Interventional Fairness : Causal Database Repair for Algorithmic Fairness

Babak Salimi[1] Luke Rodriguez [2] Bill How [2] Dan Suciu[1]

[1] Computer Science and Engineering
University of Washington
bsalimi, suciu@cs.washington.edu

[2] Information School
University of Washington
rodriglr, billhowe@uw.edu

ABSTRACT

Fairness is increasingly recognized as a critical component of machine learning systems. However, it is the underlying data on which these systems are trained that often reflect discrimination, suggesting a database repair problem. Existing treatments of fairness rely on statistical correlations that can be fooled by statistical anomalies, such as Simpson's paradox. Proposals for causality-based definitions of fairness can correctly model some of these situations, but they require specification of the underlying causal models. In this paper, we formalize the situation as a database repair problem, proving sufficient conditions for fair classifiers in terms of admissible variables as opposed to a complete causal model. We show that these conditions correctly capture subtle fairness violations. We then use these conditions as the basis for database repair algorithms that provide provable fairness guarantees about classifiers trained on their training labels. We evaluate our algorithms on real data, demonstrating improvement over the state of the art on multiple fairness metrics proposed in the literature while retaining high utility.

ACM Reference Format:
Babak Salimi[1] Luke Rodriguez [2] Bill How [2] Dan Suciu[1]. 2019. Interventional Fairness : Causal Database Repair for Algorithmic Fairness. In *2019 International Conference on Management of Data (SIGMOD '19), June 30-July 5, 2019, Amsterdam, Netherlands.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3299869.3319901

SIGMOD '19, June 30-July 5, 2019, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5643-5/19/06…$15.00
https://doi.org/10.1145/3299869.3319901

1 INTRODUCTION

In 2014, a team of machine learning experts from Amazon Inc. began work on an automated system to review job applicants' resumes. According to a recent Reuters article [12], the experimental system gave job candidates scores ranging from one to five and was trained on 10 years of recruiting data from Amazon. However, by 2015 the team realized that the system showed a significant gender bias towards male over female candidates because of historical discrimination in the training data. Amazon edited the system to make it gender agnostic, but there was no guarantee that discrimination did not occur through other means, and the project was totally abandoned in 2017.

Fairness is increasingly recognized as a critical component of machine learning (ML) systems, which make daily decisions that affect people's lives [11]. The data on which these systems are trained reflect institutionalized discrimination that can be reinforced and legitimized through automation. A naive (and ineffective) approach sometimes used in practice is to simply omit the protected attribute (say, race or gender) when training the classifier. However, since the protected attribute is frequently represented implicitly by some combination of proxy variables, the classifier still learns the discrimination reflected in training data. For example, zip code tends to predict race due to a history of segregation [21, 46]; answers to personality tests identify people with disabilities [4, 53]; and keywords can reveal gender on a resume [12]. As a result, a classifier trained without regard to the protected attribute not only fails to remove discrimination, but it can complicate the detection and mitigation of discrimination downstream via in-processing or post-processing techniques [9, 10, 18, 24, 25, 34, 43, 51], which we next describe.

The two main approaches to reduce or eliminate sources of discrimination are summarized in Fig. 1. The most popular is the in-processing, where the ML algorithm itself is modified; this approach must be reimplemented for every

ML application. The alternative is to process either the training data (pre-processing) or the output of the classifier itself (post-processing). We advocate for the pre-processing strategy, which is agnostic to the choice of ML algorithm and instead interprets the problem as a database repair task.

One needs a quantitative measure of discrimination in order to remove it. A large number of fairness definitions have been proposed (see Verma and Rubin for a recent discussion [52]), which we broadly categorize in Fig. 1. The best-known measures are based on *associative* relationships between the protected attribute and the outcome. For example, Equalized Odds requires that both protected and privileged groups have the same true positive (TP) and false positive (FP) rates. However, it has been shown that associative definitions of fairness can be mutually exclusive [9] and fail to distinguish between discriminatory, non-discriminatory and spurious correlations between a protected attribute and the outcome of an algorithm [13, 24, 34].

Example 1.1. In a well-studied case, UC Berkeley was sued in 1973 for discrimination against females in graduate school admissions when it was found that 34.6% of females were admitted in 1973 as opposed to 44.3% of males. It turned out that females tended to apply to departments with lower overall acceptance rates [44]. When broken down by department, a slight bias toward female applicant was observed, a result that did not constitute evidence for gender-based discrimination.

Such situations have recently motivated a search for a more principled measure of fairness and discrimination based on *causality* [18, 24, 25, 34, 43]. These approaches measure the discriminatory causal influence of the protected attribute on the outcome of an algorithm. However, they typically assume access to background information regarding the underlying causal model, which is unrealistic in practice. For example, Kilbertus et al. assume the underlying casual model is provided as a structural equation model [24]. Moreover, no existing proposals describe comprehensive systems for pre-processing data to mitigate causal discrimination.

This paper describes a new approach to removing discrimination by *repairing the training data* in order to remove the effect of any inappropriate and discriminatory causal relationship between the protected attribute and classifier predictions, without assuming adherence to an underlying causal models.

Our system, CAPUCHIN, accepts a dataset consisting of a protected attribute (e.g., gender, race, etc.), an outcome attribute (e.g., college admissions, loan application, or hiring decisions), and a set of *admissible variables* through which it is permissible for the protected attribute to influence the outcome. For example, the applicant's choice of department in Example 1.1 is considered admissible despite being correlated with gender. The system repairs the input data by inserting or removing tuples, changing the empirical probability distribution to remove the influence of the protected attribute on the outcome through any causal pathway that includes inadmissible attributes. That is, the repaired training data can be seen as a *sample from a hypothetical fair world*. We make this notion more precise in Section 3.1.

	Associational	Causal
In-processing (Modify the ML Algorithm)	[6, 23, 24, 34, 58]	[24, 34, 43]
Pre/post-processing (Modify the input/output Data)	[7, 16, 19, 55]	CAPUCHIN (this paper)

Figure 1: Fairness metrics and enforcement methods.

Unlike previous measures of fairness based on causality [24, 34, 43], which require the presence of the underlying causal model, our definition is based solely on the notion of *intervention* [36] and can be guaranteed even in the absence of causal models. The user need only distinguish admissible and inadmissible attributes; we prove that this information is sufficient to support the causal inferences needed to mitigate discrimination.

We use this *interventional* approach to derive in Sec. 3.1 a new fairness definition, called *justifiable fairness*. Justifiable fairness subsumes and improves on several previous definitions and can correctly distinguish fairness violations and non-violations that would otherwise be hidden by statistical coincidences, such as Simpson's paradox. We prove next, in Sec. 3.2, that, if the training data satisfies a simple saturated conditional independence, then any reasonable algorithm trained on it will be fair.

Our core technical contribution, then, consists of a new approach to repair training data in order to enforce the saturated conditional independence that guarantees fairness. The database repair problem has been extensively studied in the literature [3], but in terms of database constraints, not conditional independence. In Sec. 4 we first define the problem formally and then present a new technique to reduce it to a multivalued functional dependency MVD [1]. Finally, we introduce new techniques to repair a dataset for an MVD by reduction to the MaxSAT and Matrix Factorization problems.

We evaluate our approach in Sec 6 on two real datasets commonly studied in the fairness literature, the adult dataset [28] and the COMPAS recidivism dataset [49]. We find that our algorithms not only capture fairness situations other approaches cannot, but that they outperform the existing state-of-the-art pre-processing approaches *even on other fairness metrics for which they were not necessarily designed*. Surprisingly, our results show that our repair algorithms can mitigate discrimination as well as prohibitively aggressive approaches, such as dropping all inadmissible variables from

the training set, while maintaining high accuracy. For example, our most flexible algorithm, which involves a reduction to MaxSAT, can remove almost 50% of the discrimination while decreasing accuracy by only 1% on adult data.

We make the following contributions: We develop a new framework for causal fairness that does not require a complete causal model; We prove sufficient conditions for a fair classifier based on this framework; We reduce fairness to a database repair problem by linking causal inference to multivalued dependencies (MVDs); We develop a set of algorithms for the repair problem for MVDs; We evaluate our algorithms on real data and show that they meet our goals and outperform competitive methods on multiple metrics.

Section 2 presents background on fairness and causality, while Section 3 describes sufficient conditions for a fair classifier and derives the database repair problem. In Section 4, we present algorithms for solving the database repair problem and show, in Section 6, experimental evidence that our algorithms outperform the state-of-the-art on multiple fairness metrics while preserving high utility.

2 PRELIMINARIES

We review in this section the basic background on database repair, algorithmic fairness and models of causality, the building blocks of our paper. See the full version of this paper ([45]) for more details.

The notation used is summarized in Table 1. We denote variables (i.e., dataset attributes) by uppercase letters, X, Y, Z, V; their values with lower case letters, x, y, z, v; and denote sets of variables or values using boldface ($\mathbf{X}$ or $\mathbf{x}$). The domain of a variable X is $Dom(X)$, and the domain of a set of variables is $Dom(\mathbf{X}) = \prod_{Y \in \mathbf{X}} Dom(Y)$. In this paper, all domains are discrete and finite; continuous domains are assumed to be binned, as is typical. A *database instance* D is a relation whose attributes we denote as $\mathbf{V}$. We assume set semantics (*i.e.*, no duplicates) unless otherwise stated, and we denote the cardinality of D as $n = |D|$. Given a partition $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z} = \mathbf{V}$, we say that D satisfies the *multivalued dependency* (MVD) $\mathbf{Z} \twoheadrightarrow \mathbf{X}$ if $D = \Pi_{\mathbf{XZ}}(D) \bowtie \Pi_{\mathbf{ZY}}(D)$.

Typically, training data for ML is a bag B. We convert it into a set D (by eliminating duplicates) and a probability distribution Pr, which accounts for multiplicies;[1] We call D the support of Pr. We say that Pr is *uniform* if all tuples have the same probability. We say $\mathbf{X}$ and $\mathbf{Y}$ are *conditionally independent (CI)* given $\mathbf{Z}$, written ($\mathbf{X} \perp\!\!\!\perp_{\Pr} \mathbf{Y} | \mathbf{Z}$), or just ($\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$) if $\Pr(\mathbf{x}|\mathbf{y},\mathbf{z}) = \Pr(\mathbf{x}|\mathbf{z})$ whenever $\Pr(\mathbf{y},\mathbf{z}) > 0$. Conditional independences satisfy the Graphoid axioms [39], which are reviewed in Appendix 8.1 and are used in proofs. When $\mathbf{V} = \mathbf{XYZ}$, then the CI is said to be *saturated*. A uniform Pr satisfies a saturated CI iff its support D satisfies the MVD $\mathbf{Z} \twoheadrightarrow \mathbf{X}$. Training data usually does not have a uniform Pr, and in that case the equivalence between the CI and MVD fails [54]; we address this issue in Sec. 4.

[1] $\Pr(\mathbf{v}) \stackrel{\text{def}}{=} \frac{1}{|B|} \sum_{t \in B} 1_{t=\mathbf{v}}$.

Symbol	Meaning
X, Y, Z	attributes (variables)
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	sets of attributes
$Dom(X)$, $Dom(\mathbf{X})$	their domains
$x \in Dom(X)$, $\mathbf{x} \in Dom(\mathbf{X})$	a single value, a tuple of values
D	the database instance
$\mathbf{V}$	the attributes of the database D
G	causal DAG
$X \rightarrow Y$	an edge in G
$\mathbf{Pa}(X)$	the parents of X in G
$\mathbf{P}$	a path in G
$\mathbf{X} \perp\!\!\!\perp_{\Pr} \mathbf{Y} \mid \mathbf{Z}$ or $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$	conditional independence
$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid_d \mathbf{Z})$	d-Separation in G.
$\mathbf{MB}(X)$	The Markov boundary of X
I	Inadmissible attributes
A	Admissible attributes

Table 1: Notation used in the paper.

The *database repair problem* is the following: we are given a set of constraints Γ and a database instance D, and we need to perform a minimal set of updates on D such that the new database D' satisfies Γ [3]. The problem has been studied extensively in database theory for various classes of constraints Γ. It is NP-hard even when D consists of a single relation (as it does in our paper) and Γ consists of functional dependencies [29]. In our setting, Γ consists of conditional independence statements, and it remains NP-hard, as we show in Sec. 4.

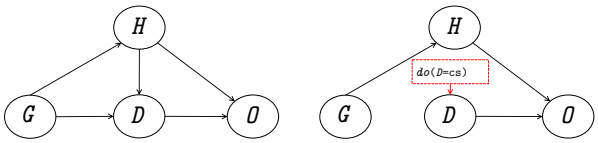
2.1 Background on Algorithmic Fairness

Algorithmic fairness considers a *protected attribute* S, a *response variable* Y, and a prediction algorithm $A : Dom(\mathbf{X}) \rightarrow Dom(O)$, where $\mathbf{X} \subseteq \mathbf{V}$, and the prediction of A is denoted O (some references denote it $\tilde{Y}$) and called *outcome*. For simplicity, we assume S classifies the population into protected $S = 1$ and privileged $S = 0$, for example, female and male. Fairness definitions can be classified as associational or causal.

Associational fairness is based on statistical measures on the variables of interest; a summary is shown in Fig. 2. *Demographic Parity* (DP) [6, 13, 22, 47, 59], requires an algorithm to classify both the protected and the privileged group with the same probability. As we saw in Example 1.1, the lack of statistical parity cannot be considered as evidence for gender-based discrimination; this has motivated the introduction of *Conditional Statistical Parity* (CSP) [10], which controls for a set of admissible factors $\mathbf{A}$. Another popular measure used for predictive classification algorithms is *Equalized Odds* (EO), which requires that both protected and privileged groups to have the same false positive (FP) rate, and the same false negative (FN) rate. Finally, *Predictive Parity* (PP) requires that both protected and unprotected groups have the same predicted positive value (PPV) It has been shown that these

Fairness Metric	Description
Demographic Parity (DP) [5, 13, 47]	$S \perp\!\!\!\perp O$
Conditional Statistical parity [10]	$S \perp\!\!\!\perp O \mid \mathbf{A}$
Equalized Odds (EO) [19, 57]	$S \perp\!\!\!\perp O \mid Y$
Predictive Parity (PP)[9, 9, 19, 47]	$S \perp\!\!\!\perp Y \mid O$

Figure 2: Common associational definitions of fairness.



a) **b)**

Figure 3: Causal DAGs related to Ex.2.2.

measures can be mutually exclusive [9]. Recently, associational fairness has bee studied in the context of statistical relational learning [14, 15].

Causal fairness [18, 24, 25, 34, 43] was motivated by the need to address difficulties generated by associational fairness and assumes an underlying causal model. We first discuss causal DAGs before reviewing causal fairness.

2.2 Background on Causal DAGs

Causal DAG. A *causal DAG* G over set of variables $\mathbf{V}$ is a directed acyclic graph that models the functional interaction between variables in $\mathbf{V}$. Each node X represents a variable in $\mathbf{V}$ that is functionally determined by: (a) its parents $\mathbf{Pa}(X)$ in the DAG, and (b) some set of *exogenous* factors that need not appear in the DAG, as long as they are mutually independent. This functional interpretation leads to the same decomposition of the joint probability distribution of $\mathbf{V}$ that characterizes Bayesian networks [36]:

$$\Pr(\mathbf{V}) = \prod_{X \in \mathbf{V}} \Pr(X|\mathbf{Pa}(X)) \tag{1}$$

***d*-Separation and Faithfulness.** A common inference question in a causal DAG is how to determine whether a CI $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})$ holds. A sufficient criterion is given by the notion of d-separation, a syntactic condition $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|_d \mathbf{Z})$ that can be checked directly on the graph. Pr and G are called *Markov compatible* if $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|_d \mathbf{Z})$ implies $(\mathbf{X} \perp\!\!\!\perp_{\Pr} \mathbf{Y}|\mathbf{Z})$; if the converse implication holds, then we say that Pr is *faithful* to G. The following is known:

Proposition 2.1. *If G is a causal DAG and* Pr *is given by Eq.(1), then they are Markov compatible.*

Counterfactuals and do Operator. A *counterfactual* is an intervention where we actively modify the state of a set of variables $\mathbf{X}$ in the real world to some value $\mathbf{X} = \mathbf{x}$ and observe the effect on some output Y. Pearl [36] described the *do* operator that allows this effect to be computed on a causal DAG, denoted $\Pr(Y|do(X = x))$. To compute this value, we assume that X is determined by a constant function $X = x$ instead of a function provided by the causal DAG. This assumption corresponds to a modified graph with all edges into $\mathbf{X}$ removed, and values of these variables are set to $\mathbf{x}$. The Bayesian rule Eq.(1) for the modified graph defines $\Pr(Y|do(\mathbf{X} = \mathbf{x}))$; the exact expression is in [36, Theorem 3.2.2]. We give an alternative and, to our best knowledge, new formula expressed by introducing some compensating factors; the proof is in Appendix 8.2:

Theorem 2.1. *Given a causal DAG G and a set of variables $\mathbf{X} \subseteq \mathbf{V}$, suppose $\mathbf{X} = \{X_0, X_1 \ldots X_m\}$ are ordered such that X_i is a non-descendant of X_{i+1} in G. The effect of a set of interventions $do(\mathbf{X} = \mathbf{x})$ is given by the following* extended adjustment formula*:*

$$\Pr(y|do(\mathbf{X} = \mathbf{x})) = \sum_{\mathbf{z} \in Dom(\mathbf{Z})} \Pr(y|\mathbf{x}, \mathbf{z}) \left(\prod_{i=0}^{m} \Pr\left(\mathbf{pa}(X_i) \middle| \bigcup_{j=0}^{i-1} \mathbf{pa}(X_j), \bigcup_{j=0}^{i-1} x_j\right)\right) \tag{2}$$

where $\mathbf{Z} = \bigcup_{X \in \mathbf{X}} \mathbf{Pa}(X)$ and $j \geq 0$.

In particular, if $\mathbf{X}$ has no parents, then intervention coincides with conditioning, $\Pr(y|do(\mathbf{X} = \mathbf{x})) = \Pr(y|\mathbf{X} = \mathbf{x})$.

Example 2.2. Continuing Example 1.1, Fig. 3(a) shows a small fragment of the causal DAG of the admission process in a college. Admissions decisions are made independently by each department and are based on a rich collection of information about the candidates, such as test scores, grades, etc. These characteristics affect not only the admission decisions, but also which department the candidate chooses to apply to. We show only a tiny fragment of the causal graph, where O = admission outcome, D = department, G = candidate's gender, and H = hobbies, which can be influenced by gender. [2] The admissions office anonymizes gender, but it does consider extracurricular activities such as hobbies, so we include an edge $H \longrightarrow O$. Since different genders apply to departments at different rates, there is an edge $G \longrightarrow D$. Some departments may tend to attract applicants with certain hobbies (e.g., the math department may attract applicants who play chess), so we also include an edge $H \longrightarrow D$. The joint distribution is given by

$$\Pr(g, h, d, o) = \Pr(g)\Pr(h|g)\Pr(d|g, h)\Pr(o|h, d) \tag{3}$$

Consider the counterfactual: *update the applicant's department to cs*. We compare the marginal probability of O, the conditional probability, and the intervention:

$$\Pr(o|D = \text{cs}) = \sum_{g,h} \Pr(g)\Pr(h|g)\Pr(D = \text{cs}|g, h)\Pr(o|D = \text{cs}, h)$$

$$\Pr(o|do(D = \text{cs})) = \sum_{g,h} \Pr(g)\Pr(h|g)\Pr(o|D = \text{cs}, h) \tag{4}$$

The expression for intervention (4), based on [36, Theorem 3.2.2] is obtained from the conditional probability by

[2] In the Amazon hiring example [12], hobbies correlated with gender, *e.g., Captain of the women's chess team.*

removing the term $\Pr(D = \text{cs}|g, h)$, or equivalently deleting the edge $G \rightarrow D$ from the graph in Fig. 3(b). Alternatively, we can express the intervention using Eq.(2) (notice that $\mathbf{Pa}(D) = \{G, H\}$):

$$\Pr(o|do(D = \text{cs})) = \sum_{g,h} \Pr(o|g, h, D = \text{cs})\Pr(h|g)\Pr(g) \quad (5)$$

2.3 Causal Fairness

Counterfactual Fairness. Kusner et al. [25, 26] (see also the discussion in [30]) defined a classifier as *counterfactually fair* if the protected attribute of an individual is not a cause of the outcome of the classifier for that individual, i.e., had the protected attributes of the individual been different, and other things being equal, the outcome of the predictor would have remained the same. However, the definition of counterfactual fairness in [25] captures individual-level fairness only under certain assumptions (see Appendix 8.1). Indeed, it is known that individual-level counterfactuals can not be estimated from data [40–42].

Proxy Fairness. To avoid individual-level counterfactuals, a common is to study population-level counterfactuals or interventional distributions that capture the effect of interventions at population level rather than individual level [38, 40, 41]. Kilbertus et. al. [24] defined proxy fairness as follows:

$$P(\tilde{Y} = 1|do(\mathbf{P} = \mathbf{p})) = P(\tilde{Y} = 1|do(\mathbf{P} = \mathbf{p}')) \quad (6)$$

for any $\mathbf{p}, \mathbf{p}' \in Dom(\mathbf{P})$, where $\mathbf{P}$ consists of proxies to a sensitive variable S (and might include S). Intuitively, a classifier satisfies proxy fairness in Eq 6, if the distribution of $\tilde{Y}$ under two interventional regimes in which $\mathbf{P}$ set to $\mathbf{p}$ and $\mathbf{p}'$ is the same. Thus, proxy fairness is not an individual-level notion. Next example shows proxy fairness fails to capture group-level discrimination in general.

Example 2.3. To illustrate the difference between counterfactual and proxy fairness, consider the college admission example. Both departments make decisions based on students' gender and qualifications, $O = f(G, D, Q)$, for a binary G and Q. The causal DAG is $G \rightarrow O, D \rightarrow O, Q \rightarrow O$. Let $D = U_D$ and $Q = U_Q$, where U_D and U_Q are exogenous factors that are independent and that are uniformly distributed, e.g., $P(U_Q = 1) = P(U_Q = 0) = \frac{1}{2}$. Further suppose $f(G, \text{'A'}, Q) = G \wedge Q$ and $f(G, \text{'B'}, Q) = (1 - G) \wedge Q$, i.e., dep. A admits only qualified males and dep. B admits only qualified females. This admission process is proxy-fair[3], because $P(O = 1|do(G = 1)) = P(O = 1|do(G = 0)) = \frac{1}{2}$. On the other hand, it is clearly individually-unfair, in fact it is group-level unfair (for all applicants to the same department). To capture individual fairness, counterfactual fairness [25, 26] is a non-standard definition that does both conditioning *and* intervention on the sensitive attribute. Conditioning "extracts information from the individual to learn the background variables" [30, pp.11, footnote 1].

Path-specific fairness. These definitions are based on graph properties of the causal graph, *e.g.*, prohibiting specific paths from the sensitive attribute to the outcome [30, 34]; however, identifying path-specific causality from data requires very strong assumptions and is often impractical [2].

3 DEFINING AND ENFORCING ALGORITHMIC FAIRNESS

In this section we introduce a new definition of fairness, which, unlike proxy fairness [24], captures correctly group-level fairness, and, unlike counterfactual fairness [25, 26] is based on the standard notion of intervention and, hence, it is testable from the data. In the next section we will describe how to repair an unfair training dataset to enforce fairness.

3.1 Interventional Fairness

This section assumes that the causal graph is given. The algorithm computes an output variable O from input variables $\mathbf{X}$ (Sec. 2.1). We begin with a definition describing when an outcome O is causally independent of the protected attribute S for any possible configuration of a given set of variables $\mathbf{K}$.

Definition 3.1 ($\mathbf{K}$-fair). Fix a set of attributes $\mathbf{K} \subseteq \mathbf{V} - \{S, O\}$. We say that an algorithm $\mathcal{A} : Dom(\mathbf{X}) \rightarrow Dom(O)$ is $\mathbf{K}$-fair w.r.t. a protected attribute S if, for any context $\mathbf{K} = \mathbf{k}$ and every outcome $O = o$, the following holds:

$$\Pr(O = o|do(S = 0), do(\mathbf{K} = \mathbf{k})) = \Pr(O = o|do(S = 1), do(\mathbf{K} = \mathbf{k})) \quad (7)$$

We call an algorithm *interventionally fair* if it is $\mathbf{K}$-fair for every set $\mathbf{K}$. Unlike proxy fairness, this notion captures correctly group-level fairness, because it ensures that S does not affect O in *any configuration* of the system obtained by fixing other variables at some arbitrary values. Unlike counterfactual fairness, it does not attempt to capture fairness at the individual level, and therefore it uses the standard definition of intervention (the do-operator). In fact, we argue that interventional fairness is the strongest notion of fairness that is testable from data, yet captures correctly group-level fairness. We illustrate with an example (see also Ex 3.6).

Example 3.2. In contrast to proxy fairness, interventional fairness correctly identifies the admission process in Ex. 2.3 as unfair at department-level. This is because the admission process fails to satisfy $\{D\}$-fairness since, $P(O = 1|do(G = 0), do(D = \text{'A'})) = 0$ but $P(O = 1|do(G = 1), do(D = \text{'A'})) = \frac{1}{2}$. Therefore, interventional fairness is a more fine-grained notion than proxy fairness. We note however that, interventional fairness does not guarantee individual faintness in general. To see this suppose the admission decisions in both departments are based on student's gender and an unobserved exogenous factor U_O that is uniformly distributed, i.e., $O = f(G, U_O)$, such that $f(G, 0) = G$ and $f(G, 1) = 1 - G$. Hence,

[3]Here D is not a proxy to G, because $D \perp\!\!\!\perp G$ by assumption.

the causal DAG is $G \to O$. Then the admission process is $\emptyset$-fair because, $P(O = 1|do(G = 1)) = P(O = 1|do(G = 0)) = \frac{1}{2}$. Therefore, it is interventionally fair (since $\mathbf{V} - \{O, G\} = \emptyset$). However, it is clearly unfair at individual level. If the variable U_0 were endogenous (i.e. known to the algorithm), then the admission process is no longer interventionally fair, because it is not $\{U_o\}$-fair: $P(O = 1|do(G = 1), do(U_o = 1)) = P(O = 1|G = 1, U_o = 1) = 0$, while $P(O = 1|do(G = 1), do(U_o = 1)) = P(O = 1|G = 0, U_o = 1) = 1$. Under the same setting counterfactual fairnesses [25, 26] fails to capture individual-level discrimination in this example (see Appendix 8.1).

In practice, interventional fairness is too restrictive, as we show below. To make it practical, we allow the user to classify variables into *admissible* and *inadmissible*. The former variables through which it is permissible for the protected attribute to influence the outcome. In Example 1.1, the user would label department as admissible since it is considered a fair use in admissions decisions, and would (implicitly) label all other variables as inadmissible, for example, hobby. Only users can identify this classification, and therefore admissible variables are part of the problem definition:

Definition 3.3 (Fairness application). A fairness application over a domain $\mathbf{V}$ is a tuple $(\mathcal{A}, S, \mathbf{A}, \mathbf{I})$, where $\mathcal{A}$ is an algorithm $Dom(\mathbf{X}) \to Dom(O)$; $\mathbf{X} \subseteq \mathbf{V}$ are its input variables; $S, O \in \mathbf{V}$ are the protected attribute and outcome, and $\mathbf{A} \cup \mathbf{I} = \mathbf{V} - \{S, O\}$ is a partition of the variables into admissible and inadmissible.

We can now introduce our definition of fairness:

Definition 3.4 (Justifiable fairness). A fairness application $(\mathcal{A}, S, \mathbf{A}, \mathbf{I})$ is *justifiability fair* if it is $\mathbf{K}$-fair w.r.t. all supersets $\mathbf{K} \supseteq \mathbf{A}$.

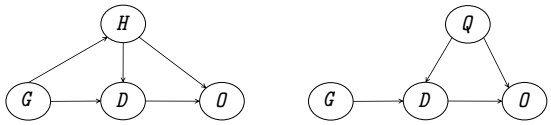
Notice that interventional fairness corresponds to the case where no variable is admissible, i.e., $\mathbf{A} = \emptyset$.

We give next a characterization of justifiable fairness in terms of the structure of the causal DAG:

Theorem 3.5. *If all directed paths from S to O go through an admissible attribute in $\mathbf{A}$, then the algorithm is justifiably fair. If the probability distribution is faithful to the causal DAG, then the converse also holds.*

To ensure interventional fairness, a sufficient condition is that there exists no path from S to O in the causal graph (because $\mathbf{A} = \emptyset$). (Hence, under faithfulness, interventional fairness implies fairness at individual-level, i.e., intervening on the sensitive attribute does not change the counterfactual outcome of individuals.) Since this is too strong in most scenarios, we adopt justifiable fairness instead. We illustrate with an example.

Example 3.6. Fig 4 shows how fair or unfair situations may be hidden by coincidences but exposed through causal analysis. In both examples, the protected attribute is gender G, and



a) College I **b) College II**

College I	Dept. A		Dept. B		Total	
	Admitted	Applied	Admitted	Applied	Admitted	Applied
Male	16	20	16	80	32	100
Female	16	80	16	20	32	100

College II	Dept. A		Dept. B		Total	
	Admitted	Applied	Admitted	Applied	Admitted	Applied
Male	10	10	40	90	50	100
Female	40	50	10	50	50	100

Figure 4: Admission process representation in two colleges where the associational notions of fairness fail (see Ex.3.6).

the admissible attribute is department D. Suppose both departments in College I are admitting only on the basis of their applicants' hobbies. Clearly, the admission process is discriminatory in this college because department A admits 80% of its male applicants and 20% of the female applicants, while department B admits 20% of male and 80% of female applicants. On the other hand, the admission rate for the entire college is the same 32% for both male and female applicants, falsely suggesting that the college is fair. Suppose H is a proxy to G such that $H = G$ (G and H are the same), then proxy fairness classifies this example as fair: indeed, since Gender has no parents in the causal graph, intervention is the same as conditioning, hence $\Pr(O = 1|do(G = \text{i})) = \Pr(O = 1|G = \text{i})$ for $i = 0, 1$. Of the previous methods, only conditional statistical parity correctly indicates discrimination. We illustrate how our definition correctly classifies this examples as unfair. Indeed, assuming the user labels the department D as admissible, $\{D\}$-fairness fails because, by Eq.(2), $\Pr(O = 1|do(G = 1), do(D = \text{'A'})) = \sum_h \Pr(O = 1|G = 1, D = \text{'A'}, h)\Pr(h|G = 1) = \Pr(O = 1|G = 1, D = \text{'A'}) = 0.8$, and, similarly $\Pr(O = 1|do(G = 0), do(D = \text{'A'})) = 0.2$. Therefore, the admission process is not justifiably fair.

Now, consider the second table for College II, where both departments A and B admit only on the basis of student qualifications Q. A superficial examination of the data suggests that the admission is unfair: department A admits 80% of all females, and 100% of all male applicants; department B admits 20% and 44.4% respectively. Upon deeper examination of the causal DAG, we can see that the admission process is justifiably fair because the only path from Gender to the Outcome goes through department, which is an admissible attribute. To understand how the data could have resulted from this causal graph, suppose 50% of each gender have

high qualifications and are admitted, while others are rejected, and that 50% of females apply to each department but more qualified females apply to department A than to B (80% v.s. 20%). Further, suppose fewer males apply to department A, but all of them are qualified. The algorithm satisfies demographic parity and proxy fairness but fails to satisfy conditional statistical parity since $\Pr(A = 1|G = 1, D = \text{A}) = 0.8$ but $\Pr(A = 1|G = 0, D = \text{A}) = 0.2)$. Thus, conditioning on D falsely indicates discrimination in College II. One can check that the algorithm is justifiably fair, and thus our definition also correctly classifies this example; for example, $\{D\}$-fairness follows from Eq.(2): $\Pr(O = 1|do(\text{G} = \text{i}), do(D = d)) = \sum_q \Pr(O = 1|G = i, d, q))\Pr(q|G = i) = \frac{1}{2}$. To summarize, unlike previous definitions of fairness, justifiable fairness correctly identifies College I as discriminatory and College II as fair.

3.2 Testing Fairness on the Training Data

In this section we introduce a sufficient condition for testing justifiable fairness, which uses only the training data $D, \Pr$ (Sec. 2) and does not require access to the causal graph G. We assume only that G and Pr are Markov compatible (Sec. 2.2). The training data has an additional response variable Y. As before, we assume a fairness application $(\mathcal{A}, S, \mathbf{A}, \mathbf{I})$ is given and that the algorithm is a good prediction of the response variable, *i.e.* $\Pr(Y = 1|\mathbf{X} = \mathbf{x}) \approx \Pr(O = 1|\mathbf{X} = \mathbf{x})$; we call the algorithm a *reasonable* classifier to indicate that it satisfies this condition. Note that this is a typical assumption in preprocessing approachs, see e.g., [7] and needed to decouple the the issues of model accuracy and fairness.

We first establish a technical condition for fairness based on the Markov boundary, and then we simplify it. Recall that, give a probability distribution Pr, the *Markov boundary* of a variable $Y \in \mathbf{V}$, denoted $\mathbf{MB}(Y)$, is a minimal subset of $\mathbf{V} - \{Y\}$ that satisfies the saturated conditional independence $(Y \perp\!\!\!\perp_{\Pr} \mathbf{V} - (\mathbf{MB}(Y) \cup \{Y\})|\mathbf{MB}(Y))$. Intuitively, $\mathbf{MB}(Y)$ shields Y from the influence of other variables. It is usually assumed that the Markov boundary of a variable is unique (see Appendix 8.1). We prove:

Theorem 3.7. *A sufficient condition for a fairness application* $(\mathcal{A}, S, \mathbf{A}, \mathbf{I})$ *to be justifiably fair is* $\mathbf{MB}(O) \subseteq \mathbf{A}$.

If Pr is faithful to the causal graph, then the theorem follows immediately from Theorem 3.5; but we prove it without assuming faithfulness in Appendix 8.2. The condition in Theorem 3.7 can be checked without knowing the causal DAG, but requires the computation of the Markov boundary; moreover, it is expressed in terms of the outcome O of the algorithm. We derive from here a sufficient condition without reference to the Markov boundary, which refers only to the response variable Y present in the training data.

Corollary 3.8. *Fix a training data* $D, \Pr$, *where* $Y \in \mathbf{V}$ *is the training label, and* $\mathbf{A}, \mathbf{I}$ *are admissible and inadmissible attributes. Then any reasonable classifier trained on a set of variables* $\mathbf{X} \subseteq \mathbf{V}$ *is justifiably fair w.r.t. a protected attribute* S, *if any of the following hold:*

(a) $\Pr$ *satisfies the CI* $(Y \perp\!\!\!\perp \mathbf{X} \cap \mathbf{I}|\mathbf{X} \cap \mathbf{A})$, *or*

(b) $\mathbf{X} \supseteq \mathbf{A}$ *and* $\Pr$ *satisfies the saturated CI* $(Y \perp\!\!\!\perp I|\mathbf{A})$.

The proof is the Appendix. While condition (a) is the weaker assumption, condition (b) has the advantage that the CI is saturated. Our method for building a fair classifier is to repair the training data in order to enforce (b).

3.3 Building Fair Classifiers

This leads us to the following methods for building justifiably fair classifiers.

Dropping Inadmissible Attributes. A naive way to satisfy Corollary 3.8(a) is to set $\mathbf{X} = \mathbf{A}$, in other words to train the classifier only on admissible attributes This method guarantees fairness, but, as we will show in Sec. 6, dropping even one inadmissible variable can negatively affect the accuracy of the classifier. Moreover, this approach cannot be used in data release situations, where all variables must be included. Releasing data that reflect discrimination can unintentionally reinforce and amplify discrimination in other contexts that data is used.

Repairing Training Data. Instead, our approach is to repair the training data to enforce the condition in Corollary 3.8(b). We consider the saturated CI $(Y \perp\!\!\!\perp \mathbf{I}|\mathbf{A})$ as an *integrity constraint* that should always hold in training data $D, \Pr$. Capuchin performs a sequence of database updates (viz., insertions and deletions of tuples) to obtain another training database D' to satisfy $(Y \perp\!\!\!\perp \mathbf{I}|\mathbf{A})$. We describe this repair problem in Sec. 4. To the causal DAG, this approach can be seen as modifying the underlying causal model to enforce the fairness constraint. However, instead of intervening on the causal DAG, which we do not know and over which we have no control, we intervene on the training data to ensure fairness. Note that minimal repairs are crucial for preserving the utility of data.

4 REPAIRING TRAINING DATA TO ENSURE FAIRNESS

We have shown in Corollary 3.8 that, if the training data D satisfies a certain saturated conditional independence (CI), then a classification algorithm trained on $D, \Pr$ is justifiably fair. We show here how to modify (repair) the training data to enforce the CI and thus ensure that any reasonable classifier trained on it will be justifiably fair.

4.1 Minimal Repair for MVD and CI

We first consider repairing an MVD. Fix an MVD $\mathbf{Z} \twoheadrightarrow \mathbf{X}$ and a database D that does not satisfy it. The minimal database repair problem is this: find another database D' that satisfies

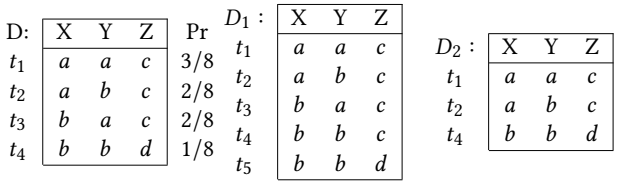


Figure 5: A simple database repair: D does not satisfy the MVD $Z \twoheadrightarrow X$. In D_1, we inserted the tuple (b, b, c) to satisfy the MVD, and in D_2 we deleted the tuple (b, a, c) to satisfy the MVD.

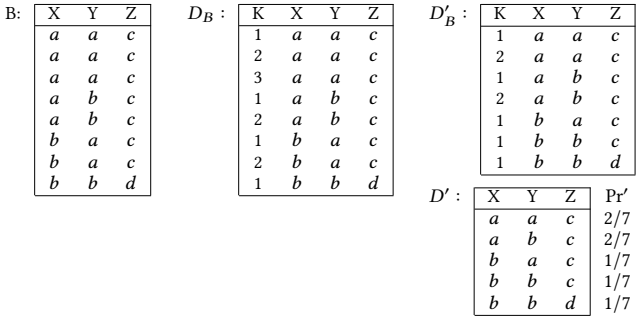


Figure 6: Repairing a conditional independence (CI).

the MVD such that the distance between D and D' is minimized. In this section, we restrict the distance function to the symmetric difference, i.e, $|\Delta(D, D')|$.

Example 4.1. Consider the database D in Fig. 5 (ignoring the probabilities for the moment), and the MVD $Z \twoheadrightarrow X$. D does not satisfy the MVD. The figure shows two minimal repairs, D_1, D_2, one obtained by inserting a tuple, and the other by deleting a tuple.

However, our problem is to repair for a saturated CI, not an MVD, since that is what is required in Corollary 3.8. The repair problem for a database constraint is well-studied in the literature, but here we need to repair to satisfy a CI, which is not a database constraint. We first formally define the repair problem for a CI then show how to reduce it to the repair for an MVD. More precisely, our input is a database D and a probability distribution Pr, and the goal is to define a "repair" $D', \Pr'$ that satisfies the given CI.

We assume that all probabilities are rational numbers. Let the *bag associated* to $D, \Pr$ to be the smallest bag B such that Pr is the empirical distribution on B. In other words, B is obtained by replicating each tuple $t \in D$ a number of times proportional to $\Pr(t)$. [4] If Pr is uniform, then $B = D$.

Definition 4.2. The minimal repair of $D, \Pr$ for a saturated CI $(\mathbf{X}; \mathbf{Y}|\mathbf{Z})$ is a pair $D', \Pr'$ such that $\Pr'$ satisfies the CI and $|\Delta(B, B')|$ is minimized, where B and B' are the bags associated to $D, \Pr$ and $D', \Pr'$, respectively.

Recall that $\mathbf{V}$ denotes the set of attributes of D. Let Pr be any probability distribution on the variables $\{K\} \cup \mathbf{V}$, where K is a fresh variable not in $\mathbf{V}$.

Lemma 4.3. *If* Pr *satisfies* $(K\mathbf{X}; \mathbf{Y}|\mathbf{Z})$*, then it also satisfies* $(\mathbf{X}; \mathbf{Y}|\mathbf{Z})$.

The lemma follows immediately from the Decomposition axiom in Graphoid (see Appendix 8.1).

We now describe our method for computing a minimal repair of $D, \Pr$ for some saturated CI. First, we compute the bag B associated to $D, \Pr$. Next, we add the new attribute K to the tuples in B and assign distinct values to $t.K$ to all duplicate tuples t, thus converting B into a set D_B with attributes $K \cup \mathbf{V}$. Importantly, we use as few distinct values for K as possible, i.e., we enumerate the instances of each unique tuple. More precisely, we define:

$$D_B = \{(i, t) | t \in B, i = 1, \ldots, |t_B|\} \tag{8}$$

were $|t_B|$ denotes the number of occurrences (or multiplicity) of a tuple t in the bag B. Then, we repair D_B w.r.t. to the MVD $\mathbf{Z} \twoheadrightarrow K\mathbf{X}$, obtaining a repaired database D'_B. Finally, we construct a new training set $D' = \Pi_{\mathbf{V}}(D'_B)$, with the probability distribution obtained by marginalizing the empirical distribution on D'_B to the variables $\mathbf{V}$. We prove the following:

Theorem 4.4. *Let D be a database and* Pr *a probability distribution on its tuples, and let B be the associated bag (with attributes $\{K\} \cup \mathbf{V}$). Fix a saturated CI $(\mathbf{X}; \mathbf{Y}|\mathbf{Z})$, and let B' be a minimal repair for the MVD $\mathbf{Z} \twoheadrightarrow K\mathbf{X}$. Then, $D', \Pr'$ is a minimal repair of $D, \Pr$ for the CI, where D' is B' with duplicates removed, and $\Pr'$ is the empirical distribution on B'.*

We illustrate with an example.

Example 4.5. In Example 4.1 we showed two repairs D_1, D_2 of the database D in Fig 5 for the MVD $Z \twoheadrightarrow X$. Consider now the probability distribution, Pr shown in the figure. Suppose we want to repair it for the CI $(X; Y|Z)$. Clearly, both D_1 and D_2, when endowed with the empirical distribution *do* satisfy this CI, but they are very poor repairs because they completely ignore the probabilities in the original training data, which are important signals for learning. Our definition captures this by insisting that the repaired bag B' be close to the bag B associated to $D, \Pr$ (see B in Fig. 6), but the sets D_1 and D_2 are rather far from B. Instead, our method first converts B into a set D_B by adding a new attribute K (see Fig. 6) then, it repairs D_B for the MVD $Z \twoheadrightarrow KX$, obtaining D'_B. The final repair $D', \Pr'$ consists of the empirical distribution on D'_B, but with the attribute K and duplicates removed.

We note that, in order for Theorem 4.4 to hold, it is critical that we use minimum distinct values for the attribute K in

[4] Equivalently, if the tuples have probabilities $p_1/q, p_2/q, \ldots$ (same denominator), then each tuple t_i occurs exactly p_i times in B.

D_B; otherwise minimal repairs of D_B are no longer minimal repairs of the original data $D, \Pr'$. For example, if we use distinct values for K, thus making K a key, then only subset of D_B that satisfies the MVD $\mathbf{Z} \twoheadrightarrow K\mathbf{X}$ is the empty set.

4.2 Reducing Minimal Repair to 3SAT

Corollary 3.8 requires us to repair the training data D to satisfy a CI. We have shown how to convert this problem into the problem of repairing a derived data D_B to satisfy an MVD. In this section we describe how to find a minimal repair for an MVD by reduction to the weighted MaxSAT problem.

We denote the database by D, the MVD by $\varphi : \mathbf{Z} \twoheadrightarrow \mathbf{X}$, and assume that D's attributes are $\mathbf{V} = \mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$. Recall that D satisfies the MVD iff $D = \Pi_{\mathbf{XZ}}(D) \bowtie \Pi_{\mathbf{YZ}}(D)$. Since we want to allow repairs that include both insertions and deletions, we start by finding an upper bound on the set of tuples that we may want to insert in the database. For example, one can restrict the set of tuples to those that have only constants that already occurring in the database, i.e., an upper bound is $ADom^k$, where $ADom$ is the active domain of D, and k is the arity of D. However, this set is too large in practice. Instead, we prove that it suffices to consider candidate tuples in a much smaller set, given by: $D^* \stackrel{\text{def}}{=} \Pi_{\mathbf{XZ}}(D) \bowtie \Pi_{\mathbf{ZY}}(D)$.

Proposition 4.1. *Any minimal repair D' of D for an MVD satisfies $D' \subseteq D^*$.*

Next, we associate the following Boolean Conjunctive query to the MVD φ:

$$Q_\varphi \leftarrow D(\mathbf{X}_1, \mathbf{Y}_1, \mathbf{Z}), D(\mathbf{X}_2, \mathbf{Y}_2, \mathbf{Z}), \neg D(\mathbf{X}_1, \mathbf{Y}_2, \mathbf{Z}) \tag{9}$$

It follows immediately that $D \not\models \varphi$ iff $D \models Q_\varphi$, and therefore the repair problem becomes: modify the database D to make Q_φ false. For that purpose, we use the lineage of the query Q_φ. By the previous proposition, we know that we need to consider as candidates for insertions only those tuples in D^*; hence we compute the lineage over the set of possible tuples D^*. We briefly review here the construction of the lineage and refer the reader to [56] and the references there for more detail. We associate a distinct Boolean variable X_t to each tuple $t \in D^*$, and consider all mappings $\theta : Var(Q_\varphi) \to ADom(D)$ such that each of the three tuples– $D^*(\theta(X_1), \theta(Y_1), \theta(Z)), D^*(\theta(X_2), \theta(Y_2), \theta(Z))$, and $D^*(\theta(X_1), \theta(Y_2), \theta(Z))$– are in D^*. Then, the lineage and its negation are:

$$\Phi_\varphi = \bigvee_\theta \left(X_{D^*(\theta(X_1),\theta(Y_1),\theta(Z))} \wedge X_{D^*(\theta(X_2),\theta(Y_2),\theta(Z))} \wedge \neg X_{D^*(\theta(X_1),\theta(Y_2),\theta(Z))} \right) \tag{10}$$

$$\neg\Phi_\varphi = \bigwedge_\theta \left(\neg X_{D^*(\theta(X_1),\theta(Y_1),\theta(Z))} \vee \neg X_{D^*(\theta(X_2),\theta(Y_2),\theta(Z))} \vee X_{D^*(\theta(X_1),\theta(Y_2),\theta(Z))} \right) \tag{11}$$

Recall that an *assignment* is a mapping from Boolean variables X_t to $\{0, 1\}$. Thus, our goal is to find an assignment satisfying the 3CNF $\neg\Phi_\varphi$, which is as close as possible to the initial assignment $X_t = 1$ for $t \in D$, $X_t = 0$ for $t \in D^* - D$.

We briefly review the weighted MaxSAT problem here. Its input is a 3CNF $\mathbf{F}$ whose clauses are partitioned into $\mathbf{F} = (\mathbf{F}_h, \mathbf{F}_s, C)$, where $\mathbf{F}_h$ are called the hard clauses, and $\mathbf{F}_s$ are the soft clauses, and a function $\mathbf{C} : \mathbf{F}_s \to R^+$ associates a non-negative cost with each soft clause. A solution to the problem finds an assignment that satisfies all hard constraints, and maximizes the weight of the satisfied soft constraints.

To ensure "closeness" to the initial assignment, we add to the Boolean formula a clause X_t for every $t \in D$, and a clause $\neg X_t$ for every $t \in D^* - D$. The final 3CNF formula is:

$$\Psi = \underbrace{(\neg\Phi_\varphi)}_{\text{hard clauses}} \wedge \underbrace{\bigwedge_{t \in D} X_t \wedge \bigwedge_{t \in D^* - D} (\neg X_t)}_{\text{soft clauses}}$$

The algorithm constructing Ψ is shown in Algorithm 1.

Algorithm 1: Converts the problem of finding a database repair w.r.t. a CI statement into solving a general CNF formula.

Input: A database D with vairables $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$ and a saturated CI $\varphi : (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z})$
Output: A 3CNF Ψ consisting of hard and soft clauses.

Compute $D^*(\mathbf{X}_1, \mathbf{Y}_2, \mathbf{Z}) = D(\mathbf{X}_1, \mathbf{Y}_1, \mathbf{Z}) \wedge D(\mathbf{X}_2, \mathbf{Y}_2, \mathbf{Z})$
for $t \in D^*$ **do**
  If $t \in D$, add the soft clause X_t to Ψ
  If $t \in D^* - D$ add the soft clause $(\neg X_t)$ to Ψ
Compute $C(\mathbf{X}_1, \mathbf{Y}_1, \mathbf{X}_2, \mathbf{Y}_2, \mathbf{Z}) = D^*(\mathbf{X}_1, \mathbf{Y}_1, \mathbf{Z}) \wedge D^*(\mathbf{X}_2, \mathbf{Y}_2, \mathbf{Z})$
for $t \in C$ **do**
  $t_1 \leftarrow t[\mathbf{X}_1, \mathbf{Y}_1, \mathbf{Z}]$; $t_2 \leftarrow t[\mathbf{X}_2, \mathbf{Y}_2, \mathbf{Z}]$; $t_3 \leftarrow t[\mathbf{X}_1, \mathbf{Y}_2, \mathbf{Z}]$
  Add the hard clause $(\neg X_{t_1} \vee \neg X_{t_2} \vee X_{t_3})$ to Ψ

Example 4.6. Continuing Ex. 4.1, we observe that $D^* = D_1$; hence, there are 5 possible tuples. The lineage expression for φ and it negation are:

$$\Phi_\varphi = (X_{t_1} \wedge X_{t_4} \wedge \neg X_{t_2}) \vee (X_{t_2} \wedge X_{t_3} \wedge \neg X_{t_1}) \vee (X_{t_3} \wedge X_{t_2} \wedge \neg X_{t_4}) \vee (X_{t_4} \wedge X_{t_1} \wedge \neg X_{t_3})$$

Hence,

$$\neg\Phi_\varphi = (\neg X_{t_1} \vee \neg X_{t_4} \vee X_{t_2}) \wedge (\neg X_{t_2} \vee \neg X_{t_3} \vee X_{t_1}) \wedge (\neg X_{t_3} \vee \neg X_{t_2} \vee X_{t_4}) \wedge (\neg X_{t_4} \vee \neg X_{t_1} \vee X_{t_3})$$

The reader can check that the repairs D_1 and D_2 in Ex. 4.1 are corresponded to some satisfying assignment of $\neg\Phi_\varphi$, e.g., D_2 obtained from the truth assignment $\sigma(X_{t_1}) = \sigma(X_{t_2}) = 1$, $\sigma(X_{t_3}) = \sigma(X_{t_5}) = 0$; both satisfy all clauses in $\neg\Phi_\varphi$. The formula Ψ that we give as input to the weighted MaxSAT consists of $\neg\Phi_\varphi$ plus these five clauses: $X_{t_1} \wedge X_{t_2} \wedge X_{t_3} \wedge X_{t_4} \wedge \neg X_{t_5}$, each with cost 1. MaxSAT will attempt to satisfy as many as possible, thus finding a repair that is close to the initial database D.

Note that repairing a database w.r.t. a CI $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})$ can be reduced to repairing subsets $\sigma_{\mathbf{Z}=\mathbf{z}}(D)$ for $\mathbf{z} \in Dom(Z)$ w.r.t. the marginal independence $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y})$. Therefore, the problem is highly parallelizable. CAPUCHIN partition subsets $\Pi_{\mathbf{Z}}(D)$ into chunks of even size (if possible) and repairs them in parallel (see Sec 6.4).

Algorithm 2: Repair using Matrix Factorization.

Input: A bag B with attributes $\mathbf{V} = \mathbf{XYZ}$ a CI statment $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})$.
Output: B' a repair of B

```
for z ∈ Dom(Z) do
    M_X^{B'}, M_Y^{B'} ← Factorize(M_{X,Y}^{B_z})
    M_{X,Y}^{B'z} ← (1/|B_Z|) M_X^{B'⊤} M_Y^{B'}
return B' associated with M_{X,Y,Z}^{B'} = {M_{X,Y}^{B'z}}
```

4.3 Repair via Matrix Factorization

In this section, we use matrix factorization to repair a bag w.r.t. a CI statement. We are given a bag B to which we associate the empirical distribution $\Pr(\mathbf{v}) = \frac{1}{|B|}\sum_{t\in B} 1_{t=\mathbf{v}}$, and a CI statement $\varphi : (\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})$ such that B is inconsistent with φ, meaning that φ does not hold in Pr. Our goal is to find a repair of B, i.e., a bag B' that is close to B such that $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|_{\Pr'}\mathbf{Z})$, where $\Pr'$ is the empirical distribution associated to B'.

First, we review the problem of non-negative rank-one matrix factorization. Given a matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$, the problem of *rank-one nonnegative matrix factorization (NMF)* is the minimization problem: $\text{argmin}_{\mathbf{U}\in\mathbb{R}_+^{n\times 1},\mathbf{V}\in\mathbb{R}_+^{1\times m}} \|\mathbf{M} - \mathbf{UV}\|_F$, where $\mathbb{R}_+$ stands for non-negative real numbers and $\|.\|_F$ is the Euclidean norm of a matrix.[5]

We express the connection between our repair problem and the NMF problem using contingency matrices. Given three disjoint subsets of attributes $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, let $m = |Dom(\mathbf{X})|$, $n = |Dom(\mathbf{Y})|$, $k = |Dom(\mathbf{Z})|$ and $B_{\mathbf{z}} = \sigma_{\mathbf{Z}=\mathbf{z}}(B)$. A *multiway-contingency matrix* over $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ consists of k $n \times m$ matrices $\mathbf{M}^B_{\mathbf{X},\mathbf{Y},\mathbf{Z}} = \{\mathbf{M}^{B_{\mathbf{z}}}_{\mathbf{X},\mathbf{Y}} | \mathbf{z} \in Dom(\mathbf{Z})\}$ where, $\mathbf{M}^{B_{\mathbf{z}}}_{\mathbf{X},\mathbf{Y}}(ij) = \sum_{t\in B} 1_{t[\mathbf{XY}]=ij}$. Intuitively, $\mathbf{M}^{B_{\mathbf{z}}}_{\mathbf{X},\mathbf{Y}}(ij)$ represents the joint frequency of $\mathbf{X}$ and $\mathbf{Y}$ in a subset of bag with $\mathbf{Z} = \mathbf{z}$.

The following obtained immediately from the connection between independence and rank of a contingency matrix.

PROPOSITION 4.2. *Let B be a bag and* Pr *be the empirical distribution associated to B. It holds that $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|_{\Pr}\mathbf{Z})$ iff each contingency matrix $\mathbf{M} \in \mathbf{M}^B_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$ is of rank-one.*

We illustrate with an example.

Example 4.7. Let $\mathbf{M}_1 = \begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}$, $\mathbf{M}_2 = \begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}$, $\mathbf{M}_3 = \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}$, $\mathbf{M}_4 = \begin{bmatrix}1 & 1\\ 0 & 0\end{bmatrix}$. The following contingency matrices are associated to D, D_1 and D_2 in Ex. 4.1: $\mathbf{M}^D_{\mathbf{X},\mathbf{Y},\mathbf{Z}} = \{\mathbf{M}_1, \mathbf{M}_2\}$, $\mathbf{M}^{D_1}_{\mathbf{X},\mathbf{Y},\mathbf{Z}} = \{\mathbf{M}_3, \mathbf{M}_2\}$ and $\mathbf{M}^{D_2}_{\mathbf{X},\mathbf{Y},\mathbf{Z}} = \{\mathbf{M}_4, \mathbf{M}_2\}$. The reader can verify that $\mathbf{M}_2, \mathbf{M}_3$ and $\mathbf{M}_4$ are of rank-one but $\mathbf{M}_1$ is not. It is clear that, D is inconsistent with φ but D_1 and D_2 are consistent with φ.

The following implied from NP-hardness of NMF [50].

PROPOSITION 4.3. *The problem of repairing a database w.r.t. a single CI is NP-hard in general.*

Based on Prop 4.2, we propose Algorithm 2 for repairing a bag w.r.t. a single CI $\varphi : (\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})$. The algorithm works as follows: for each $\mathbf{z} \in Dom(\mathbf{Z})$, it uses the **Factorize** subroutine to factorize the $n \times m$ contingency matrix $\mathbf{M}^{B_{\mathbf{z}}}_{\mathbf{X},\mathbf{Y}}$ into a $1 \times n$ matrix $\mathbf{M}^{B'}_{\mathbf{X}}$ and a $1 \times m$ matrix $\mathbf{M}^{B'}_{\mathbf{Y}}$. Then, it uses the product of $\mathbf{M}^{B'\intercal}_{\mathbf{X}}$ and $\mathbf{M}^{B'}_{\mathbf{X}}$ to construct a new bag B'. It is clear that $\mathbf{M}^{B'\intercal}_{\mathbf{X}}\mathbf{M}^{B'}_{\mathbf{Y}}$ is of rank-one by construction; thus, the algorithm always returns a bag B' that is consistent with φ. Note that any off-the-shelf NMF algorithm (such as [17]) can be used in Algorithm 2, to minimize the Euclidean distance between Pr and Pr′, the empirical distributions associated to B and B', respectively. In addition, we use the simple factorization of $\mathbf{M}^{B_{\mathbf{z}}}_{\mathbf{X},\mathbf{Y}}$ into $\mathbf{M}^{B_{\mathbf{z}}}_{\mathbf{X}}$ and $\mathbf{M}^{B_{\mathbf{z}}}_{\mathbf{Y}}$, i.e., the marginal frequencies of $\mathbf{X}$ and $\mathbf{Y}$ in $B_{\mathbf{z}}$. We refer to this simple factorization as *Independent Coupling (IC)*. It is easy to see that KL-divergence between Pr and Pr′ is bounded by conditional mutual information $I(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z})$.

5 DISCUSSION

Generalizability to Unseen Test Data. In the following we briefly discuss the generalizability of the proposed repair algorithm to unseen test data. Recall that the bag B represents the training data, B' its repair, and let T be the unseen test data. We prove the following in Appendix 8.2:

LEMMA 5.1. *If the repaired data satisfies $(Y \perp\!\!\!\perp S, \mathbf{I}|_{\Pr_{B'}}\mathbf{A})$ and the unseen test data satisfies $\Pr_T(s,\mathbf{i}|\mathbf{a}) = \Pr_{B'}(s,\mathbf{i}|\mathbf{a})$, then the unseen test data also satisfies $(Y \perp\!\!\!\perp S, \mathbf{I}|_{\Pr_T}\mathbf{A})$*

The goal of repair is precisely to satisfy $(Y \perp\!\!\!\perp S, \mathbf{I}|_{\Pr_{B'}}\mathbf{A})$, hence the classifier trained on the repaired data B' will be justifiable fair on the test data T provided that $\Pr_T(s,\mathbf{i}|\mathbf{a}) = \Pr_{B'}(s,\mathbf{i}|\mathbf{a})$. It is generally assumed that the test and training data are drawn from the same distribution Pr. By the law of large numbers, the empirical distribution of i.i.d samples of size $N \to \infty$ converges to Pr, hence $\Pr_T = \Pr_B$ in the limit. Therefore, the algorithm will be justifiable fair on the test data, provided that the repair is done such that $\Pr_B(s,\mathbf{i}|\mathbf{a}) = \Pr_{B'}(s,\mathbf{i}|\mathbf{a})$. This condition is satisfied by the IC repair method which simply repair data by coupling marginal distributions, because it holds by construction that $\Pr_{B'}(y,s,\mathbf{i},\mathbf{a}) = \Pr_B(y,\mathbf{a})\Pr_B(s,\mathbf{i},\mathbf{a})/\Pr_B(\mathbf{a})$. In contrast, the condition is only approximately satisfied by the MaxSAT

[5] Recall that a matrix is of rank-one if and only if it can be represented by the outer product of two vectors.

and MF approaches, translating to slightly weaker fairness guarantees on unseen test data. Nevertheless, we empirically show in Sec 6 that MaxSAT and MF approaches maintain a significantly better balance between accuracy and fairness.

Scalability. As shown in Sec 4, repairing data w.r.t. a single CI is an NP-complete problem. Therefore, the scalability of our proposed repair methods is equal to that of MaxSAT solvers and approximation algorithms for matrix factorization. However, our repair problem is embarrassingly parallel and can be scaled to large datasets by partitioning data into small chunks formed by the conditioning set (see Sec 6). In this paper we focused on a single CI, which suffices for many real world fairness applications. We leave the natural extension to future work.

6 EXPERIMENTAL RESULTS

This section presents experiments that evaluate the feasibility and efficacy of CAPUCHIN. We aim to address the following questions. **Q1**: What is the end-to-end performance of CAPUCHIN in terms of utility and fairness, with respect to our different algorithms? **Q2**: To what extent are the repaired datasets modified by the repair process of CAPUCHIN? **Q3**: How does CAPUCHIN compare to state-of-the-art preprocessing methods for enforcing fairness in predictive classification algorithms? Table 2 reports the running time of the repair algorithms.

6.1 Degree of Discrimination

To assess the effectiveness of the proposed approaches, we next propose a metric that quantifies the degree of discrimination of a classification algorithm.

If we have access to the causal DAG, we could directly compute the *degree of interventional discrimination* of an algorithm: given admissible variables $\mathbf{A}$, for each $\mathbf{K} \supseteq \mathbf{A}$, compute the ratio of the LHS and RHS of Eq. 17 using Theorem 2.1, and average the results. However, in many practical settings we must make judgments about the fairness of an algorithm whose inputs are unknown. We cannot assume access to an underlying causal DAG in these situations. Instead, we propose a new metric for discovering evidence of potential discrimination from data that uses the causal framework we described but is still applicable in situations where all we know is which attributes in the Markov boundary of O are admissible.

Definition 6.1. Given afairness application $(\mathcal{A}, S, \mathbf{A}, \mathbf{I})$, let $\mathbf{A}_b = \mathbf{MB}(O) - \mathbf{I}$. We quantify the *ratio of observational discrimination (ROD)* of $\mathcal{A}$ against S in a context $\mathbf{A}_b = \mathbf{a}_b$ as $\delta(S; O|\mathbf{a}_b) \stackrel{\text{def}}{=} \frac{\Pr(O=1|S=0,\mathbf{a}_b)\Pr(O=0|S=1,\mathbf{a}_b)}{\Pr(O=0|S=0,\mathbf{a}_b)\Pr(O=1|S=1,\mathbf{a}_b)}$.

Intuitively, ROD calculates the effect of membership in a protected group on the odds of the positive outcome of $\mathcal{A}$ for subjects that are similar on $\mathbf{A}_b = \mathbf{a}_b$ ($\mathbf{A}_b$ consists of admissible attributes in the Markov boundary of the outcome). If $\delta(S; O|\mathbf{a}_b) = 1$, then there is no observational evidence that $\mathcal{A}$ is discriminatory toward subjects with similar characteristics $\mathbf{a}_b$. If $\delta(S; O|\mathbf{a}_b) > 1$, then the algorithm potentially discriminates against the protected group, and vice versa if $\delta(S; O|\mathbf{a}_b) < 1$. ROD is sensitive to the choice of a context $\mathbf{A}_b = \mathbf{a}_b$ by design. The overall ROD denoted by $\delta(S, O|\mathbf{A}_b)$ can be computed by averaging $\delta(S, O|\mathbf{a}_b)$ for all $\mathbf{a}_b \in \mathbf{A}_b$. It is easy to see for faithful distributions that ROD=1 coincides with justifiable fairness (see Prop 8.2 in the Appendix 8.2).

Dataset	Att. [#]	Rows[#]	IC	MF	MS(H.)	MS(S.)
Adult [28]	10	48k	12	20	40	30
Binned Adult [7]	4	48k	2	3	20	NA
COMPAS [49]	7	7k	2	3	7	8
Binned COMPAS [7]	5	7k	2	3	9	NA

Table 2: Runtime in seconds for experiments in Sec. 6.3.

6.2 Setup

The datasets used for experiments are listed in Table 2. We implemented our MaxSAT encoding algorithm in Python. For every instance of the input data, our algorithm constructed the appropriate data files in WCNF format. We used the Open-WBO [33] solver to solve the weighted MaxSAT instances.

We report the empirical utility of each classifier using Accuracy (ACC) = $\frac{TP+TN}{TP+FP+FN+TN}$ via 5-fold cross-validation. We evaluate using three classifiers: Linear Regression (LR), Multi-layer Perceptron (MLP), and Random Forest (RF).

We evaluated using the fairness metrics in Table 3. For computing these metrics, conditional expectations were estimated as prescribed in [44]. We used standard techniques in meta-analysis to compute the pooled odds ratio [8], and its statistical significance, needed to compute ROD. Specifically, we reported the p-value of the ROD, where the null hypothesis was ROD=1; (low p-values suggest the observed ROD is not due to random variation). We combined the p-values from cross-validation test datasets using Hartung's method [20]; p-values were dependent due to the overlap in cross-validation tests. We normalized ROD between 0 and 1, where 0 shows no observational discrimination. We reported the absolute value of the averages of all metrics computed from each test dataset, where the smaller the value, the less the discrimination exhibited by the classifier.

6.3 End-To-End Results

In the following experiments, a fairness constraint was enforced on training data using CAPUCHIN repair algorithms (cf. Sec 4). Specifically, each dataset was split into five training and test datasets. All training data were repaired separately using Matrix Factorization (MF), Independent Coupling (IC) and two versions of the MaxSAT approach: MS(Hard), which

Metric	**Description** and **Definition**
ROD	Ratio of Observation Discrimination: (See Sec.6.1)
DP	Demographic Parity: $Pr(O=1\|S=1) - Pr(O=1\|S=0)$
TPB	True Positive Rate Balance: $Pr(O=1\|S=1, Y=1) - Pr(O=1\|S=0, Y=1)$
TNB	True Negative Rate Balance: $Pr(O=0\|S=1, Y=0) - Pr(O=0\|S=0, Y=0)$
CDP	Conditional Statistical Parity: $\mathbb{E}_{\mathbf{a}}[Pr(O=1\|S=1, \mathbf{a}) - Pr(O=1\|S=0, \mathbf{a})]$
CTPB	Conditional TPRB: $\mathbb{E}_{\mathbf{a}}[Pr(O=1\|S=1, Y=1, \mathbf{a}) - Pr(O=1\|S=0, Y=1, \mathbf{a})]$
CTNB	Conditional TNRB: $\mathbb{E}_{\mathbf{a}}[\Pr(O=0\|S=1, Y=0, \mathbf{a}) - \Pr(O=0\|S=0, Y=0, \mathbf{a})]$

Table 3: Fairness metrics used in our experiments.

feeds all clauses of the lineage of a CI into MaxSAT, and MS(Soft), which only feeds small fraction of the clauses. We tuned MaxSAT to enforce CIs approximately. We then measured the utility and discrimination metrics for each repair method as explained in Sec 6.2. For all datasets, the chosen training variables included the Markov boundary of the outcome variables, which were learned from data using the Grow-Shrink algorithm [32] and permutation [44].

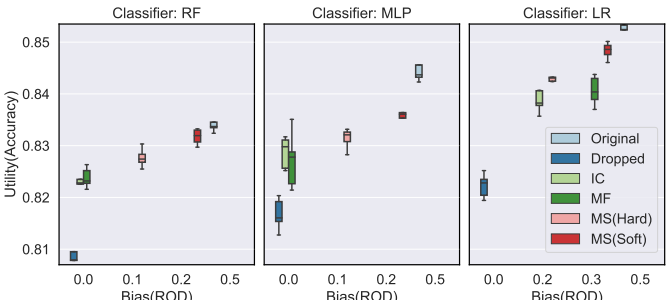


Figure 7: Performance of CAPUCHIN on Adult data.

Adult data. Using this dataset, several prior efforts in algorithmic fairness have reported gender discrimination based on a strong statistical dependency between income and gender in favor of males [31, 48, 60]. However, it has been shown that Adult data is inconsistent [44] because its income attribute reports household income for married individuals, and there are more married males in data. Furthermore, data reflects the historical income inequality that can be reinforced by ML algorithms. We used CAPUCHIN to remove the mentioned sources of discrimination from Adult data. Specifically, we categorized the attributes in Adult data as follows: (**S**) sensitive attributes: gender (male, female); (**A**) admissible attributes: hours per week, occupation, age, education, etc.; (**N**) inadmissible attributes: marital status; (Y) binary outcome: high income. As is common in the literature, we assumed that the potential influence of gender on income through some or all of the admissible variables was fair; However, the direct influence of gender on income, as well as its indirect influence on income through marital status, were assumed to be discriminatory. To remove the bias, we enforced the CI ($Y \perp\!\!\!\perp \mathbf{S}, \mathbf{N} | \mathbf{D}$) on training datasets using the CAPUCHIN repair algorithms. Then, we trained the classifiers on both original and repaired training datasets using the set of variables $\mathbf{A} \cup \mathbf{N} \cup \mathbf{S}$. We also trained the classifiers on original data using only **A**, i.e., we dropped the sensitive and inadmissible variables.

Fig. 7 compares the utility and bias of CAPUCHIN repair methods on Adult data. As shown, all repair methods successfully reduced the ROD for all classifiers. As shown in Fig. 8 (see also Fig 15 in the appendix), the repaired data also improved associational fairness measures: the CAPUCHIN repair methods had an effect similar to dropping the sensitive and inadmissible variables completely, but they delivered much higher accuracy (because the CI was enforced approximately). The residual bias after repair was expected since: (1) the classifier was only an approximation, and (2) we did not repair the test data. However, as shown in most cases, the residual bias indicated by ROD was not statistically significant. This shows that our methods are robust (by design) to the mismatch between the distribution of repaired data and test data. These repair methods delivered surprisingly good results: when partially repairing data using the MaxSAT approach, i.e, using MS(Soft), almost 50% of the bias was removed while accuracy decreased by only 1%. We also note that the residual bias generally favored the protected group (as opposed to the bias in the original data).

COMPAS. For the second experiment, we used the ProPublica COMPAS dataset [27]. This dataset contains records for all offenders in Broward County, Florida in 2013 and 2014. We categorized the attributes in COMPAS data as follows: (**S**) protected attributes: race (African American, Caucasian); (**A**) admissible attributes: number of prior convictions, severity of charge degree, age; **(Y)** binary outcome: a binary indicator of whether the individual is a recidivist. As is common in the literature, we assumed that it was fair to use the admissible attributes to predict recidivism even though they can potentially be influenced by race, and our only goal in this experiment was to address the direct influence of race. We pursued the same steps as explained in the first experiment. Fig. 9 compares the bias and utility of CAPUCHIN repair methods to original data. As shown, all repair methods successfully reduced the ROD. However, we observed that MF and IC performed better than MS on COMPAS data (as opposed to Adult data); see 6.4 for an explanation.

6.4 Comparing CAPUCHIN Repair Methods

To compare CAPUCHIN repair methods beyond the utility experiments in Sec 6.3, we compared the number of tuples

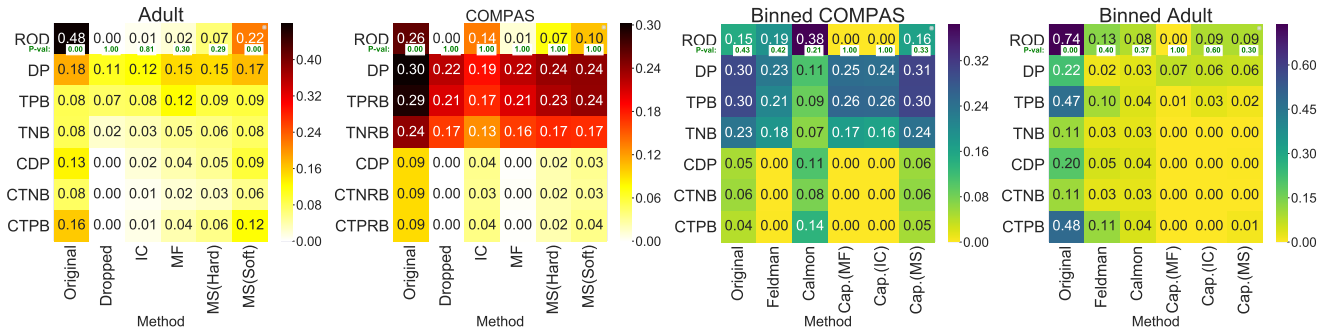


Figure 8: Bias reduction performance of CAPUCHIN for MLP classifier.

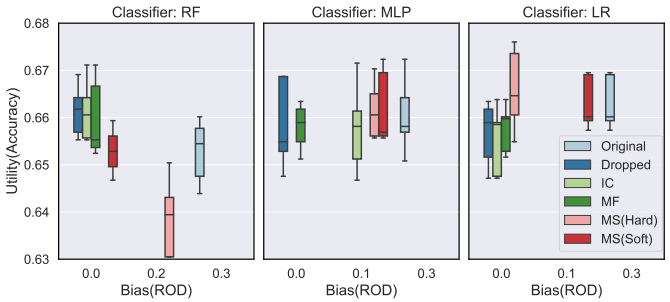


Figure 9: Performance of CAPUCHIN on COMPAS data.

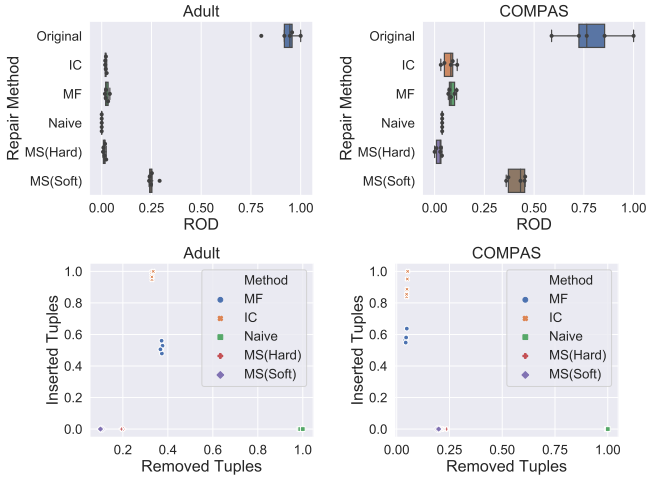


Figure 10: Comparison of different repair methods.

added and deleted for each method, as well as the bias reduction on training data. Fig 10 reports these measures for the experiments in Sec 6.3. Note that all numbers were normalized between 0 and 1, where ROD=1 shows no discrimination. For Adult data, we tuned the MS approach to repair data only by tuple deletion and compared it to a naive approach that repaired data using lineage expression but without using the MaxSat solver. As shown in Fig 10, the MaxSat approach removed up to 80% fewer tuples than the naive approach.

In general, the MaxSAT approach was the most flexible repair method (since it can be configured for partial repairs). Further, it achieved better classification accuracy, and it balanced tuple insertion and deletion. In terms of the utility of classification, the MS approach performed better on sparse data in which the conditioning groups consisted of several attributes. Figure 11 shows that repairing a very small fraction of inconsistencies (i.e., clauses in the lineage expression of the associated CI) in the experiment conducted on Adult data (Sec 6.3) led to a significant discrimination reduction. This optimization makes the MS approach more appealing in terms of balancing bias and utility. However, for dense data, IC and MD performed better. This difference was because the size of the lineage expression grew very large when the conditioning sets of CIs consisted of only a few attributes.

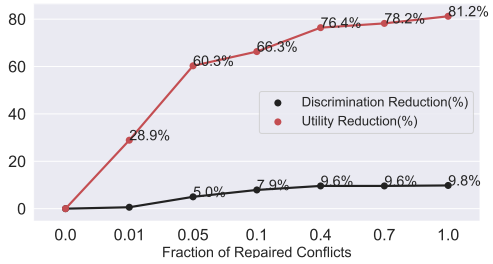
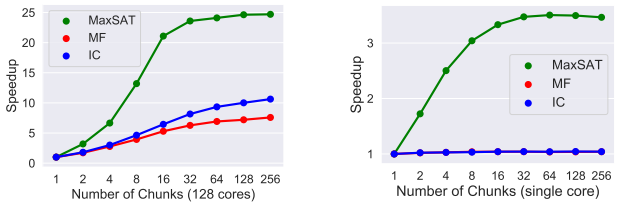


Figure 11: Bias-utility trade off in MaxSAT approach.

To evaluate the effect of partitioning and parallelizing on different methods, we replicated the experiment in sec 6.3 and partitioned Adult data into several chunks of approximately equal sizes; we then repaired the chunks in parallel on a cluster of 128 cores. Fig 12 shows the achieved speed up; all approaches were parallelizable. Parallel processing was most appealing for MaxSAT since MaxSAT solvers were much more efficient on smaller input sizes. While partitioning had no effect on MF and IC on a single-core machine,



(a) (b)

Figure 12: Speed up achieved (a) by partitioning and parallel processing on 128 cores; (b) by partitioning on a single core.

as shown in Fig 12(b), it sped up MaxSAT . Note that partitioning data into several small chunks does not necessarily speed up the MaxSAT approach, since MaxSAT solver must be called for several small inputs. Hence, performance does not increase linearly by increasing the number of chunks. In general partitioning data into several instance of medium size delivers the best performance.

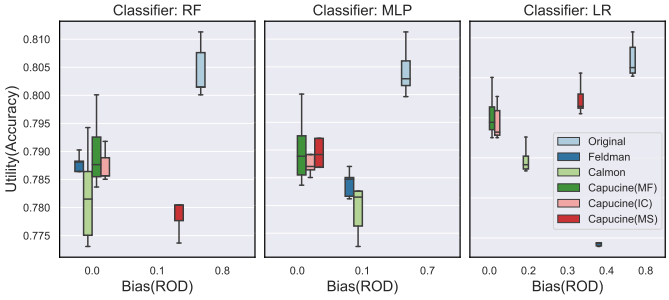


Figure 13: Comparison with other methods on Binned Adult data.

6.5 Comparing Capuchin to Other Methods

We compared Capuchin with two reference pre-processing algorithms, Feldman et al. [16] and Calmon et al. [7]. Feldman's algorithm modifies each attribute so that the marginal distributions based on the subsets of the attribute with a given sensitive value are all equal. Calmon's algorithm randomly transforms all variables except for the sensitive attribute to reduce the dependence between training labels and the sensitive attribute subject to the following constraints: (1) the joint distribution of the transformed data is close to the original distribution, and (2) individual attributes are not substantially distorted. Individual distortion is controlled for using a distortion constraint, which is domain dependent.

We used these algorithm only to repair training datasets and compare their bias and utility to Capuchin. In addition, since the distortion function required in Calmon's algorithm is completely arbitrary, we replicated the same experiments conducted in [16] using binned Adult data and binned COMPAS data. We note that the analysis in [16] was restricted to only a few attributes, and the data was excessively binned to few categories (to facilitate the definition of distortion function). As a result, the bias and utility obtained in this experiment was mismatched with Sec 6.3. For binned Adult data, the analysis was restricted to age, education and gender. For both datasets, we assumed all attributes were admissible.

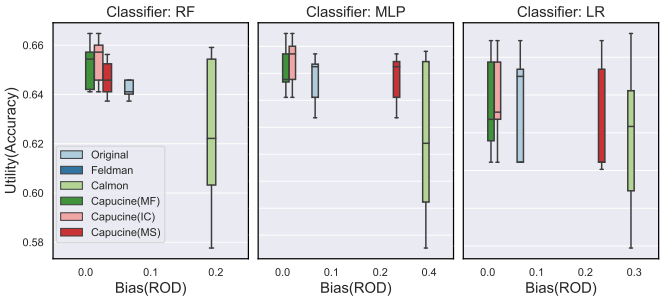


Figure 14: Comparison with other methods on Binned COMPAS.

Figs. 13 and 14 compares the utility and bias of Capuchin to the reference algorithms. The insights obtained from this experiment follow. For binned Adult data, all methods significantly reduced ROD, even though the goal of Calmon's and Fledman's algorithm is essentially to reduce DP. Similarly, Capuchin reduced DP and other associational metrics as a side effect. However, Capuchin outperformed both methods in terms of utility. Because COMPAS data was excessively binned, the ROD in training labels became insignificant for COMPAS, and accuracy dropped by 2%. We observe that both reference algorithms enforced DP at the cost of increasing ROD; however, in some cases the introduced bias was not statistically significant. In terms of utility, all methods of Capuchin (except for MaxSAT) performed better than Feldman's algorithm, and all Capuchin methods outperformed Calmon's algorithm quite significantly. This experiment shows that enforcing DP, while unnecessary, can severely affect the accuracy of a classifier and, even more importantly, introduce bias in sub-populations.

7 CONCLUSIONS

We considered a causal approach for fair ML, reducing it to a database repair problem. We showed that conventional associational and causal fairness metrics can over- and under-report discrimination. We defined a new notion of fairness, called as *justifiable fairness*, that addresses shortcoming of the previous definitions and arguably the strongest notion of fairness that is testable from data. We then proved sufficient properties for justifiable fairness and use these results to translate the properties into saturated conditional independences that we can be seen as multivalued dependencies with which to repair the data. We then propose multiple

algorithms for implementing these repairs. Our experimental results show that our algorithms not only outperform state-of-the-art pre-processing approaches for fairness on our own metrics, but that they are also competitive with existing approaches on conventional metrics. We empirically show that our methods are robust to unseen test data.

Acknowledgments: We thank the anonymous reviewers for their feedback. We also thank Prof. Ricardo Silva and Prof. Matt Kusner for useful discussions on counterfactual fairness. This work is supported by the National Science Foundation through NSF grants NSF III-1703281, III-1614738, NSF AITF 1535565 and III-1740996.

REFERENCES

[1] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[2] Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. 2005.

[3] Leopoldo E. Bertossi. *Database Repairing and Consistent Query Answering*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2011.

[4] Matthew T Bodie, Miriam A Cherry, Marcia L McCormick, and Jintong Tang. The law and policy of people analytics. *U. Colo. L. Rev.*, 88:961, 2017.

[5] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*, pages 13–18. IEEE, 2009.

[6] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[7] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3992–4001. Curran Associates, Inc., 2017.

[8] Bei-Hung Chang and David C Hoaglin. Meta-analysis of odds ratios: Current good practices. *Medical care*, 55(4):328, 2017.

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

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

[11] Rachel Courtland. Bias detectives: the researchers striving to make algorithms fair. *Nature*, 558, 2018.

[12] Jeffrey Dastin. Rpt-insight-amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, 2018. https://www.reuters.com/article/amazoncom-jobs-automation/rpt-insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSL2N1WP1RO.

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[14] Golnoosh Farnadi, Behrouz Babaki, and Lise Getoor. Fairness-aware relational learning and inference. In *Third International Workshop on Declarative Learning Based Programming (DeLBP) at thirty-second AAAI conference on Artificial Intelligence*, 2018.

[15] Golnoosh Farnadi, Behrouz Babaki, and Lise Getoor. Fairness in relational domains. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

[16] 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. ACM, 2015.

[17] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456, 2011.

[18] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pages 498–510. ACM, 2017.

[19] 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.

[20] Joachim Hartung. A note on combining dependent tests of significance. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 41(7):849–855, 1999.

[21] David Ingold and Spencer Soper. Amazon doesn't consider the race of its customers. should it? *Bloomberg*, 2016. www.bloomberg.com/graphics/2016-amazon-same-day/.

[22] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*, pages 1–6. IEEE, 2009.

[23] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

[24] 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*, pages 656–666, 2017.

[25] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

[26] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *CoRR*, abs/1703.06856, 2017.

[27] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9, 2016.

[28] M. Lichman. Uci machine learning repository, 2013.

[29] Ester Livshits, Benny Kimelfeld, and Sudeepa Roy. Computing optimal repairs for functional dependencies. In *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, Houston, TX, USA, June 10-15, 2018*, pages 225–237, 2018.

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

[31] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 502–510. ACM, 2011.

[32] Dimitris Margaritis. Learning bayesian network model structure from data. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2003.

[33] Ruben Martins, Vasco Manquinho, and Inês Lynce. Open-wbo: A modular maxsat solver. In *International Conference on Theory and Applications of Satisfiability Testing*, pages 438–445. Springer, 2014.

[34] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2018, page 1931. NIH Public Access, 2018.

[35] Richard E Neapolitan et al. *Learning bayesian networks*, volume 38. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

[36] Judea Pearl. *Causality*. Cambridge university press, 2009.

[37] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.

[38] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

[39] Judea Pearl and Azaria Paz. *Graphoids: A graph-based logic for reasoning about relevance relations*. University of California (Los Angeles). Computer Science Department, 1985.

[40] Donald B Rubin. *The Use of Matched Sampling and Regression Adjustment in Observational Studies*. Ph.D. Thesis, Department of Statistics, Harvard University, Cambridge, MA, 1970.

[41] Donald B Rubin. Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.

[42] Donald B Rubin. Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, 103(484):1350–1353, 2008.

[43] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, pages 6414–6423, 2017.

[44] Babak Salimi, Johannes Gehrke, and Dan Suciu. Bias in olap queries: Detection, explanation, and removal. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1021–1035. ACM, 2018.

[45] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Capuchin: Causal database repair for algorithmic fairness. *arXiv preprint arXiv:1902.08283*, 2019.

[46] Andrew D Selbst. Disparate impact in big data policing. *Ga. L. Rev.*, 52:109, 2017.

[47] Camelia Simoiu, Sam Corbett-Davies, Sharad Goel, et al. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

[48] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *Security and Privacy (EuroS&P), 2017 IEEE European Symposium on*, pages 401–416. IEEE, 2017.

[49] Jennifer Valentino-Devries, Jeremy Singer-Vine, and Ashkan Soltani. Websites vary prices, deals based on usersâĂŹ information. *Wall Street Journal*, 10:60–68, 2012.

[50] Stephen A Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2009.

[51] Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 440:1–440:14, New York, NY, USA, 2018. ACM.

[52] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, pages 1–7, New York, NY, USA, 2018. ACM.

[53] Lauren Weber and Elizabeth Dwoskin. Are workplace personality tests fair? *Wall Strreet Journal*, 2014.

[54] SK Michael Wong, Cory J. Butz, and Dan Wu. On the implication problem for probabilistic conditional independency. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(6):785–805, 2000.

[55] Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1920–1953, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

[56] Jane Xu, Waley Zhang, Abdussalam Alawini, and Val Tannen. Provenance analysis for missing answers and integrity repairs. *Data Engineering*, page 39, 2018.

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

[58] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 962–970, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

[59] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

[60] Indre Žliobaite, Faisal Kamiran, and Toon Calders. Handling conditional discrimination. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 992–1001. IEEE, 2011.

8 APPENDIX

8.1 Additional Background

Implication Problem for CIs. The implication problem for CI is the problem of deciding whether a CI φ is logically follows from a set of CIs Σ, meaning that in every distribution in which Σ holds, φ also holds. The following set of sound but incomplete axioms, known as *Graphoid*, are given in [39] for this implication problem.

- **(Symmetry)**
$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}) \rightarrow (\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{Z}) \tag{12}$$
- **(Decomposition)**
$$(\mathbf{X} \perp\!\!\!\perp \mathbf{WY}|\mathbf{Z}) \rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{W}|\mathbf{Z}) \tag{13}$$
- **(Weak Union)**
$$(\mathbf{X} \perp\!\!\!\perp \mathbf{WY}|\mathbf{Z}) \rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{ZW}) \tag{14}$$
- **(Contraction)**
$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{WZ}) \wedge (\mathbf{X} \perp\!\!\!\perp \mathbf{W}|\mathbf{Z}) \rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{YW}|\mathbf{Z}) \tag{15}$$

For strictly positive distribution in addition to the above the following axiom also holds:

- **(Intersection)**
$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{WZ}) \wedge (\mathbf{X} \perp\!\!\!\perp \mathbf{W}|\mathbf{YZ}) \rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{YW}|\mathbf{Z}) \tag{16}$$

Markov Blanket. We briefly review the notion of Markov blanket, which used in Sec 3.2.

Definition 8.1. *[37] Fix a joint probability distribution* $\Pr(\mathbf{v})$ *and a variable* $X \in \mathbf{V}$. *A set of variables* $\mathbf{B}(X) \subseteq \mathbf{V} - \{X\}$ *is called a* Markov Blanket *of* X *if* $(X \perp\!\!\!\perp \mathbf{V} - \mathbf{B}(X) - \{X\} \,|\mathbf{B}(X))$*; it is called a* Markov Boundary *if it is minimal w.r.t. set inclusion, denoted* $\mathbf{MB}(X)$.

In the admission process in Fig 2.2 $\mathbf{MB}(X) = \{D, H\}$, simply because $O \perp\!\!\!\perp G|H, D$ (since $\{H, D\}$ d-separate O and G).

It is known that if Pr is a strictly positive distribution (i.e., forall $\mathbf{v} \in Dom(\mathbf{V}), P(\mathbf{v}) > 0$), then $\mathbf{MB}(V)$ is unique for all $V \in \mathbf{V}$ and can be learned from data in polynomial time [32]. Strictly positive distributions do not allows for logical functional dependencies between their variables. The requirement can be satisfied in data by removing logical dependencies [44]. Note that under the faithfulness assumption, the Markov boundary of a node X in the causal graph consists of the parents of X, the children of X, and the parents of the children of X [35].

Counterfactual Fairness. Given a set of features $\mathbf{X}$, a protected attribute S, an outcome variable Y, and a set of unobserved background variables $\mathbf{U}$, Kusner et al. [25] defined a predictor $\tilde{Y}$ to be *counterfactually fair* if for any $\mathbf{x} \in Dom(\mathbf{X})$:

$$P(\tilde{Y}_{S\leftarrow 0}(\mathbf{U}) = 1|\mathbf{X} = \mathbf{x}, S = 1) = P(\tilde{Y}_{S\leftarrow 1}(\mathbf{U}) = 1|\mathbf{X} = \mathbf{x}; S = 1) \tag{17}$$

where, $\tilde{Y}_{S\leftarrow s}(\mathbf{U})$ means intervening on the protected attribute in an unspecified configuration of the exogenous factors. The definition meant to capture the requirement that the protected attribute S should not be a cause of $\tilde{Y}$ at individual level. However, it fails on the simple example in Ex 3.2. This is because, $P(O_{G\leftarrow g}(U_O) = 1) = P(U_O = 1)P(Y_{G\leftarrow g}(U_O) = 1|U_O = 1) = \frac{1}{2}$ for $g = \{0, 1\}$. We note that the stricter version of counterfactual fairness in [26] also fails to capture the individual-level unfairness in this example. We report that this observation has been confirmed by the authors of [26]. We defer the full comparison for future work.

Proof of Lemma 5.1. Because the classifier is a deterministic function trained on $\Pr_{B'}$, it follows that $\Pr_T(o, s, i, a) = \Pr_{B'}(o, s, i, a) = \Pr_{B'}(o|a)\Pr_{B'}(i, s, a)$. Hence it is sufficient to show that $D_{KL}(\Pr_T(o|s, i, a) \,||\, \Pr_T(o|a)) = 0$ or $\Pr_T(o|s, i, a) = \Pr_T(o|a)$. We show this in the following steps:

$$\Pr_T(o|s, i, a) = \frac{\Pr_T(o, s, i, a)}{\Pr_T(s, i, a)} \tag{18}$$
$$= \frac{\Pr_{B'}(o, s, i, a)}{\Pr_T(s, i, a)} \tag{19}$$
$$= \frac{\Pr_{B'}(o|a)\Pr_{B'}(s, i, a)}{\Pr_T(s, i, a)} \tag{20}$$

$$\Pr_T(o|a) = \frac{\Pr_T(o, a)}{\Pr_T(a)} \tag{21}$$
$$= \frac{\sum_{s,i} \Pr_{B'}(o, i, s, a)}{\Pr_T(a)} \tag{22}$$
$$= \frac{\sum_{s,i} \Pr_{B'}(o|a)\Pr_{B'}(s, i, a)}{\Pr_T(a)} \tag{23}$$
$$= \frac{\Pr_{B'}(o|a)\Pr_{B'}(a)}{\Pr_T(a)} \tag{24}$$

Hence,

$$D_{KL}(pr_T(o|s, i, a) \,||\, pr_T(o|a)) \tag{25}$$
$$= -\sum \Pr_T(o|i, s, a) \log \frac{\Pr_T(o|a)}{\Pr_T(o|s, i, a)} \tag{26}$$
$$= -\sum \Pr_T(o|i, s, a) \log \frac{\frac{\Pr_{B'}(o|a)\Pr_{B'}(a)}{pr_T(a)}}{\frac{\Pr_{B'}(o|a)\Pr_{B'}(i,s,a)}{\Pr_T(s,i,a)}} \tag{27}$$
$$= -\sum \Pr_T(o|i, s, a) \log \frac{\Pr_{B'}(s, i|a)}{\Pr_T(s, i|a)} \tag{28}$$

Thus, $D_{KL}(\Pr_T(o|s, i, a) \,||\, \Pr_T(o|a)) = 0$ if $\Pr_T(s, i|a)\Pr_{B'}(s, i|a)$, which implies $\Pr_T(o|s, i, a) = \Pr_T(o|a)$ or equivalently that $(Y \perp\!\!\!\perp S, \mathbf{I}|_{\Pr_T}\mathbf{A})$. This completes the proof. □

8.2 Proofs and Supplementary Propositions and graphs

Proof of Theorem 2.1. Recall that a causal G admits the following factorization of the observed distribution:

$$\Pr(\mathbf{v}) = \prod_{V\in\mathbf{V}} \Pr(v|\mathbf{pa}(V)) \tag{29}$$

Now, each atomic intervention $do(X = x)$ modifies the causal DAG G by removing parents of X from G. Therefore, the probability distribution $P(\mathbf{v}|do(\mathbf{X} = \mathbf{x}))$ can be obtained from the observed distribution $P(\mathbf{v})$ by removing all factors $\Pr(x|\mathbf{pa}(X))$, for $X \in \mathbf{X}$, from $P(\mathbf{v})$, i.e.,

$$\Pr(\mathbf{v}|do(\mathbf{X} = \mathbf{x})) = \frac{\Pr(\mathbf{v})}{\prod_{i=0}^{m} \Pr(x_i|\mathbf{pa}(X_i))} \tag{30}$$

The following holds according to the chain rule of probability:

$$\Pr(\mathbf{v}) = \prod_{i=0}^{m} \left(\Pr\left(\mathbf{pa}(X_i)|\bigcup_{j=0}^{i-1}\mathbf{pa}(X_j), \bigcup_{j=0}^{i-1} x_j\right)\right)\left(\Pr\left(x_i|\bigcup_{j=0}^{i}\mathbf{pa}(X_j), \bigcup_{j=0}^{i-1} x_j\right)\right) \Pr(\mathbf{w}|\mathbf{x}, \mathbf{z}) \tag{31}$$

where, $\mathbf{Z} = \bigcup_{X\in\mathbf{X}} \mathbf{Pa}(X)$, $\mathbf{W} = \mathbf{V} - (\mathbf{X} \cup \mathbf{Z})$ and $j \geq 0$. It holds that in a causal DAG G, any node $X \in \mathbf{V}$ is independent of its non-descendant condition on it parents $\mathbf{P}a(X)$ (known as Markov property [37]). This is simply because $\mathbf{P}a(X)$ d-separates X from its non-descendants. Therefore, the following is implied from the assumption that X_i is an non-descendant of X_{i+1}:

$$\Pr(x_i|\bigcup_{j=0}^{i}\mathbf{pa}(X_j), \bigcup_{j=0}^{i-1} x_j) = \Pr(x_i|\mathbf{pa}(X_i)) \text{ for } i = 0, m \tag{32}$$

Hence,

$$\Pr(\mathbf{v}) = \left(\prod_{i=0}^{m} \Pr\left(\mathbf{pa}(X_i)|\bigcup_{j=0}^{i-1}\mathbf{pa}(X_j), \bigcup_{j=0}^{i-1} x_j\right)\right)\left(\prod_{i=0}^{m} \Pr(x_i|\mathbf{P}a(X_i))\right) \Pr(\mathbf{w}|\mathbf{x}, \mathbf{z}) \tag{33}$$

The following implied from Eq. 33 and 30.

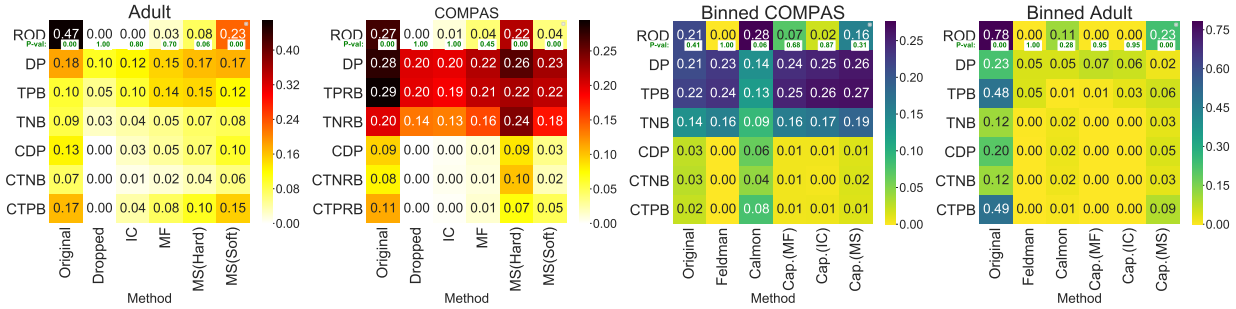
$$\Pr(\mathbf{v}|do(\mathbf{X} = \mathbf{x})) = \frac{\Pr(\mathbf{v})}{\prod_{i=0}^{m} \Pr(x_i|\mathbf{pa}(X_i))} = \left(\prod_{i=0}^{m} \Pr\left(\mathbf{pa}(X_i)|\bigcup_{j=0}^{i-1}\mathbf{pa}(X_j), \bigcup_{j=0}^{i-1} x_j\right)\right)\Pr(\mathbf{w}|\mathbf{x}, \mathbf{z}) \tag{34}$$

Now, by summation over all variables except for Y and $\mathbf{X}$ in Eq. 34 we obtain the following, which proves the theorem.

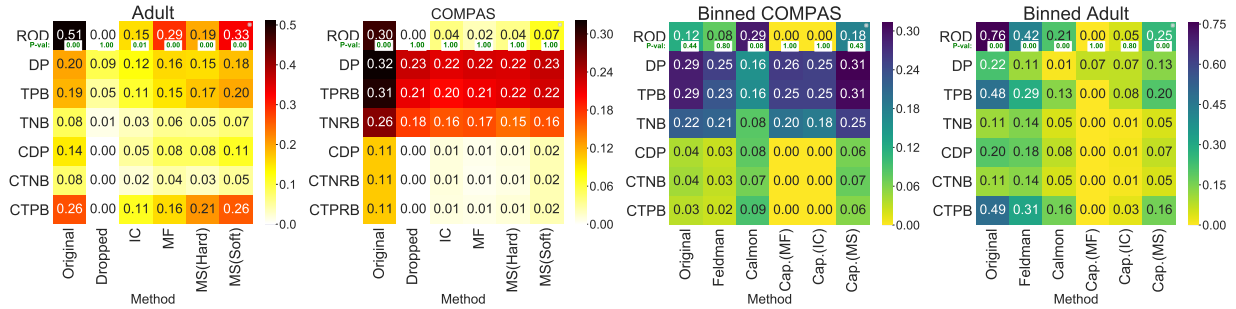
$$P(y|do(\mathbf{X} = \mathbf{x})) = \sum_{\mathbf{z}\in\mathbf{Z}} \Pr(y|\mathbf{x}, \mathbf{z})\left(\prod_{i=0}^{m} \Pr\left(\mathbf{pa}(X_i)|\bigcup_{j=0}^{i-1}\mathbf{pa}(X_j), \bigcup_{j=0}^{i-1} x_j\right)\right) \tag{35}$$

□

Proof of Proposition 3.5. In one direction, we note that, for any choice of $\mathbf{K}$, the causal graph corresponding to an intervention $do(\mathbf{K} = \mathbf{k})$ disconnects S and O, and therefore intervening on S does not affect O. In the other direction, let $\mathbf{P}$ be a path from S to O s.t. $\mathbf{P} \cap \mathbf{A} = \emptyset$, and let $\mathbf{K}$ be the set of all variables not in $\mathbf{P}$; in particular, $\mathbf{A} \subseteq \mathbf{K}$. The causal graph corresponding to an intervention on $\mathbf{K}$ consists of a single path $S \rightarrow^* O$ because all other edges are removed by



a) Fairness Comparison of CAPUCHIN for RF classifier.



b) Fairness Comparison of CAPUCHIN for LR classifier.

Figure 15: Extra Graphs for experiments in Sec 6.

the intervention. Since S has no parents, intervening on S is the same as conditioning on S, and, since Pr is faithful, we have $\Pr(O = o|S = 0) \neq \Pr(O = 0|S = 1)$ for some outcome $O = o$, contradicting the assumption of $\mathbf{K}$-fairness. □

PROOF OF THEOREM 3.7. We show that an algorithm $\mathcal{A}$ is $\mathbf{A}$-fair if $\mathbf{MB}(O) \subseteq \mathbf{A}$. From Theorem 2.1, we obtain:

$$\Pr(O = o|do(S = i), do(\mathbf{A} = \mathbf{a})) = \sum_{\mathbf{z} \in Dom(\mathbf{Z})} \Pr(y|S = i, \mathbf{A} = \mathbf{a}, \mathbf{z}) \left(\prod_{i=0}^{m} \Pr(\mathbf{pa}(X_i) \Big| \bigcup_{j=0}^{i-1} \mathbf{pa}(X_j), \bigcup_{j=0}^{i-1} x_j) \right) \quad (36)$$

where, $\mathbf{Z} = \bigcup_{A \in \mathbf{A}} \mathbf{Pa}(A)$. Without loss of generality assume $\mathbf{Z} \cap \mathbf{A}' = \emptyset$. Let $\mathbf{A} = \mathbf{MB}(O) \cup \mathbf{A}'$ and $\mathbf{V}' = \mathbf{V} - \{\mathbf{A}' \cup \mathbf{Z} \cup \{S\}\}$. From the definition of Markov boundary we have $(O \perp\!\!\!\perp \mathbf{V}', \mathbf{A}', S, \mathbf{Z}|\mathbf{MB}(O))$. It follows from Decomposition and Weak Union axioms in Graphoid that $(O \perp\!\!\!\perp S, \mathbf{Z}|\mathbf{MB}(O), \mathbf{A}')$, hence $(O \perp\!\!\!\perp S, \mathbf{Z}|\mathbf{A})$. We obtain the following for $i = \{0, 1\}$:

$$\begin{aligned} \Pr(O = o|do(S = i), do(\mathbf{A} = \mathbf{a})) &= \Pr(y|A = \mathbf{a}) \\ &\sum_{\mathbf{z} \in Dom(\mathbf{Z})} \left(\prod_{i=0}^{m} \Pr(\mathbf{pa}(X_i) \Big| \bigcup_{j=0}^{i-1} \mathbf{pa}(X_j), \bigcup_{j=0}^{i-1} x_j) \right) \\ &= \Pr(y|A = \mathbf{a}) \end{aligned} \quad (37)$$

Note that (37) obtained by the fact that each product inside the summation becomes 1 (simply because $\sum_X \Pr(X|Y) = 1$) This proves the $\mathbf{A}$-fairness of $\mathcal{A}$. $\mathbf{K}$-fairness for each $\mathbf{K} \supseteq \mathbf{A}$ can be proved in a similar way. □

PROOF OF COROLLARY 3.8. Without loss of generality, suppose $\mathbf{V} = Y\mathbf{DZWU}$ with $\mathbf{X} = \mathbf{A} \cup \mathbf{Z}$ and $\mathbf{A} = \mathbf{W} \cup \mathbf{Z}$. Since the classifier is trained on $\mathbf{X}$, there is a functional dependency $\mathbf{X} \rightarrow O$, which implies $(O \perp\!\!\!\perp \mathbf{Y}, \mathbf{W}, \mathbf{U}|\mathbf{A}, \mathbf{Z})(1)$, i.e., $\mathbf{X}$ forms a Markov blanket for O. It is also implied from the assumptions $\Pr(Y = 1|\mathbf{X} = \mathbf{x}) \approx \Pr(O = 1|\mathbf{X} = \mathbf{x})$ and $(Y \perp\!\!\!\perp \mathbf{X} - \mathbf{A}|\mathbf{A} \cap X)$ that $(O \perp\!\!\!\perp \mathbf{A}|\mathbf{Z})$ approximately holds (2). By applying the Contraction axiom in Graphoid to (1) and (2), we obtain $(O \perp\!\!\!\perp Y\mathbf{A}, \mathbf{W}, \mathbf{U}|\mathbf{Z})$ i.e., $\mathbf{MB}(O) \subseteq \mathbf{A}$. Therefore, $\mathcal{A}$ is justifiably fair according to Theorem 3.7. This completes the proof of part (a). Part (b) is implied from part(1), definition of Markov boundary and Decomposition axiom in Graphoid. □

PROPOSITION 8.2. *Given a fairness application $(\mathcal{A}, S, \mathbf{A}, \mathbf{I})$, suppose the probability distribution of $\mathcal{A}$ is faithful to the causal DAG. Then, the application is justifiably fair iff $\delta(S; O|\mathbf{MB}(O) \cap \mathbf{A}) = 1$.*

PROOF OF PROPOSITION 8.2. It is easy to see $\delta(S; O|\mathbf{MB}(O) \cap \mathbf{A}) = 1$ iff $S \perp\!\!\!\perp O|\mathbf{MB}(O) \cap \mathbf{A}$. Under the faithfulness assumption, we obtain $\mathbf{MB}(O) \cap \mathbf{A}$ and d-separates S and O. Hence, all directed paths from S to O go thorough $\mathbf{MB}(O) \cap \mathbf{A}$. Therefore, the algorithm is justifiably fair according to Theorem 3.5. The converse is immediate from the natural assumption that O does not have any descendants in the causal DAG; hence, its Markov boundary consists of the algorithm's inputs. □

PROOF OF PROPOSITION 4.1. The proposition follows from three facts, all easily verified. (1) $D \subseteq D^*$, (2) D^* satisfies the

MVD $\mathbf{Z} \twoheadrightarrow \mathbf{X}$, and (3) If two databases D_1, D_2 satisfy the MVD then so does $D_1 \cap D_2$. Indeed, the three facts imply that, for any repair D', the database $D^* \cap D'$ is also a repair and $|\Delta(D, D^* \cap D')| \leq \Delta(D, D')|$, hence, if D' is a minimal repair, then $D' \subseteq D^*$. □