R)\leftarrow(g_0,r_0)} = f_Y(x_i, g_0, r_0) + \epsilon_i^Y$$

for all i. In this case the conditional distribution of these scores (5) is invariant under counterfactual transformations $(g, r) \leftarrow (g', r')$ because x_i is held fixed, (g', r') will be substituted with the fixed baseline values (g_0, r_0), and the error term is exogenous and in particular its distribution does not change under transformations of (g, r). If X is not resolving then we use

$$(Y_i)_{(G,R)\leftarrow(g_0,r_0)} = f_Y(f_X(g_0, r_0) + \epsilon_i^X, g_0, r_0) + \epsilon_i^Y$$

Under counterfactual transformations $(g, r) \leftarrow (g', r')$ all of the inputs above stay fixed except for the error terms, and, as before, these errors do not depend on (g, r) so the training data scores have the desired distributional invariance. The semi-resolving case is similar.

For score based ranking $\hat{\boldsymbol{\tau}}$ sorts the counterfactual scores, denoted by $(Y_i)_{(G,R)\leftarrow(g_0,r_0)}$. Since the distributions of these scores are unchanged under counterfactual transformations as we just established, the probability for any score to equal a given rank k is also unchanged, hence $\hat{\boldsymbol{\tau}}$ is a counterfactually fair ranking. In cf-LTR, at test time the test data is first transformed to the intervened version $\mathbf{D}^{\text{test}}_{(G,R)\leftarrow(g_0,r_0)}$ before inputting to $\hat{\boldsymbol{\tau}}$. As before, the distribution of the predicted rank for observation i under any counterfactual transformation $(G, R) \leftarrow (g', r')$ is fixed to that of the distribution under $(G, R) \leftarrow (g_0, r_0)$, which depends only on the exogenous errors.

Finally, we relax the assumption that the functions f_X, f_Y are known. Since we have assumed the causal model is identifiable and correctly specified (in particular, it satisfies sequential ignorability in cases where the model has mediators), these functions can be estimated on the (training) data via any appropriate causal inference method. Hence, counterfactually fair ranking condition will hold approximately due to plug-in estimation error. ◀

A.2 Additional experimental results: score-based ranking

In this section, we show evaluation results of using our method on a more complicated data under a different causal model: a synthetic dataset with three sensitive attributes and one of them is a continuous or numeric attribute (e.g., age) under an assumed causal model $\mathcal{M}_5$ in Figure 2e.

Moving company + age is a variant of *moving company* dataset with $10,000$ records drawn from the causal model $\mathcal{M}_5$ in Figure 2e, with three sensitive attributes: gender G, race R, and age A, with edge weights $w(G \rightarrow X) = 0.95$, $w(R \rightarrow X) = 0$, $w(A \rightarrow X) = 0.05$, $w(G \rightarrow Y) = 0.1$, $w(R \rightarrow Y) = 0.1$, $w(A \rightarrow Y) = 0.1$, and $w(X \rightarrow Y) = 0.7$. Age A affects the weight-lifting ability score X and the qualification score Y in a piece-wise linear fashion, with X and Y decreasing for ages A above some thresholds. Specifically, the effect of age on X is negligible for ages below 45, then slightly negative, and more strongly negative above age 55. The mean age for White and Black individuals are 35 and 45 respectively. We use this dataset to showcase the applicability of our framework to cases with more than two sensitive attributes, and to cases where sensitive attributes may be continuous.

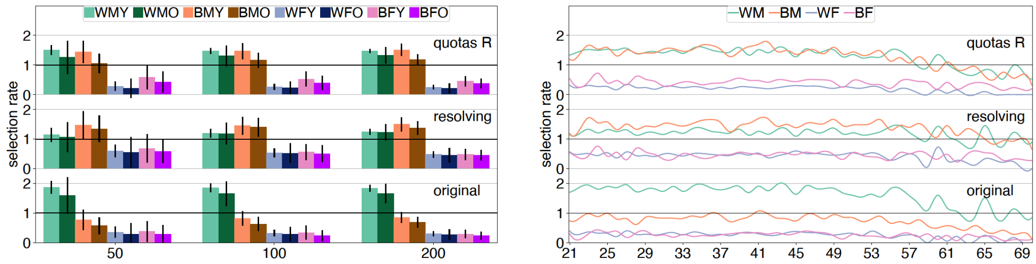
Figure 5 shows the performance of our methods in terms of demographic parity on *moving company+age* (over 100 executions), focusing on three different rankings: *original*, *resolving*, and *quotas R*. Recall that *moving company+age* includes a continuous sensitive attribute age in addition to gender and race. We present selection rates for two age groups, younger (age < 45) and older (age ≥ 45) in Figure 5a, and at each age in Figure 5b. We observe that in the *original* ranking, the bottom set of bars in Figure 5a, younger applicants are selected at a higher rate compared to older applicants within each intersectional group. For example,young White males and young Black males are both selected at higher rate than their older counterparts old White males and old Black males. Further, selection rates for racial and gender groups differ in the *original* ranking. For example, White males are selected at a much higher rate than other intersectional groups. These disparities in selection rates are preserved in the *quotas R* ranking, shown as the top set of bars in Figure 5a. Recall that the goal for *moving company+age* is to compute a ranking of the applicants that is free of racial and age discrimination while allowing for a difference in weight-lifting ability X between gender groups, thus treating X as resolving variable. In the counterfactually fair ranking *resolving*, the middle set of bars in Figure 5a, we observe an increase in selection rates for Black males, and also note that the age of the applicants does not materially affect their selection rates.

Figure 5b presents selection rates for each value of age, for each intersectional group on gender and race, at the top-200. Observe that the *original* ranking, shown in the bottom set of lines, exhibits a disparity in selection rates between the Black and the White applicants for all age values, and that selection rates drop substantially for all groups around age ≥ 50. The *quotas R* ranking, the top set of lines in Figure 5b, reduces the disparity in selection rates between racial groups (e.g., there is no gap between the lines for White males and Black males for any age), but it still shows a disparity by age, meaning that selection rates drop for all groups around age ≥ 50, just as they did in the original ranking. Finally, the counterfactually fair ranking *resolving*, shown as the middle set of lines in Figure 5b, reduces disparities in selection rates by both race and gender.

We also computed utility loss at top-k, based on the original Y scores (see Section 3.1 for details). For *moving company+age*, the loss of the counterfactually fair ranking *resolving* and of the *quotas R* ranking is at most 1% across the values of k.

A.3 Additional experimental results: rank-aware fairness measures

In this section, we report evaluation results of using a rank-aware fairness measure called **in-group fairness ratio (IGF-Ratio)** on *moving company*, *moving company+age*, and *COMPAS*. In-group fairness ratio (IGF-Ratio) is the simpler of two in-group fairness measures proposed in [49]. It captures an important intersectional concern that arises when an input ranking must be re-ordered (and thus suffer a utility loss) to satisfy some fairness or



(a) **(b)** $k = 200$.

Figure 5 Demographic parity on the *moving company+age* dataset. The X-axis shows the top-k values of the rankings for (a) and shows the value of the attribute age A for (b). For both subplots, the Y-axis shows the selection rate, while each span of Y-axis represents different rankings and each color represents an intersectional group. The assumed causal model is $\mathcal{M}_5$ in Figure 2e. Figure 5a shows the results for the binarized attribute age A according to a threshold: younger ($\mathbf{Y}$): age < 45 and older ($\mathbf{O}$): age ≥ 45.

diversity constraint. Specifically, IGF-Ratio compares the amount of re-ordering within each intersectional groups, and considers a ranking fair if the corresponding loss is balanced across groups. Let us denote by $\boldsymbol{\tau}_{1\ldots k}$ the *set* of the top-k items in $\boldsymbol{\tau}$. For a given intersectional group g and position k, $\mathit{IGF\text{-}Ratio}_k(\boldsymbol{\tau}, g)$ is the ratio of lowest score of any item from g in $\boldsymbol{\tau}_{1\ldots k}$ and the highest score of an item from g not in $\boldsymbol{\tau}_{1\ldots k}$. IGF-Ratio requires non-negative scores and ranges from $[0, 1]$, with higher values implying better in-group fairness. To make the scores non-negative, we increase the values of Y by $|\min(Y)|$.

Table 1 shows the results of in-group fairness ratio (IGF-Ratio) in counterfactually fair score-based ranking derived using our method on *moving company* (over 100 executions), *moving company+age* (over 100 executions), and *COMPAS*. To compute this measure, we cannot have any ties in the ranking. For *COMPAS*, we broke the ties by Y-score by randomly permuting the items within an equivalence class by score. Recall that IGF-Ratio ranges between 0 and 1 and that a higher value is better, since it indicates that the ratio of the score of the lowest-scoring selected item among the top-k and of the highest-scoring item not among the top-k is close to 1. Observe that most IGF-Ratio values are close to 1, meaning that there is only a limited amount of re-ordering of individuals within each intersectional group. Further, in-group fairness loss in terms of IGF-Ratio is balanced among intersectional groups in all cases, while some groups (e.g. White males) face a slightly lower but acceptable IGF-Ratio in the fair ranking *non-resolving*.

Table 1 IGF-Ratio on *moving company*, *moving company+age*, and *COMPAS*. A higher value is better: it indicates that the ratio of scores of the lowest-scoring selected item among the top-k and of the highest-scoring item not among the top-k is close to 1. In the table, $k_{1,2,3} = 50, 100, 200$ for *moving company* ($n = 2000$) and *moving company+age* ($n = 10,000$), and $k_{1,2,3} = 500, 1000, 1500$ for *COMPAS* ($n = 4162$). N/A is used when a particular intersectional group is not represented among the top-k.

Dataset	Ranking	k_1				k_2				k_3			
		WM	BM	WF	BF	WM	BM	WF	BF	WM	BM	WF	BF
moving company	non-res	0.98	0.94	0.93	0.94	0.96	0.95	0.91	0.92	0.94	0.94	0.89	0.89
	resolving	0.95	0.95	0.98	0.98	0.93	0.93	0.96	0.96	0.92	0.92	0.93	0.94
moving company+age	non-res	0.82	0.9	0.99	0.99	0.8	0.99	0.99	0.98	0.99	0.98	0.99	0.97
	resolving	0.99	0.99	0.99	0.94	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
COMPAS	non-res	N/A	1.00	N/A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	resolving	N/A	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00