Individual Fairness in Pipelines

Cynthia Dwork
Harvard John A Paulson School of Engineering and Applied Sciences, Cambridge, MA, USA
Radcliffe Institute for Advanced Study, Cambridge, MA, USA
Microsoft Research, Mountain View, CA, USA
dwork@seas.harvard.edu

Christina Ilvento
Harvard John A Paulson School of Engineering and Applied Sciences, Cambridge, MA, USA
cilvento@g.harvard.edu

Meena Jagadeesan
Harvard University, Cambridge, MA, USA
mjagadeesan@college.harvard.edu

Abstract

It is well understood that a system built from individually fair components may not itself be individually fair. In this work, we investigate individual fairness under pipeline composition. Pipelines differ from ordinary sequential or repeated composition in that individuals may drop out at any stage, and classification in subsequent stages may depend on the remaining "cohort" of individuals. As an example, a company might hire a team for a new project and at a later point promote the highest performer on the team. Unlike other repeated classification settings, where the degree of unfairness degrades gracefully over multiple fair steps, the degree of unfairness in pipelines can be arbitrary, even in a pipeline with just two stages.

Guided by a panoply of real-world examples, we provide a rigorous framework for evaluating different types of fairness guarantees for pipelines. We show that naïve auditing is unable to uncover systematic unfairness and that, in order to ensure fairness, some form of dependence must exist between the design of algorithms at different stages in the pipeline. Finally, we provide constructions that permit flexibility at later stages, meaning that there is no need to lock in the entire pipeline at the time that the early stage is constructed.

2012 ACM Subject Classification Theory of computation → Machine learning theory

Keywords and phrases algorithmic fairness, fairness under composition, pipelines

Digital Object Identifier 10.4230/LIPIcs.FORC.2020.7

Related Version `https://arxiv.org/abs/2004.05167`

Funding *Cynthia Dwork*: Research supported in part by NSF grant CCF-1763665 and Microsoft Research.
Christina Ilvento: This work was funded by the Sloan Foundation.

1 Introduction

As algorithms reach ever more deeply into our daily lives, there is increasing concern that they be *fair*. The study of the theory of algorithmic fairness was initiated by Dwork et al. [5], who introduced the solution concept of *individual fairness*. Roughly speaking, individual fairness requires that similar individuals receive similar distributions on outcomes. Dwork and Ilvento [6] examined the behavior of individual fairness (and various group notions of fairness) under composition. They showed that although competitive composition, i.e., when two different tasks "compete" for individuals, can result in arbitrarily bad behavior under composition, fairness under simple repeated or sequential classifications (for the same task) degrades gracefully, similar to degradation of differential privacy loss under multiple computations.[1] In this work we expand the investigation of individual fairness under sequential composition to the case of *cohort pipelines*. Cohort pipelines differ from ordinary sequential composition in that each stage of the pipeline considers only the remaining cohort of individuals and may change its classification strategy conditioned on the set of individuals remaining.

Cohort pipelines are common: many data-driven systems consist of a sequence of cohort selection or filtering steps, followed by decision or scoring steps. A running exemplary scenario in this work will be a two-stage cohort pipeline: a company hires a team (cohort) of individuals to work on a project and subsequently promotes the highest performer on the team to a leadership position. Although the team selection may be fair in the sense that similarly qualified candidates have similar chances of being chosen for the team, the selection of the highest performer critically depends on the *other members of the team*. As we will see, being compared fairly to other members of the cohort in each stage doesn't imply fairness of the entire pipeline, as the competitive landscape can vary between similar individuals.

Indeed, a fair cohort selection mechanism [6] can exploit the "myopic" nature of the promotion stage to skew overall fairness. This can happen either through good intentions (e.g., choosing teams so that members of a minority group always have a mentor on the team) or malice (e.g., ensuring that minority candidates are almost always paired with a more qualified majority candidate): in both these cases minorities suffer significantly reduced chances of promotion.[2] Unlike other repeated classification settings in which the degree of unfairness of multiple fair steps degrades gracefully, the degree of unfairness in cohort pipelines can be arbitrary, even in a pipeline with just two stages. Furthermore, we demonstrate that construction of malicious pipelines under naïve auditing of fairness is straightforward and both computationally and practically feasible.

In this work we examine the subtle issues that arise in cohort-based pipelines, focusing on short pipelines consisting of a single cohort selection step followed by a scoring step. We formalize fairness desiderata capturing the issues unique to pipelines (not shared by ordinary sequential composition), and give constructions for *robust* cohort selection mechanisms that behave well under (i.e., are robust to) pipeline composition with a variety of future scoring policies. In particular, we demonstrate that it is possible to design cohort selection mechanisms that are robust to a rich family of subsequent scoring functions given a simple description of a *policy* governing the behavior of the family.[3] This provides, for example, a means for enabling a company to choose an individually fair hiring procedure that will be robust to many possible compensation functions (all adhering to the policy) chosen at a later date. Guided by a panoply of real-world examples, this work provides a rigorous framework for evaluating and ensuring different types of fairness guarantees for pipelines.

We now summarize our contributions. First, we formalize what it means for the outcomes of a pipeline, which include both the outcome of the initial cohort selection step and the score conditioned on being chosen, to be fair.[4] We then extend this fairness notion to describe how a cohort selection mechanism can be *robust* to a scoring policy, i.e., to compose fairly with any cohort scoring function chosen from a permissible set. Although the choice of scoring function may not depend on the cohort, the scores assigned to any individual may be highly

---

[1] Note that although fairness degrades gracefully in these scenarios, it does not rule out the existence of feedback loops which arbitrarily amplify unfairness, see e.g., [12, 20].

[2] See Appendix A for additional examples.

[3] Formally, we can think of a policy as a description of a set of permitted scoring functions.

[4] Bower et al. consider fairness in pipelines for a group-based definition of fairness, and primarily consider the accuracy of the final pipeline decision [1].

dependent on their cohort "context." Second, we determine how the scoring policy imposes conditions on the cohort selection mechanism. In particular, we show that there is a natural way to describe the set of cohort *contexts* in which similar individuals are treated similarly by all functions permitted by the policy, and we demonstrate that assigning similar individuals to similar *distributions* over cohort contexts is sufficient (and sometimes necessary) to ensure pipeline robustness. Third, we provide constructions for cohort selection mechanisms which are both robust to a rich set of practical scoring policies and permit flexibility in selection of the original cohort.

2 Model and Definitions

2.1 Preliminaries

We base our model on individual fairness, as proposed in [5]. The intuition behind individual fairness is that "similar individuals should be treated similarly." What constitutes similarity for a particular classification task is provided by a metric which captures society's best understanding of who is similar to whom. Below we formally define individual fairness as in [5] with a natural Lipschitz relaxation.

▶ **Definition 1** (α-Individual Fairness [5]). *Given a universe of individuals U, and a metric $\mathcal{D} : U \times U \to [0,1]$ for a classification task with outcome set O, and a distance metric $d : \Delta(O) \times \Delta(O) \to [0,1]$ over distributions over outcomes, a randomized classifier $C : U \to \Delta(O)$ is $\alpha-$individually fair if and only if for all $u, v \in U$, $d(C(u), C(v)) \leq \alpha\mathcal{D}(u, v)$.*

We use the phrase "similar individuals are treated similarly" as a shorthand for the individual fairness Lipschitz condition. Individual fairness was originally proposed in the context of independent classification, i.e., each individual is classified exactly once, independently of all others. However, in many practical settings individuals cannot be classified independently, particularly when there are a limited number of positive classifications available (e.g., a university which can only accept a limited number of students each year, an advertiser with a limited budget). Dwork and Ilvento formalized this problem as the "cohort selection problem," in which a set of exactly n individuals must be selected such that the probabilities of selection conform to individual fairness constraints [6].

▶ **Definition 2** (Cohort Selection Problem [6]). *Given a universe U of individuals, an integer n, and a task with metric $\mathcal{D}$, select a cohort $C \subseteq U$ of exactly n individuals such that $|\Pr[u \in C] - \Pr[v \in C]| \leq \mathcal{D}(u, v)$. We call such a mechanism an individually fair cohort selection mechanism.*

Our work extends the investigation into fair composition by considering composition within a *pipeline* of cohort selection and scoring steps. We focus on the case of a two-step pipeline, and we assume for simplicity that the metric for the cohort selection and scoring functions are the same.

▶ **Definition 3** (Two-stage Cohort pipeline). *Given a universe of individuals U, a two-stage cohort pipeline consists of: a set of permissible cohorts $\mathcal{C} \subseteq \mathsf{Pow}(U) \backslash \emptyset$ (where $\mathsf{Pow}(U)$ indicates the power set of U), a single (randomized) cohort selection mechanism A which outputs a single cohort $C \subseteq \mathcal{C}$, a set of scoring functions $\mathcal{F} : \mathcal{C} \times U \to [0,1]$, and a scoring function $f \in \mathcal{F}$. The two-stage cohort pipeline procedure is $A \circ f$.*

Table 1 Terminology.

Term	Definition
U	The universe of individuals
$\mathcal{D} : U \times U \to [0, 1]$	The individual fairness metric
$\mathcal{C} \subseteq \mathsf{Pow}(U) \backslash \emptyset$	The set of permissible cohorts
$\mathcal{F}$	The family of permitted scoring functions.
$f : \mathcal{C} \times U \to [0, 1]$	A scoring function. $f(C, x)$ is *undefined* whenever $x \notin C$, and throughout this work, whenever we write $f(C, x)$, where x is any element in U, it is the case that $x \in C$.
$A : U \to \mathcal{C}$	An individually fair cohort selection mechanism.
$\mathbb{A}(C) \in [0, 1]$	The probability that A outputs the cohort C.
$\mathcal{C}_u \subseteq \mathcal{C}$	The subset of permissible cohorts containing u.
$p(u) \in [0, 1]$	The probability A outputs a cohort containing U.

We now briefly introduce supporting terminology (summarized in Table 1). For $C \in \mathcal{C}$, let $\mathbb{A}(C)$ denote the probability that A outputs C, where the probability is over the randomness in the cohort selection mechanism A operating on the universe U. We denote the set of cohorts containing u as $\mathcal{C}_u$, and the probability that A selects u can be expressed $p(u) = \sum_{C \in \mathcal{C}_u} \mathbb{A}(C)$. As initial constraints on A and $\mathcal{F}$, we assume that A is an individually fair cohort selection mechanism and that each $f \in \mathcal{F}$ is individually fair within the cohort it observes, i.e., it is *intra-cohort individually fair*:

▶ **Definition 4** (Intra-cohort individual fairness). *Given a cohort C, a scoring function $f : \mathcal{C} \times U \to [0, 1]$ is intra-cohort individually fair if for all $C \in \mathcal{C}$, $\mathcal{D}(u, v) \geq |f(C, u) - f(C, v)|$ for all $u, v \in C$.*

Although intra-cohort fairness constrains f to be individually fair *within* a particular cohort, $f(C_1, u)$ can differ arbitrarily from $f(C_2, u)$ if $C_1 \neq C_2$. For ease of exposition we sometimes refer to C as the "cohort context" or simply the "context" of u for $u \in C$.

▶ Remark 5 (Intra-cohort individual fairness is insufficient). A pipeline consisting of an individually fair cohort selection mechanism and intra-cohort individually fair scoring function may result in arbitrarily unfair treatment. For example, suppose $\mathcal{X} = \{X_1, X_2, \ldots\}$ is a partition of U, and A chooses a cohort X_i uniformly at random. Suppose f assigns score 1 to all members of the cohort corresponding to X_*, and otherwise assigns score 0. A is not only individually fair, it selects each element with an equal probability; f is not only intra-cohort individually fair, it treats all members of a given cohort equally; yet the pipeline can result in arbitrarily large differences in scores for similar individuals. Furthermore, this observation holds for *any* partition including adversarially chosen partitions. Although this abstract example suffices to prove the point, we include an extensive set of realistic pipeline examples, analogous to the "Catalog of Evils" of [5], in Appendix A. We also include a practical method for *malicious* pipeline construction in Appendix B of the full version.

An important part of the pipeline definition is the contextual behavior of f, i.e., the behavior of the second stage of the pipeline may depend on the selected cohort C. The simplistic solution to this problem is to design and evaluate the whole pipeline for fairness as a single unit, i.e., requiring that similar individuals have similar distributions over $\Delta(O_{pipeline})$. Although such evaluation would catch unfairness, it (1) doesn't provide explicit guidance for designing any given component, (2) may miss certain pipeline-specific fairness issues (see Examples 7 and 9), and (3) "locks" the pipeline into a single monolithic strategy, which is highly impractical. For example, employers frequently need to change compensation policies due to changing market conditions. However, changing compensation policies due to disliking a particular member of a cohort, e.g., switching to equal bonuses for all team members if the company does not like the individual who would have received the largest bonus, is not permitted in our model. Indeed, later stages in the pipeline may be completely ignorant of the existence of prior stages, e.g., a manager deciding on employee compensation may be unaware of automated resume screening.

This motivates our design goal of *robustness*: designing the cohort selection mechanism A which composes well with *every* function in $\mathcal{F}$, rather than expecting the scoring function to properly analyze and respond to the choices made in the original cohort selection mechanism design. As a result, the only communication necessary between the steps is the description of $\mathcal{F}$. With this in mind, a deceptively(!) simple extension of Definition 1 gives our fairness desideratum for pipelines.

▶ **Definition 6** (α-Individual Fairness and Robustness for Pipelines (informal)). *Consider the pipeline consisting of $(\mathcal{C}, A, \mathcal{F})$, with outcome space $O_{pipeline}$. For $f \in \mathcal{F}$, the pipeline instantiated with f satisfies $\alpha-$individual fairness with respect to the similarity metric $\mathcal{D}$ and a distance measure $d : \Delta(O_{pipeline}) \times \Delta(O_{pipeline}) \to [0,1]$ if $\forall u, v \in U$, $d([f \circ A](u), [f \circ A](v)) \leq \alpha\mathcal{D}(u, v)$.*

*If the pipeline satisfies $\alpha-$individual fairness with respect to every $f \in \mathcal{F}$, i.e., if $\forall f \in \mathcal{F}$ and $\forall u, v \in U$, $d([f \circ A](u), [f \circ A](v)) \leq \alpha\mathcal{D}(u, v)$, we say that A is $\alpha-$*robust *to $\mathcal{F}$ with respect to $d, \mathcal{D}$.*

We model the contextual nature of the problem by allowing the behavior of each $f \in \mathcal{F}$ to depend on the cohort, rather than allowing f to be chosen adaptively in response to the selected cohort. This modeling choice still allows us to capture the contextual nature of scoring policies, while keeping our abstractions clean.[5]

2.2 Fairness of pipelines

Lurking in this informal definition are two subtle choices critical to pipeline fairness: (1) how should distributions over $O_{pipeline}$ be interpreted, and (2) what distance measure d is appropriate for measuring differences in distributions over $O_{pipeline}$. In the remainder of this section, we consider these two questions and frame the notion of robustness parametrized by the two axes: distribution and distance measure over distributions.

2.2.1 Choosing the interpretation of the distribution

To account for the fact that individuals not selected by A never receive a score from f the relevant outcome space is the union of possible scores and "not selected," i.e., $O_{pipeline} := [0,1] \cup \{\perp\}$. Thus conditioning on whether an individual was selected or not changes the interpretation of the distribution over the outcome space and, more importantly, changes the *perception* of fairness.

▶ **Example 7** (Perception of conditional probability). Suppose Alice (a) and Bob (b) are similar but not equal job candidates, i.e., $\mathcal{D}(a, b) \in (0, 0.1]$. Consider an individually fair cohort selection mechanism, A which either selects a cohort containing one of Alice or Bob or neither and satisfies $p(a) = p(b) = p^*$. Consider the fairness constraint on the scoring function f for the unconditional distribution over $O_{pipeline}$: $|p(a)f(a) - p(b)f(b))| \leq \mathcal{D}(a, b)$, which

[5] See Appendix A for explicit examples of modeling adaptation to changing market conditions.

simplifies to $p^*|f(a) - f(b))| \leq \mathcal{D}(a, b)$. (Note: as Alice and Bob never appeared together in a cohort, there is no intra-cohort fairness condition.) The constraint on the difference in treatment by f is essentially diluted by a factor of p^*.

Enforcing fairness on the unconditional distribution essentially allows the company to hand out job offers of the following form: "Congratulations you are being offered a position at Acme Corp., you can expect a promotion after one year with probability x%." Alice and Bob may *receive* offers will equal probability, but the values of x printed on the offer may be wildly different, and as such they will perceive the value of the job offer differently.

The choice of conditional or unconditional distribution boils down to what perception of fairness is important. In the case of bonuses or promotions awarded long after hiring, the conditional perception may be particularly important. However, on shorter time frames or if the only consequential outcome is the final score, the unconditional distribution may be more appropriate (e.g., resume screening immediately followed by interviews).[6] We consider two approaches which capture these different perspectives: the **unconditional distribution** $S_u^{N,A,f}$, treats the $\perp$ outcome as a score of 0 and the **conditional distribution** $S_u^{C,A,f}$ conditions on u being selected in the cohort. More formally:

▶ **Definition 8** (Conditional and unconditional distributions)**.** *Let $S_u^{A,f} \in \Delta(O_{pipeline})$ be the distribution over outcomes arising from the pipeline, i.e., $f \circ A$. $S_u^{A,f}$ places a probability of $1 - p(u)$ on $\perp$, and for $s \in [0, 1]$, $S_u^{A,f}$ places a probability of $\sum_{C \in \mathcal{C}} \Pr[f(C, u) = s]A(C)$ on s.*

- *The **unconditional distribution** $S_u^{N,A,f}$ is identical to $S_u^{A,f}$ with the exception that it treats the $\perp$ outcome as if it had score 0. That is, for $0 < s \leq 1$, $S_u^{N,A,f}$ places a probability of $\sum_{C \in \mathcal{C}} \Pr[f(C, u) = s]\mathbb{A}(C)$ on s; at $s = 0$, $S_u^{N,A,f}$ has a probability of $1 - p(u) + \sum_{C \in \mathcal{C}} \Pr[f(C, u) = 0]\mathbb{A}(C)$.*
- *The **conditional distribution** $S_u^{C,A,f}$ has probability $\frac{\sum_{C \in \mathcal{C}} \Pr[f(C,u)=s]\mathbb{A}(c)}{p(u)}$ for each score $s \in [0, 1]$, i.e., it is $S_u^{A,f}$ conditioned on the positive outcome of $A(C)$.*[7]

Each of these approaches can be viewed as a method for converting a distribution $S_u^{A,f}$ over $O_{pipeline}$ to distributions $S_u^{C,A,f}$ and $S_u^{N,A,f}$ over $[0, 1]$.

2.2.2 Distance measures over distributions

The natural approach for measuring distances between distributions would be to use expectation: that is, $d^{uncond,\mathbb{E}}(S_u^{A,f}, S_v^{A,f}) := |\mathbb{E}[S_u^{N,A,f}] - \mathbb{E}[S_v^{N,A,f}]|$ and $d^{cond,\mathbb{E}}(S_u^{A,f}, S_v^{A,f}) := |\mathbb{E}[S_u^{C,A,f}] - \mathbb{E}[[S_v^{C,A,f}]|$. Difference in expectation generally captures the unfairness in the examples discussed thus far. However, a subtle issue can arise from the *certainty* of outcomes, which requires greater insight into the distribution of scores.

[6] Although in this work we consider pipelines with a single relevant metric, the conditional versus unconditional question is critically important when metrics differ between stages of the pipeline. For example, the metric for selecting qualified members of a team may be different than the metric for choosing an individual from the team to be promoted to a management role, as the two stages in the pipeline require different skillsets.

[7] This definition is not defined if $p(u) = 0$, since it does not make sense to consider a "conditional distribution" if u is never selected to be in the cohort (and thus never receives a score). In defining robustness of a cohort selection mechanism, we should thus restrict to considering $u \in U$ where $p(u) > 0$ (and individual fairness of the cohort selection mechanism on its own would provide fairness guarantees over the probabilities $p(u)$). For simplicity, we do not explicitly mention this modification.

▶ **Example 9** (Certainty of outcomes). Consider two equally qualified job candidates, Charlie and Danielle. As these two candidates are equally qualified, they should clearly be offered jobs and promotions with equal probability. Recall the company's pleasant form letter for job offers from Example 7, "Congratulations you are being offered a position at Acme Corp., you can expect a promotion after one year with probability $x\%$." Danielle receives an offer with $x = 70\%$ (with probability p^*), but Charlie receives either an offer with $x = 100\%$ (with probability $0.7p^*$) or an offer with $x = 0\%$ (with probability $0.3p^*$). Although both are offered jobs with equal probability and their expectations of promotion are equal, Charlie's offers have *certainty* of promotion (or no promotion) whereas Danielle's promotion fate is uncertain.

As Example 9 illustrates, expected score does not entirely capture problems related to the *distribution* of scores rather than the average score. Although total-variation distance is a natural choice for evaluating such distributional differences, it is too strong for this setting. For example, if Charlie receives a score of 0.7 with probability 1 (over randomness of the entire pipeline), while Danielle receives a score of $0.7 - \epsilon$ with probability 0.5 and a score of $0.7 + \epsilon$ with probability 0.5, then the total variation distance would be 1, though these outcomes are intuitively very similar. We therefore introduce the notion of *mass-moving distance* over probability measures. Mass-moving distance combines total variation distance with earthmover distance to reflect that similar individuals should receive similar distributions over close (rather than identical) sets of scores.

▶ **Definition 10** (Mass-moving distance). *Let γ_1 and γ_2 be probability mass functions over finite sets $\Omega_1 \subseteq [0,1]$ and $\Omega_2 \subseteq [0,1]$, respectively. Let $V \subseteq [0,1]$ be the set of real values $v \in [0,1]$ such that there exist probability mass functions $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ over $[0,1]$ with finite supports $\tilde{\Omega}_1$ and $\tilde{\Omega}_2$, respectively, where:*

1. *Nothing moves far and mass is conserved. For $i = 1, 2$, there is a function $Z_i : [0,1] \to \Delta(\tilde{\Omega}_i)$ such that:*
   a. *Nothing moves far. For all $x \in [0,1]$ and $y \in Supp(Z_i(x))$, it holds that $|x - y| \leq 0.5v$.*
   b. *Mass is conserved. For all $y \in \tilde{\Omega}_i$, it holds that $\tilde{\gamma}_i(y) = \sum_{x \in \Omega_i} z_i^x(y)\gamma_i(x)$, where z_i^x is the probability mass function of the distribution $Z_i(x)$.*
2. *Total variation distance is small. It holds that $0.5v \geq TV(\tilde{\gamma}_1, \tilde{\gamma}_2) := \frac{1}{2}\sum_{w \in \tilde{\Omega}_1 \cup \tilde{\Omega}_2} |\tilde{\gamma}_1(w) - \tilde{\gamma}_2(w)|$.*

Then we let $MMD(\gamma_1, \gamma_2) = \inf(V)$.

A simple way to think about mass-moving distance is to break the definition down into two steps: (1) transforming the original distributions over scores into distributions over a single shared set of *adjusted scores* and (2) moving mass between the distributions over adjusted scores.

Since there is a natural association between probability distributions over $[0,1]$ and probability mass functions over $[0,1]$, Definition 10 also gives a notion of distance between probability distributions.[8] In the example of Charlie and Danielle receiving scores of 0.7 or $0.7 \pm \varepsilon$ described above, the mass-moving distance is at most 2ϵ since $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ can both be taken to be the probability measure that places the full mass of 1 on 0.7.

[8] We slightly abuse notation and use $MMD(\mathcal{X}_1, \mathcal{X}_2)$ for probability distributions $\mathcal{X}_1$ and $\mathcal{X}_2$, to denote $MMD(\gamma_1, \gamma_2)$ where γ_1 is the probability mass function associated to $\mathcal{X}_1$ and γ_2 is the probability mass function associated to $\mathcal{X}_2$.

Using mass-moving distance, we specify two additional complementary distance measures $d^{cond,MMD}(S_u^{A,f}, S_v^{A,f}) := MMD(S_u^{C,A,f}, S_v^{C,A,f})$ and $d^{uncond,MMD}(S_u^{A,f}, S_v^{A,f}) := MMD(S_u^{N,A,f}, S_v^{N,A,f})$.

2.3 Robustly fair pipelines

Recall our informal notion that a cohort selection mechanism A is robust to a family of scoring functions $\mathcal{F}$ if the composition of A and any $f \in \mathcal{F}$ is individually fair. We can now formalize robustness as either **conditional** or **unconditional** with respect to either **expected score** or **mass moving distance** over score distributions. By evaluating the properties of each combination of distribution and distance measure, we can capture a range of subtle fairness desiderata in pipelines.[9]

▶ **Definition 11** (Robust pipeline fairness). *Given a universe U, a metric $\mathcal{D}$, let A be an individually fair cohort-selection mechanism and let $\mathcal{F}$ be a collection of intra-cohort individually fair scoring functions $\mathcal{C} \times U \to [0,1]$. Choose $d \in \{d^{cond,\mathbb{E}}, d^{uncond,\mathbb{E}}, d^{cond,MMD}, d^{uncond,MMD}\}$, a distance measure over $S_u^{A,f}$. We say A is α-**robust w.r.t $\mathcal{F}$** for d if $d(S_u^{A,f}, S_v^{A,f}) \leq \alpha\mathcal{D}(u,v)$ for all $u, v \in U$ and for all $f \in \mathcal{F}$.*

Throughout the rest of this work, we will examine robustness properties in terms of particular settings of d. As one might expect, mass moving distance over score distributions is a stronger condition than expected score, and conditional robustness implies unconditional robustness up to a Lipschitz relaxation.[10]

3 Conditions for Success

In this section, we describe conditions on A that will result in our desired robustness properties with respect to a class of scoring functions $\mathcal{F}$. We first consider the description of $\mathcal{F}$ available to A, i.e., the policy. The simplest method of specifying the policy by describing all $f \in \mathcal{F}$ prohibits adding f with similar or identical fairness properties to $\mathcal{F}$ at a later point and is highly unrealistic (and potentially intractable). In practice, we expect policies to govern how differently f can treat individuals within different contexts, rather than enumerating the permitted functions. To that end, we propose policies in the form of a *distance function over (cohort, individual) pairs*, $\delta^{\mathcal{F}} : (\mathcal{C} \times U) \times (\mathcal{C} \times U) \to [0,1]$. This distance function specifies the maximum difference in score between two (cohort, individual) pairs $\delta^{\mathcal{F}}((C_1,u),(C_2,v)) := \sup_{f\in\mathcal{F}} |f(C_1,u) - f(C_2,v)|$. $\delta^{\mathcal{F}}$ captures the salient fairness behavior of the family of scoring functions, while being succinct in comparison to maintaining a list of all supported f directly. In fact, as we will show in Lemma 13, a partial description or an overestimate of $\delta^{\mathcal{F}}$ will also suffice. To illustrate our policy descriptions, consider the following two families:

1. $\mathcal{F}_1$ ignores the cohort context entirely, and treats each $u \in U$ the same regardless of the cohort, i.e., $\mathcal{F}_1 = \{f \mid \exists g : U \to [0,1] \text{ s.t. } f(C,u) = g(u) \text{ for all } (C,u) \in \mathcal{C} \times U\}$.
2. $\mathcal{F}_2$ treats u and v similarly within the same context, but has no constraint on treatment in different contexts, i.e., $\mathcal{F}_2 = \{f \mid f((C\backslash\{u\}) \cup \{v\}, v) - f(C,u)| \leq \mathcal{D}(u,v) \text{ for all } u, v \in U$ and $\forall C \in \mathcal{C}$ s.t. $u \in C, v \notin C\}$.

[9] Note that these choices for d are not the only possible choices, and the framework can be extended to different choices of distribution and distance measure to address other fairness concerns.

[10] See Propositions E.2 and E.1 in the full version. Interestingly, we show in the full version that for some classes of score functions, guaranteeing individual fairness w.r.t mass-mover distance fairness is “equivalent” to guaranteeing individual fairness w.r.t expected score.

Recall that intra-cohort individual fairness requires that the scoring functions in both families must treat u and v similarly if they appear in the same cohort, i.e., $\mathcal{D}(u,v) \geq |f(C,u) - f(C,v)|$.

For the family $\mathcal{F}_1$, we observe that $\delta^{\mathcal{F}_1}((C_1,u),(C_2,v)) = \mathcal{D}(u,v)$, and, intuitively, the designers of A will not need to consider the behavior of $\mathcal{F}$ in their design of A. On the other hand, for $\mathcal{F}_2$, we observe that $\delta^{\mathcal{F}_2}((C,u),(C,v)) = \mathcal{D}(u,v)$ for any cohort C, but $\delta^{\mathcal{F}_2}((C,u),(C',v))$ can be much greater than $\mathcal{D}(u,v)$ for $C' \neq C$. For this reason, composition planning for A is non-trivial. As one would expect, $\delta^{\mathcal{F}}$ heavily influences the strength of conditions on A.

3.1 A's Task: Designing Mechanisms Compatible with $\delta^{\mathcal{F}}$

We now describe how to design A to guarantee robustness with respect to $\mathcal{F}$, given (possibly overestimates of) the distance function $\delta^{\mathcal{F}}$ over (cohort, individual) pairs describing $\mathcal{F}$. The conditions on A will roughly consist of making sure that A assigns *similar individuals to similar distributions over cohort contexts*, where similarity of (cohort, individual) pairs is defined with respect to $\delta^{\mathcal{F}}$.

Although $\delta^{\mathcal{F}}$ is a succinct description of a policy, it is more intuitive when designing with composition in mind to translate $\delta^{\mathcal{F}}$ into a set of "mappings" specifying which (cohort, individual) pairs will be treated similarly by $f \in \mathcal{F}$. That is, for each pair $u, v \in U$, we can describe $\delta^{\mathcal{F}}$ as a partitioning $\mathcal{P}_{u,v}$ of $(\mathcal{C}_u \times u) \cup (\mathcal{C}_v \times v)$ such that each partition or "cluster" has small diameter with respect $\delta^{\mathcal{F}}$, i.e., within a cluster $\delta^{\mathcal{F}}((C_1,u),(C_2,v)) \leq \mathcal{D}(u,v)$. The collection of partitions over all pairs of individuals then defines the mapping.

▶ **Definition 12** (Mapping based on δ)**.** *For each pair of distinct individuals u and v, consider the subset $\mathcal{P}_{u,v} := (\mathcal{C}_u \times \{u\}) \cup (\mathcal{C}_v \times \{v\})$ of (cohort, individual) pairs. Consider a partition of $\mathcal{P}_{u,v}$ into clusters that respects δ, i.e., that satisfies the following condition: if $(C_1,x),(C_2,y)$ are in the same cluster*[11]*, then $\delta((C_1,x),(C_2,y)) \leq \mathcal{D}(u,v)$. Let $n_{u,v}$ (and $n_{v,u}$) be the number of clusters of the partition. We call a collection of such partitions for each pair $u, v \neq U$ a* ***mapping*** *of $\mathcal{C}$ that* ***respects $\boldsymbol{\delta}$****.*

Mappings interact well with distance functions δ' that overestimate $\delta^{\mathcal{F}}$, as larger distances between (cohort, individual) pairs imposes more strict conditions on cluster membership. Lemma 13 states that a mapping that respects δ' will also respect $\delta^{\mathcal{F}}$, although the resulting conditions on the mapping might be more restrictive.

▶ **Lemma 13.** *Let $\delta' : (\mathcal{C} \times U) \times (\mathcal{C} \times U) \to [0,1]$ be a distance function. Suppose that δ' has the property that for all pairs of cohort contexts $(C_1,x),(C_2,y) \in \mathcal{C} \times U$, it holds that $\delta'((C_1,x),(C_2,y)) \geq \delta^{\mathcal{F}}((C_1,x),(C_2,y))$. If a mapping respects δ', then the mapping also respects $\delta^{\mathcal{F}}$.*

Proof. Consider any pair of individuals u and v, and consider any mapping that respects δ'. In the partition corresponding to u and v, if (C_1,x) and (C_2,y) are in the same cluster, then it holds that $\delta^{\mathcal{F}}((C_1,x),(C_2,y)) \leq \delta'((C_1,x),(C_2,y)) \leq \mathcal{D}(u,v)$. Thus, the mapping respects $\delta^{\mathcal{F}}$, as desired. ◀

[11] Note that $x, y \in \{u,v\}$. Recall that (C_1,u) and (C_2,u) may appear in the same cluster, and thus it is possible that $x = y$.

Table 2 Policy and mapping terminology.

Term	Definition
$\delta^{\mathcal{F}} : (\mathcal{C} \times U) \times (\mathcal{C} \times U) \to [0,1]$.	distance function specifying the maximum difference in treatment between (cohort,individual) pairs by any $f \in \mathcal{F}$. $\delta^{\mathcal{F}}((C_1,u),(C_2,v))$ is undefined if $u \notin C_1$ or $v \notin C_2$.
$M_{u,v} : \mathcal{C}_u \to \mathbb{N}$	a mapping of the cohorts containing u to clusters containing (C,u).
$n_{u,v}$	The number of clusters in a mapping
$\mathcal{M}_\delta$	the set of all mappings which respect δ.

We now briefly introduce supporting terminology for policies and mappings (summarized in Table 2). To succinctly refer to the clusters in a mapping, we define label functions $M_{u,v} : \mathcal{C}_u \to \mathbb{N}$ and $M_{v,u} : \mathcal{C}_v \to \mathbb{N}$ such that $M_{u,v}(C)$ is the label of the cluster containing (C,u) and $M_{v,u}(C)$ is the label of the cluster containing (C,v). We use $n_{u,v}$ (or $n_{v,u}$) to denote the number of clusters in a mapping. We also refer to the set of functions $(M_{u,v})_{u \neq v \in U}$, which entirely specify the partitions, as a mapping. Valid mappings for δ are not necessarily unique, as there may be more than one way to partition $\mathcal{P}_{u,v}$ into clusters with diameter bounded by $\mathcal{D}(u,v)$. We let $\mathcal{M}_\delta$ be the set of mappings that respect δ.

Given a mapping of $\delta^{\mathcal{F}}$ (or of an overestimate δ'), we can now interpret "distributions over cohorts" induced by A as "distributions over clusters" induced by A. Formally, we convert the distributions over cohorts into measures over $[n_{u,v}]$ for each pair $(u,v) \in U \times U$. As a result, "similar distributions over cohorts" will turn out to mean "similar measures over $[n_{u,v}]$."

▶ **Definition 14.** *Let $(M_{u,v})_{u \neq v \in U}$ be a mapping of $\mathcal{C}$. For $u, v \in U$, we define measures $q^1_{u,v}$ and $q^2_{u,v}$ over the sample space $[n_{u,v}]$ as follows:*

1. *The* ***unconditional measure over cohorts*** $\mathbf{q^1_{u,v}}$ *on the sample space $[n_{u,v}]$ for each (u,v) ordered pair is defined as follows. For $i \in [n_{u,v}]$, we let $q^1_{u,v}(i) = \sum_{C \in \mathcal{C}_u | M_{u,v}(C)=i} \mathbb{A}(C)$.*[12]
2. *The* ***conditional measure over cohorts*** $\mathbf{q^2_{u,v}}$ *on the sample space $[n_{u,v}]$ for each (u,v) ordered pair is defined as follows. For $i \in [n_{u,v}]$, we let $q^2_{u,v}(i) = \frac{\sum_{C \in \mathcal{C}_u | M_{u,v}(C)=i} \mathbb{A}(C)}{p(u)}$.*[13][14]

We now specify sufficient conditions for robustness in terms of distances between these measures over $[n_{u,v}]$. The conditions require that for each pair $u, v \in U$, A assigns similar probabilities to cohorts containing u and cohorts containing v within each cluster.

▶ **Definition 15** (α-Notions 1 and 2)**.** *Let $(M_{u,v})_{u \neq v \in U}$ be a mapping of $\mathcal{C}$. For $u, v \in U$, let $q^1_{u,v}$ and $q^2_{u,v}$ be defined as in Definition 14. We define α-Notions 1 and 2 as follows:*

1. *For $\alpha \geq 0.5$, we say that A satisfies α-**Notion 1** if for all $u, v \in U$ such that $\mathcal{D}(u,v) < 1$, $TV(q^1_{u,v}, q^1_{v,u}) \leq (\alpha - 0.5)\mathcal{D}(u,v)$. (The 0.5 arising in Notion 1 comes from having to "smooth out" $q^1_{u,v}$ to a probability measure in a later step.)*
2. *For $\alpha \geq 0$, we say that A satisfies α-**Notion 2** if for all $u, v \in U$ such that $\mathcal{D}(u,v) < 1$, $TV(q^2_{u,v}, q^2_{v,u}) \leq \alpha\mathcal{D}(u,v)$.*

---

[12] This is not necessarily a probability measure, since the total sum on the sample space is $p(u) \leq 1$, but it is finite.

[13] Observe that this is in fact a probability measure since $p(u) = \sum_{C \in \mathcal{C}_u} \mathbb{A}(C) = \sum_{i=1}^{M_{u,v}} \sum_{C \in \mathcal{C}_u | M_{u,v}(C)=i} \mathbb{A}(C)$.

[14] Like in Definition 8, this definition is not defined if $p(u) = 0$, since it does not make sense to consider a "conditional distribution" if u is never selected to be in the cohort (and thus never receives a score). We should thus restrict to considering $u \in U$ where $p(u) > 0$ (and individual fairness of the cohort selection mechanism on its own would provide fairness guarantees over the probabilities $p(u)$). For simplicity, in this extended abstract, we do not explicitly mention this modification.

Our main result is that these conditions guarantee pipeline robustness for composition with any $f \in \mathcal{F}$ with respect to mass-moving distance (and thus expected score).[15] Theorem 16 states that so long as A satisfies Notion 1 (resp. 2) for the mappings associated with $\mathcal{F}$, then A will be robust with respect to $\mathcal{F}$.

▶ **Theorem 16** (Robustness to Post-Processing). *Let $\mathcal{F}$ be a class of scoring functions, let $\alpha \geq 0.5$ be a constant. Suppose that $(M_{u,v})_{u \neq v \in U}$ is in $\mathcal{M}_{\frac{1}{2\alpha}\delta^{\mathcal{F}}}$. If A is individually fair and satisfies α-Notion 1 (resp. α-Notion 2) for $(M_{u,v})_{u \neq v \in U}$, then A is 2α-robust w.r.t. $\mathcal{F}$ for $d^{uncond,MMD}$ (resp. $d^{cond,MMD}$).*

The proof of Theorem 16 appears in Appendix B.1 of the full version.

Furthermore, these conditions are necessary both for mass-moving distance and the weaker condition of expected score for certain rich classes of scoring functions.

▶ **Theorem 17** (Informal). *Let d be any metric in $\{d^{uncond,MMD}, d^{cond,MMD}, d^{cond,\mathbb{E}}, d^{uncond,\mathbb{E}}\}$. Loosely speaking, given $\mathcal{F}$ described by mappings such that inter-cluster distances are much larger than intra-cluster distances, the requirements on A in Theorem 16 are* ***necessary*** *for achieving robustness w.r.t. d.*

We formalize Theorem 17 in Appendix B of the full version.[16]

4 Robust Mechanisms

Although the conditions specified in the previous section are quite strict, and indeed some pathological scoring function families admit no robust solutions, we can nonetheless construct robust cohort selection mechanisms for rich classes of scoring policies.[17] We exhibit mechanisms robust to two broad classes of policies:

1. **Individual interchangeability:** replacing a single individual in the cohort does not change treatment of the cohort too much, i.e., policies like $\delta^{\mathcal{F}_2}$.
2. **Quality-based treatment:** cohorts with similar quality "profiles" are treated similarly. That is, the scoring function only considers the set of qualifications represented within a cohort and is agnostic to the specific individual(s) exhibiting a given qualification.

These policies cover a wide range of realistic scenarios and allow for significant flexibility and adaptability in the choice of f. In this section, we demonstrate that these policies also admit a variety of efficient and *expressive* constructions for A, i.e., A that may assign a wide range of probabilities $p(u)$ to individuals.

▶ Remark 18. As previously noted, robustness is trivial for the class of scoring functions which ignore the cohort context ($\mathcal{F}_1$). We formalize this observation in the following proposition:

▶ **Proposition 19.** *Consider the mapping that, for each pair of individuals u and v, places all of the cohort contexts in $(\mathcal{C}_u \times \{u\}) \cup (\mathcal{C}_v \times \{v\})$ into the same cluster. If A is individually fair, then A satisfies 0.5-Notion 1 and 0.5-Notion 2 w.r.t. this mapping.*

[15] See Corollary B.1.1 in the full version for a formal statement of the relationship between MMD and expected score.
[16] See Theorem B.5 and Theorem B.6.
[17] See Appendix D.1 of the full version for an example of $\mathcal{F}$ which admits no robust A.

4.1 Individual interchangeability

To describe the interchangeability policy, we specify a distance function $\delta^{\mathsf{int}} : (\mathcal{C} \times U) \times (\mathcal{C} \times U) \to [0,1]$ that requires that "swapping" any individual in a cohort does not result in significantly different treatment. More formally:

▶ **Definition 20** (Individual interchangeability policy).

$$\delta^{\mathsf{int}}((C,u),(C',v)) = \begin{cases} \mathcal{D}(u,v) & \text{if } C = C' \\ \mathcal{D}(u,v) & \text{if } C' = (C \setminus \{u\}) \cup \{v\} \,. \\ 1 & \text{otherwise.} \end{cases}$$

δ^{int} can be viewed as an overestimate of $\delta^{\mathcal{F}_2}$, or as a partial specification of the distance function on a subset of $(\mathcal{C} \times U) \times (\mathcal{C} \times U)$, trivially completed to 1 on other pairs of cohort context pairs. δ^{int} is naturally translated into a simple mapping: for any pair of individuals u and v, the partition corresponding to u and v in the mapping consists of clusters of size 2 consisting of "corresponding" (cohort, individual) pairs. This follows from observing that if an individual u receives some score $f(C,u)$ in a cohort C, if u were replaced by $v \notin C$, then v would receive a score in $[f(C,u) - \mathcal{D}(u,v), f(C,u) + \mathcal{D}(u,v)]$. More formally:

▶ **Definition 21** (Swapping Mapping). *Let $\mathcal{C}$ be the set of all subsets of U with exactly k individuals. The* ***swapping mapping*** *is defined as follows. For each pair of individuals $u, v \in C$:*

1. *For $C \in \mathcal{C}$ such that $u, v \in C$, the partition includes the cluster $\{(C,u),(C,v)\}$.*
2. *For $C \in \mathcal{C}$ such that $u \in C$ and $v \notin C$, the partition includes the cluster $\{(C,u),((C \setminus \{u\}) \cup \{v\}, v)\}$.*

It is straightforward to verify that the swapping mapping respects δ^{int}.

For the swapping mapping, there is a simple condition under which cohort selection mechanisms satisfy unconditional robustness (Notion 1): monotonicity.

▶ **Definition 22** (Monotonic cohort selection). *Suppose that $\mathcal{C}$ is the set of cohorts of size k. A cohort selection mechanism A is* ***monotonic*** *if for all pairs of individuals $u, v \in U$, for any $C' \subseteq U$ such that $|C'| = k-1$ and $u, v \notin C'$, if $p(u) \le p(v)$ then $\mathbb{A}(C' \cup \{u\}) \le \mathbb{A}(C' \cup \{v\})$.*

The intuition for the link between the monotonicity property and the swapping mapping is that the probability masses on a cohort containing u and a cohort containing v that are paired in the swapping mapping are directionally aligned and cannot diverge by more than $\mathcal{D}(u,v)$.

▶ **Lemma 23.** *Suppose that $\mathcal{C}$ is the set of cohorts of size k. If A is monotonic, then A satisfies* 0.5*-Notion 1 for the swapping mapping.*

Both PermuteThenClassify and WeightedSampling, cohort selection mechanisms proposed in [6], are monotonic, efficient and have a high degree of expressivity.[18]

However, monotonicity alone is not sufficient to guarantee conditional robustness (Notion 2) for the swapping mapping (see Appendix B of the full version). Borrowing intuition from PermuteThenClassify, we give a novel, efficient, individually fair cohort selection mechanism that achieves conditional robustness (Notion 2) for the swapping mapping:

[18] See Appendix B of the full version for detailed descriptions of these mechanisms and formal proofs of the monotonicity property.

▶ **Mechanism 24** (Conditioning Mechanism)**.** *Given a target cohort size k, a universe U and a distance metric $\mathcal{D}$, initialize an empty set S. For each individual $u \in U$:*

1. *Assign a weight $w(u)$ such that $|w(u) - w(v)| \leq \mathcal{D}(u, v)$, i.e., the weights are individually fair.*
2. *Draw from $\mathbb{1}_u \sim Bern(w(u))$, (i.e., flip a biased coin with weight $w(u)$). If $\mathbb{1}_u$, add u to S.*

If $|S| \geq k$, return a uniformly random subset of S of size k.[19] *Otherwise, repeat the mechanism.*

We show that under mild conditions, the Conditioning Mechanism is satisfies Notion 2, concludes in a small number of rounds, and allows for a high degree of expressivity. (See Appendix D of the full version for a formal statement and proof details.)

4.2 Quality-based treatment

One downside of the monotonic mechanisms proposed for δ^{int} is that they require that any cohort with a single individual swapped is considered with nearly the same probability as the original cohort. In practice, this is problematic when A needs to ensure that each cohort has a certain structure. For example, when hiring a team of software engineers, designers and product managers, the proportion of each type of team member is important, and arbitrary swaps are not desirable from the perspective of team structure. By restricting to scoring functions that only consider the quality profile of a cohort, i.e., how many individuals from each quality group are represented in a cohort, A can construct highly *structured* cohorts, so long as the structure of the cohort is valid with respect to the fairness metric $\mathcal{D}$.

We now consider robust mechanisms for policies predicated on additional structure within the metric over U. In particular, we assume the existence of a partition of the universe U into one or more "quality groups" $q_1, \ldots, q_n$. These quality groups satisfy the property that the distances within a quality group are smaller than distances between quality groups. *How much* smaller is determined by a parameter β. More formally,

▶ **Definition 25.** *Let $\beta \leq 1$ be a constant and $n \geq 1$ be an integer. Consider a partitioning of a U into subsets $q_1, \ldots, q_n$, i.e., "quality groups", and let $\mathcal{D}^*$ be a metric on U. Now, we define metrics D on $\{1, \ldots, n\}$ and $\mathcal{D}^i$ for $1 \leq i \leq n$ on q_i as follows: we let $D(i, j) = \inf_{u \in q_i, v \in q_j} \mathcal{D}^*(u, v)$ and $\mathcal{D}^i$ be the restriction of $\mathcal{D}^*$ to q_i. We call the metric $\mathcal{D}^*$ endowed with quality groups $q_1, \ldots, q_n$* ***β-quality-clustered*** *if for all $1 \leq i \leq n$, we have that*

$$\max_{u,v \in q_i} \mathcal{D}^i(u, v) \leq \beta \min_{j \neq i} D(i, j).$$

Notice that any metric $\mathcal{D}^*$ is trivially 1-clustered with respect to the trivial quality group $q_1 = U$. The benefit of endowing $\mathcal{D}^*$ with a greater number of quality groups is to exploit additional structure of the metric, when any exists.

For simplicity in the specification of the relevant policy and family of scoring functions we introduce a **quality profile function** P to count the number of individuals in each quality group in a cohort: that is, $P : 2^U \to \{(x_1, \ldots, x_n) \mid x_i \in \mathbb{Z}^{\geq 0}\}$, and the ith coordinate of

[19] One might imagine a mechanism that conditions on exactly k individuals being chosen, but this mechanism can be arbitrarily far from individually fair. Consider $k-1$ individuals with weight 1 and $|U| - k - 1$ individuals with weight 0.9. Conditioning exactly k individuals would cause $|p(u) - p(v)|$ to diverge arbitrarily for $w(u) = .9$ and $w(v) = 1$.

$P(C)$ is $|C \cap q_i|$. Loosely speaking, the quality-based treatment policy requires that the only information about a cohort utilized by the scoring functions is its quality profile. We now formally define $\mathcal{F}_3$ and an associated policy $\delta^{\mathsf{quality}}$:

▶ **Definition 26.** *Let $\beta \leq 1$ be a constant. Suppose that $\mathcal{D}$ is endowed with quality groups $q_1, \ldots, q_n$ and $\mathcal{D}$ is β-quality-clustered. We define $\mathcal{F}_3$ to be the set of intra-cohort individually fair score functions $f : \mathcal{C} \times U \to [0,1]$ satisfying the following conditions:*

1. *For $C, C' \in \mathcal{C}$ satisfying $P(C) = P(C')$, if u and v that are in the same quality group, then $f(C,u) = f(C',v)$.*
2. *For integers $1 \leq i \neq j \leq n$, $C, C' \in \mathcal{C}$ satisfying $P(C) = P(C')$, and any individuals $u \in q_j$ and $v \in q_j$, it holds that $|f(C,u) - f(C',v)| \leq D(i,j)$.*

When each quality group is homogeneous in terms of individual "quality", this corresponds to score functions that are determined purely by "quality".[20] As in Section 4.1, we specify a distance function $\delta^{\mathsf{quality}} : (\mathcal{C} \times U) \times (\mathcal{C} \times U) \to [0,1]$ that overestimates $\delta^{\mathcal{F}_3}$, but still preserves enough of the fairness structure to construct the desired mapping.

▶ **Definition 27** (Quality-based treatment policy)**.** *Given a universe U, a set of permissible cohorts $\mathcal{C}$ and distance metrics and quality groups as in Definition 26,*

1. *For $C, C' \in \mathcal{C}$ satisfying $P(C) = P(C')$, if $u \in q_j$ and $v \in q_j$, then $\delta^{\mathsf{quality}}((C,u),(C',v)) = 0$.*
2. *For integers $1 \leq i \neq j \leq n$, $C, C' \in \mathcal{C}$ satisfying $P(C) = P(C')$, and any individuals $u \in q_j$ and $v \in q_j$, we set $\delta^{\mathsf{quality}}((C,u),(C',v)) = D(i,j)$.*

The core intuition is that a nice mapping exists when $\mathcal{C}$ is "symmetric with respect to individuals in each quality group." It is helpful here to consider a bipartite graph $G = (A, B, E)$, where A has one vertex for each subset of the universe U, B has one vertex for each possible profile of a subset of U, and there is an edge $(a,b) \in E$ precisely when b is the profile of a, that is $b = P(a)$.

Fix any $\mathcal{C}$, and consider the subgraph $G' = (A', B', E')$ of G induced by the vertices in A corresponding to members of $\mathcal{C}$, the edges incident on these vertices, and the subset of B induced by these edges. We say that $\mathcal{C}$ is **quality-symmetric** if for all $b' \in B'$ it is the case that E' contains all the edges in E (in the original graph) incident on b'.

That is, $\mathcal{C}$ contains all cohorts obtained by swapping out individuals from the same quality group. If $\mathcal{C}$ is quality-symmetric, then consider the following mapping.

▶ **Definition 28** (Quality-Based Mapping)**.** *Let $\beta \leq 1$ be a constant. Suppose that $\mathcal{D}$ is endowed with quality groups $q_1, \ldots, q_n$ and $\mathcal{D}$ is β-quality-clustered. Suppose $\mathcal{C}$ is quality-symmetric. The **quality-based mapping** is defined as follows. For each pair of individuals $u, v \in C$, let $\mathcal{P}_{u,v} = (\mathcal{C}_u \times \{u\}) \cup (\mathcal{C}_v \times \{v\})$. For each $(x_1, \ldots, x_n) \in P(\mathcal{C}_u \cup \mathcal{C}_v)$, the partitioning of $\mathcal{P}_{u,v}$ contains a cluster of the form $\{(C,x) \in \mathcal{P}_{u,v} \mid P(C) = (x_1, \ldots, x_n)\}$.*

We verify that the quality-based mapping indeed respects $\delta^{\mathsf{quality}}$ (and thus respects $\delta^{\mathcal{F}_3}$ by Lemma 13). If u and v are in the same quality group, then the diameter of each cluster under $\delta^{\mathsf{quality}}$ is 0, which is trivially upper bounded by $\mathcal{D}(u,v)$. On the other hand, if u and v are in different quality groups q_i and q_j respectively, then the diameter of each cluster is no more than $D(i,j) \leq \mathcal{D}(u,v)$. Thus, the properties of a mapping are satisfied by the quality-based mapping.

[20] In this case, $\mathcal{F}_3$ includes Equal Treatment, Promotion, Stack Rank, and Fixed Bonus (discussed in Appendix A) when scores are based on the "quality" of "performance" of individuals.

In this scenario, the quality-based *mapping* captures the intuition for the fairness structure of $\mathcal{F}_3$ much better than $\delta^{\mathsf{quality}}$. The mapping groups together all cohorts with the same quality profile (i.e., the same number of individuals in each quality group), capturing the intuition that the only information that a score function in $\mathcal{F}_3$ utilizes about a cohort is the quality profile.

As the score function behavior does not depend on the specific individuals in a quality group, A should have significant freedom to choose individuals within each quality group while still satisfying robustness w.r.t. $\mathcal{F}_3$. We will show that once the number of members of each quality group in the cohort is decided, utilizing any individually fair cohort selection mechanism within each quality group will satisfy our conditions. Moreover, our mechanisms have some flexibility in deciding the quality profile as well.

▶ **Mechanism 29** (Quality Compositional Mechanisms)**.** *Let $\beta \leq 1$ be a constant, and suppose that $\mathcal{D}$ endowed with quality groups $q_1, \ldots, q_n$ is β-quality-clustered. Suppose also that $\mathcal{C}$ is quality-symmetric. For each $1 \leq i \leq n$ and each $1 \leq x_i \leq |q_i|$, let A_{i,x_i} be a $\mathcal{D}^i$-individually fair mechanism selecting x_i individuals in q_i. We define the **quality compositional mechanism** for $\{A_{i,x_i}\}$ as follows. Let $\mathcal{X}$ be any distribution over n-tuples of nonnegative integers $(x_1, \ldots, x_n) \in P(\mathcal{C})$.*

1. *Draw $(x_1, \ldots, x_n) \sim \mathcal{X}$.*
2. *Independently run A_{i,x_i} for each $1 \leq i \leq n$, and return the union of the outputs of all of these mechanisms.*

In the next lemma, we show that when a quality composition mechanism only selects cohorts whose quality projection vectors $(x_1, \ldots, x_n)$ are "close" to an inter-quality group distance multiple of $(|q_1|, \ldots, |q_n|)$, Notion 1 is achieved. (This requirement essentially says that the relative proportion of selected individuals in each quality group needs to be approximately reflective of the relative proportion of individuals in each quality group in the universe, scaled by the difference between the quality groups in the original metric. This type of requirement turns out be necessary for basic individual fairness guarantees, by the constrained cohort impossibility result in [6].) Moreover, under stronger conditions, we show that Notion 2 is also achieved.

▶ **Lemma 30.** *Let $\beta \leq 0.5$ be a constant, and suppose that $\mathcal{D}$ endowed with quality groups $q_1, \ldots, q_n$ is β-quality-clustered. Suppose also that $\mathcal{C}$ is quality-symmetric, and let $\mathcal{X}$ be any distribution over $(x_1, \ldots, x_n) \in P(\mathcal{C})$ such that $|\frac{x_i}{|q_i|} - \frac{x_j}{|q_j|}| \leq (1-2\beta)D(i,j)$. If A is a quality compositional mechanism, then:*

1. *A is always individually fair.*
2. *A always satisfies 0.5-Notion 1.*
3. *A satisfies 0.5-Notion 2 for $\mathcal{D}$ and $\delta^{\mathcal{F}}$ if **either** of the following conditions hold:*
   a. *(One set) $|Supp(\mathcal{X})| = 1$ (i.e., one "canonical" $(x_1, \ldots, x_n)$), or*
   b. *(0-1 metric) $D(i,j) = 1$ for $1 \leq i \neq j \leq n$ and $\mathcal{D}^i(u,v) = 0$ for $1 \leq i \leq n$.*

The quality compositional mechanisms provide a greater degree of structure in cohort selection than the monotone mechanisms giving in Section 4.1. The Conditioning Mechanism and similar monotone mechanisms are forced to select individuals essentially independently, with the only dependence stemming from the cohort size constraint. However, structured cohorts are necessary in a number of practical applications, as previously noted. Although $\delta^{\mathsf{quality}}$ imposes more constraints on the permitted $\mathcal{F}$ than δ^{int}, the basis for these constraints is likely to be tolerated well in legitimate use cases in which structure is important.

Moreover, the company has flexibility in selecting individuals within each experience group, as any individually fair mechanism can be utilized. This offers significantly more flexibility than selecting members in each quality group uniformly at random. Such flexibility is particularly crucial, for example, if a company further wants to ensure that tech company teams have a mixture of software engineers and product managers. The individually fair mechanisms within each quality group can help achieve this balance through selecting balanced subsets of engineers and product managers. In essence, the quality compositional mechanisms allow flexibility in cohort selection while still satisfying robustness for $\mathcal{F}_3$, due to restrictions on the behavior of scoring functions in $\mathcal{F}_3$.

5 Discussion and Future Work

We have presented a framework for evaluating the robustness of cohort selection as part of a pipeline. We've demonstrated that naive auditing strategies concerning average cohort quality or score are unable to uncover significant fairness problems. We've also shown that many reasonable policies for cohort selection and subsequent scoring can conflict with each other resulting in very poor fairness outcomes. Furthermore, we've demonstrated that a malicious pipeline designer can easily use composition problems to disguise bad behavior. Despite these hurdles, we've shown that it is possible to construct pipelines that are fair. In particular we've shown that constructing cohort selection mechanisms that are robust to composition with a family of scoring functions is possible. By framing the problem in terms of robustness, we address the concern that placing requirements on future designs is nearly unenforceable, whereas designing the current stage to be robust to a large class of potential future policies can give much better practical guarantees. Finally, we've shown robust cohort selection mechanisms that compose well with reasonable scoring function families.

In the process of exploring robustness and fairness in pipelines, we uncovered a number of interesting questions for future work. **Policy complexity**: we have considered a set of concise and practical policies in this work, but the trade-off between policy complexity and the expressiveness of cohort selection has not been fully characterized. **Fair Matching**: choosing a cohort is very similar to the problem of assigning an individual to an existing cohort. However, in the traditional matching literature, significant emphasis is placed on individuals' and teams' preferences over placements, rather than external fairness criteria. Is it possible to simultaneously achieve a good matching, in the sense of satisfying preferences or stability, and individual fairness? **Quantifying the tradeoff**: There are significant differences in the difficulty for constructing mechanisms which satisfy the conditional, versus unconditional, notion of robustness. Is it possible to more directly quantify the tradeoff in mechanism expressivity between these two settings? **Different metrics**: Handling different metrics in the pipeline: we considered just one metric throughout the entire pipeline, but using different metrics for different stages of the pipeline may be valid. For example, in the case of promoting an individual contributor to a management position, the metric for "manager" may be different. **Ranking instead of scoring**: although ranking with hard cutoffs does not satisfy individual fairness, it is frequently used in practice. Can the model we have outlined with respect to scoring be translated to ranking, e.g., incorporating the results of [7]?

6 Related Work

There is a wide variety of work concerning fairness in machine learning [9, 14, 24, 17, 3, 18, 25, 22, 10, 16, 15, 11, 12, 20, 19, 4, 23, 5]. Individual fairness was introduced by Dwork et al. [5]. Dwork and Ilvento studied composition of combination of individually fair and group fair classifiers [6]. Two other recent lines of work have also considered composition problems and fair systems. First, several works have studied the problem of feedback loops, in which decisions that previous time steps, such as where to send law enforcement officers, influence outcomes at later time steps potentially unrelated to the original decision [12, 8, 21]. Bower et al. study fairness in a pipeline of decisions under a group-based notion of fairness [1]. They primarily consider the combination of multiple non-adaptive sequential decisions, evaluating fairness at the end of the pipeline. Second, several works have considered competitive scenarios, such as advertising, in which many (potentially fair or unfair) classifiers compete for individuals [2, 13]. Although not explicitly addressing composition, recent work considering fairness in rankings, e.g., [7], also address fairness in a setting in which outcomes, in this case rankings, naturally depend on the outcomes of others.

References

1. Amanda Bower, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alexander Vargas, and Suresh Venkatasubramanian. Fair pipelines. *CoRR*, abs/1707.00391, 2017. `arXiv:1707.00391`.
2. L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. Toward controlling discrimination in online ad auctions. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 4456–4465, 2019.
3. Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.
4. Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.
5. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226, 2012.
6. Cynthia Dwork and Christina Ilvento. Fairness under composition. In *10th Innovations in Theoretical Computer Science Conference, ITCS 2019, January 10-12, 2019, San Diego, California, USA*, pages 33:1–33:20, 2019.
7. Cynthia Dwork, Michael Kim, Omer Reingold, Guy Rothblum, and Gal Yona. Learning from outcomes: Evidence-consistent rankings. In *60th Annual IEEE Symposium on Foundations of Computer Science November 9-12, 2019, Baltimore, Maryland*, pages 106–125, 2019.
8. Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, pages 160–171, 2018.
9. Stephen Gillen, Christopher Jung, Michael J. Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 2605–2614, 2018.
10. Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

11 Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1944–1953. PMLR, 2018.

12 Lily Hu and Yiling Chen. Fairness at equilibrium in the labor market. *CoRR*, abs/1707.01590, 2017. `arXiv:1707.01590`.

13 Christina Ilvento, Meena Jagadeesan, and Shuchi Chawla. Multi-category fairness in sponsored search auctions. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 348–358, 2020.

14 Christopher Jung, Michael J. Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. Eliciting and enforcing subjective individual fairness. *CoRR*, abs/1905.10660, 2019. `arXiv:1905.10660`.

15 Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 2569–2577, 2018.

16 Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 656–666, 2017.

17 Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 4847–4857, 2018.

18 Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, pages 43:1–43:23, 2017.

19 Anja Lambrecht and Catherine Tucker. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, 65(7):2966–2981, 2019.

20 Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 6196–6200, 2019.

21 Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

22 David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 3381–3390, 2018.

23 Ya'acov Ritov, Yuekai Sun, and Ruofei Zhao. On conditional parity as a notion of non-discrimination in machine learning. *arXiv preprint*, 2017. `arXiv:1706.08519`.

24 Gal Yona and Guy N. Rothblum. Probably approximately metric-fair learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 5666–5674, 2018.

25 Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 325–333, 2013.

A Extended Motivating Examples

In each example, we consider a universe U comprised of individuals belonging to two groups, a majority group S and a minority group T, such that the majority group is k times as large as the minority group (i.e., $k|T| = |S|$). For the particular employment task in question, there is a known metric $\mathcal{D}$ which specifies who is similar to whom for the purposes of this task. For simplicity, we assume that $\mathcal{D}$ is one-dimensional, i.e., each individual u has a qualification $q_u \in [0, 1]$, and $\mathcal{D}(u, v) := |q_u - q_v|$. We assume that S and T have an equal distribution of talents: more specifically, for every qualification level q, there are exactly k times as many individuals with qualification q in S as there are in T. We assume that there is a nontrivial range of qualifications in $[0, 1]$, and we will generally assume that the company prefers to hire the most highly qualified candidates, but in order to fill the number of positions open cannot hire only maximally qualified candidates. We use Q_H to refer to the subset of individuals who are highly qualified.

Our examples are based on a set of facially neutral company compensation policies. We now give precise descriptions of these policies in the form of a scoring function, and indicate where the scoring policies must be adjusted to give intra-cohort individual fairness. (As we will see later, even adjusting the policies to be intra-cohort individually fair won't be enough to prevent bad behavior under composition.)

1. **Fixed Bonus Pool**: A fixed pool of bonus money B is assigned to each team and is split between the members of each team, with the highest achieving members receiving larger portions of the pool. More formally, given a cohort of individuals $C = \{x_1, \ldots, x_c\}$ of size c with qualifications $\{q_{x_1}, \ldots, q_{x_c}\}$, the scoring function f_B assigns a bonus share b_i to each individual x_i such that $\sum_{u \in C} b_u = 1$, optimized to ensure that individuals with higher qualification receive larger bonuses.
   In particular, f_B can either be a simple proportional mechanism, e.g., $f_B(u) \propto q_u$, or it can be optimized for specific goals, e.g., maximizing the difference in compensation between the most and least qualified individuals, creating an even spread of compensations, etc. For example, the company could choose f_B using the following optimization to choose the largest "weighted spread" to maximize the objective of increasing the difference in compensation based on difference in qualification: $argmax_{\{b_u \in [0,1]\}}\{\sum_{u,v \in C}(b_u - b_v)(q_u - q_v)\}$ subject to $|b_u - b_v| \leq |q_u - q_v|$ for all $u, v \in C$ and $\sum_{u \in C} b_u = 1$.
   This optimization will tend to choose bonus shares that maximize the differences in bonuses between individuals with significantly different qualifications within the cohort. Notice that the scoring function has no way of knowing what other cohorts may or may not appear and with what probabilities, and so it only optimizes within the particular cohort C.
2. **Stack Rank**: The bottom 10% of each team may be fired or put on "performance plans". Formally, $f(C, u) := \begin{cases} 1 \text{ if } \frac{|\{v | q_u > q_v\}|}{|C|} \leq 0.1, \\ 0 \text{ otherwise} \end{cases}$
   However, this strict cut off violates intra-cohort individual fairness, as two nearly equally qualified individuals might find themselves on opposite sides of the cutoff. Alternatively, we can construct a scoring function which closely approximates the desired policy but still satisfies intra-cohort individual fairness, by optimizing subject to the intra-cohort fairness constraints. For example, taking $\mathbb{O}_u$ to be the indicator that u is in the bottom 10% of the cohort, one could use the following optimization to maximize the probability that

only the bottom 10% are placed on performance plans: $argmax_f \sum_{u \in C} f(C,u)\mathbb{O}_u + (1 - f(C,u))(1-\mathbb{O}_u)$subject to $|f(C,u) - f(C,v)| \leq |q_u - q_v|$ for all $u, v \in C$. Alternatively, if exactly 10% of the cohort should be put on performance plans, Permute-Then-Classify can be applied or an additional constraint on the expected number of employees placed on performance plans could be added to the optimization above in order to satisfy *intra-cohort* individual fairness.

3. **Equal Treatment**: Each team's bonus is determined by average performance of the team (assumed to be proportional to average quality) and awarded equally to each member. Formally, the scoring function f first chooses the total bonus amount $B_C \propto B \sum_{u \in C} q_u$, and then assigns $b_u = \frac{B_C}{|C|}$ for all $u \in C$. Intra-cohort individual fairness for f is trivial, as every individual is treated equally.
4. **Promotion**: Choose the single most qualified person on the team to promote, based on performance. As in the case of stack ranking, strictly implementing this policy will violate intra-cohort individual fairness, as nearly equal individuals may be treated very differently. As above we can satisfy *intra-cohort* individual fairness by posing the relevant optimization question, and Permute-then-Classify (see Appendix B of the full version) can be used to select exactly one individual for promotion.

We now show that these compensation policies can cause significant unfairness for T when combined with simple hiring protocols. In each case, we state the set of cohorts the company intends to select from, and we assume that the company uses a method similar to the one described in Appendix C.3 of the full version to derive a fair set of weights to use to sample a single cohort in an individually fair way.[21]First, we consider the "packing" hiring protocol.

▶ **Example 31** (Packing). Suppose that in the past, the company had a particular problem retaining employees from the minority group T and in order to address this problem, the company ensures that individuals with high potential from T are always hired together into the same team for mutual support. On the other hand, talented members of S are spread out between the other teams, to make sure that there is at least one highly talented individual on each team. Formally, the company specifies the set of cohorts $\mathcal{C}_{packing} = \{C \in \mathcal{C} \mid (|C \cap T \cap Q_H| > 1 \wedge |C \cap S \cap Q_H| = 0) \oplus (|C \cap T \cap Q_H| = 0 \wedge |C \cap S \cap Q_H| = 1)\}$, where Q_H is the set of highly qualified candidates, and samples a single cohort from the set such that individual fairness is satisfied.

"Packing" results in lower compensation for T for Fixed Bonus Pool, Stack Rank, and Promotion compensation policies

"Packing" causes talented members of T to be on teams of higher average quality than those with talented members of S. As a result, members of T will receive lower bonuses and promoted less often than members of S. Thus, this seemingly beneficial practice can backfire when composed with certain compensation policies.

One may imagine that utilizing a "splitting" strategy, where qualified members of T are separated from other qualified members to increase their chance of "standing out" on teams, would solve this issue.

[21] We omit the details of the method and the particulars of the conditions on the set of cohorts specified as they are easy to fulfill in these settings.

▶ **Example 32** (Splitting). The company chooses teams where highly qualified members of T are always the only highly qualified member of their team, giving them the opportunity to stand out and be recognized for their talent. More formally, the company chooses from the set of cohorts $\mathcal{C}_{splitting} = \{C \in \mathcal{C} \mid (|C \cap T \cap Q_H| = 1 \wedge |C \cap S \cap Q_H| = 0) \oplus (|C \cap T \cap Q_H| = 0 \wedge |C \cap S \cap Q_H| \geq 1)\}$. In each cohort containing a highly qualified member of T, there are no other highly qualified individuals (from either T or S).

Though this policy no longer leads to lower compensation for T for Stack Rank, Fixed Bonus Pool, and Promotion, "Splitting" results in lower compensation for T for Equal Treatment, because the practice causes talented members of T to be on teams of lower average quality than talented members of S. As a result, with Equal Treatment, qualified S will receive greater compensation than qualified T. Splitting can also occur when members of T are primarily hired via outreach. For example, suppose that a company has been trying to form a team to work on a difficult or low prestige task (e.g., Fortran code maintenance). All of the talented candidates in S pass on the job offer because they are confident they can do better, so HR reaches out more aggressively to candidates in T. These candidates may be more willing to take the job because they are less confident about their other options. Thus, even without an explicit policy in place to choose minority candidates to be the singular most qualified member on a less qualified team, these situations can still arise from the interactions between the hiring procedure and the job market.

▶ Remark 33. The motivation for both of these policies could be malicious, and determining whether the stated goals or justifications were legitimate aims of the policy would be difficult.

One may imagine that these issues could be addressed by ensuring that qualified members of T and qualified members of S appearing on teams with similar average quality. However, a malicious company can still cause members of T to receive lower compensation.

▶ **Example 34** (Adversarial ranking). Suppose that the company did not want any member of the T to be chosen for promotion or wished to depress their compensation relative to the members of S. The company decides to choose teams such that, for each team, there is a correspondence between the members of T and S included in the team, such that the members of S are almost always more talented than their counterparts in T. (Given the equal distribution of talents of T and S, there may be an excess member of T that is allowed to be the most qualified, but this is a singular case.) More formally, the company chooses from $\mathcal{C}_{adv.ranking} = \{C \in \mathcal{C} \mid \exists G : C \cap T \to C \cap S \text{ s.t. } \forall u \in C \cap T,\, q_u < q_{G(u)}\}$.

"Adversarial Ranking" is particularly catastrophic for T for Promotion or Stack Ranking if the hard cutoff (not intra-cohort individually fair) versions are used. Although ensuring intra-cohort individual fairness helps, members of T will always be seeing depressed levels of promotion, higher levels of firing, and lower levels of compensation except in the case of Equal Treatment. Thus "Adversarial Ranking" keenly illustrates that average team quality is not sufficient to ensure that individuals are truly being treated fairly in cohort-based pipelines. We stress that Adversarial Ranking can also be efficiently achieved using the procedure described in Appendix C.3 of the full version.

Sample Cohorts

To illustrate these issues, we include Figures 1a and 1b to compare the example scoring functions for a pair of cohorts, demonstrating the issues outlined above.

	Qualification	Fixed Pool Bonus	Equal Bonus
Cohort 1			
Alice	0.8	35	60
Bob	0.7	25	60
Charlie	0.5	5	60
Dan	0.2	0	60
Eve	0.8	35	60
Cohort 2			
Frank	0.8	57	40
George	0.6	36	40
Harriet	0.1	0	40
Ivan	0.2	0	40
Julia	0.3	7	40

(a) Bonus score function comparisons for two cohorts, each containing five individuals of varying qualifications. Cohort 1 has an average qualification of 0.6, and Cohort 2 has an average qualification of 0.4. In the fixed pool bonus, a total pool of 100 is split between the members of the cohorts. The same optimization is used for both cohorts, that is according the maximum possible bonus to the most qualified individual(s). Notice that in Cohort 1, Alice and Eve have to share the top bonus (35 each), but in Cohort 2, Frank doesn't have to split the top bonus (57). Notice also that George and Julia receive higher bonuses than Bob and Charlie, even though they are (much) less qualified. On the other hand, in the equal bonus setting Frank receives a lower bonus than both Alice and Eve, even though he's equally qualified.

	Qualification	Promotion	Stack Rank (IF)	Stack Rank (exact, not IF)
Cohort 1				
Alice	0.8	35%	0	0
Bob	0.7	25%	10%	0
Charlie	0.5	5%	30%	0
Dan	0.2	0	60%	1
Eve	0.8	35%	0	0
Cohort 2				
Frank	0.8	57%	0	0
George	0.6	36%	0	0
Harriet	0.1	0	43%	1
Ivan	0.2	0	33%	0
Julia	0.3	7%	24%	0

(b) Promotion score function comparison of the cohorts from Figure 1a. The promotion policy attempts to maximize the probability of promotion for the most qualified individuals, subject to the individual fairness constraints and that the expected number of promotions is 1. In this case, essentially the same observations apply as in the fixed pool bonus setting. In the case of Stack rank, both cohorts are optimized to maximize the probability of placing the least qualified person on a performance plan. Notice that Dan is much more likely to be placed on a performance plan than the equally qualified Ivan, due to the larger number of less qualified individuals in Cohort 2. Although it might seem that the exact stack rank policy, rather than the individually fair version, would be less likely to have this problem, in fact in this case Dan is still treated differently than Ivan.