A Causal Approach for Unfair Edge Prioritization and Discrimination Removal

Pavan Ravishankar   pavan.rshankar@gmail.com
Pranshu Malviya   cs19s031@cse.iitm.ac.in
Balaraman Ravindran   ravi@cse.iitm.ac.in
Robert Bosch Center for Data Science and Artificial Intelligence, IIT Madras, Chennai, India
Department of Computer Science and Engineering, IIT Madras, Chennai, India

Abstract

In budget-constrained settings aimed at mitigating unfairness, like law enforcement, it is essential to prioritize the sources of unfairness before taking measures to mitigate them in the real world. Unlike previous works, which only serve as a caution against possible discrimination and de-bias data after data generation, this work provides a toolkit to mitigate unfairness during data generation, given by the *Unfair Edge Prioritization* algorithm, in addition to de-biasing data after generation, given by the *Discrimination Removal* algorithm. We assume that a non-parametric Markovian causal model representative of the data generation procedure is given. The edges emanating from the sensitive nodes in the causal graph, such as race, are assumed to be the sources of unfairness. We first quantify *Edge Flow* in any edge $X \to Y$, which is the belief of observing a specific value of Y due to the influence of a specific value of X along $X \to Y$. We then quantify *Edge Unfairness* by formulating a non-parametric model in terms of edge flows. We then *prove* that cumulative unfairness towards sensitive groups in a decision, like race in a bail decision, is non-existent when edge unfairness is absent. We prove this result for the non-trivial non-parametric model setting when the cumulative unfairness cannot be expressed in terms of edge unfairness. We then measure the *Potential to mitigate the Cumulative Unfairness* when edge unfairness is decreased. Based on these measurements, we propose the *Unfair Edge Prioritization* algorithm that can then be used by policymakers. We also propose the *Discrimination Removal Procedure* that de-biases a data distribution by eliminating optimization constraints that grow exponentially in the number of sensitive attributes and values taken by them. Extensive experiments validate the theorem and specifications used for quantifying the above measures.

Keywords: Causal Inference, Fairness, and Public Policy

1. INTRODUCTION

Motivation and Problem: Anti-discrimination laws in the U.S. prohibit unfair treatment of people based on sensitive features, such as gender or race (Act, 1964). The fairness of a decision process is based on *disparate treatment* and *disparate impact*. Disparate treatment, referred to as intentional discrimination, is when sensitive information is explicitly used to make decisions. Disparate impact, referred to as unintentional discrimination, is when decisions hurt certain sensitive groups even when the policies are neutral. For instance, only candidates with a height of 6 feet and above are selected for basketball teams. This might
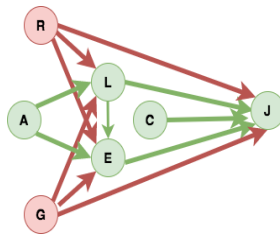
Figure 1: Bail Decision Causal Graph. Each node is generated from its parents. **Edges:** Fair Edges in Green, Unfair Edges in Red; **Nodes:** Sensitive nodes in Red, Not sensitive nodes in Green; R: Race; A: Age, G: Gender, L: Literacy, E: Employment, C: Case characteristics, J: Judicial Bail decision

eliminate players of a certain race. Unjustifiable disparate impact is unlawful (Barocas and Selbst, 2016). In high-risk decisions, such as in the criminal justice system, it is imperative to mitigate unfairness resulting from either disparate treatment or disparate impact. Considering that agencies operate in a budget-constrained scenario owing to limited resources, it is essential to prioritize potential sources of unfairness before we take measures to mitigate them. This paper proposes the Unfair Edge Prioritization methodology for prioritizing these sources before mitigating unfairness during the data generation phase. Further, this paper also proposes the Discrimination Removal procedure to de-bias data distribution after data is generated.

We motivate our problem through the following illustration using Fig. 1. Consider the problem of reducing unfairness in the *bail decision J* towards a specific *racial* group R. We use an unfair edge as the potential source of unfairness as in Chiappa and Isaac (2018). An unfair path contains at least one unfair edge. The unfairness propagates along all the unfair paths from the racial group R to the bail decision J. Although discrimination has been quantified in previous works (Zhang et al., 2017), it only serves as a caution against possible discrimination. There is utility when such notes of caution, like "discrimination exists in the bail decision J towards the racial group R", are augmented with tangible information to mitigate discrimination, like "unfairness in the unfair edge $R \to L$ is responsible for discrimination in the bail decision J towards the racial group R". Then, the agencies can attempt to address the real-world issues underlying $R \to L$, such as lack of scholarships for racial group R. The challenge lies in providing such tangible information and methodologies for mitigating discrimination such as the amount of unfairness present in an edge, the measure of how edge unfairness affects cumulative unfairness [1], prioritizing the unfair edges, and removing discrimination.

This paper attempts to provide such "tangible" information using Pearl's framework of Causal Inference (Pearl, 2009). The contributions of this paper are as follows,

---

1. Cumulative unfairness captures discrimination due to unequal influences of $R = r$ on $J = j$ as compared to $R = r'$ on $J = j$ via the directed paths from R to J.

1. **Quantify Edge Flow** in any edge $X \to Y$, which is the belief of observing a specific value of Y due to the influence of a specific value of X along the edge $X \to Y$.
2. **Quantify Edge Unfairness** in any edge $X \to Y$, which is the average difference in conditional probability of Y given its parents $Pa(Y)$, $\mathbb{P}(Y|Pa(Y))$, with and without edge flow along $X \to Y$. It measures average unit contribution of edge flow in $X \to Y$ to $\mathbb{P}(Y|Pa(Y))$. We formulate a non-parametric model for $\mathbb{P}(Y|Pa(Y))$ in terms of edge flows along the parental edges of Y (see Theorem 10).
3. **Prove** that the discrimination in any decision towards any sensitive groups is non-existent when edge unfairness is eliminated. The proof is non-trivial in the non-parametric model setting of CPTs as cumulative unfairness cannot be expressed in terms of edge unfairness. We derive this result by upper bounding the absolute value of cumulative unfairness and showing that the upper bound becomes zero when edge unfairness is zero (see Theorem 12 and Corollary 13).
4. **Quantify the Potential to Mitigate Cumulative Unfairness** by calculating the derivative of the upper bound w.r.t edge unfairness. We do this as cumulative unfairness cannot be expressed in terms of edge unfairness in a non-parametric model setting of *CPTs*.
5. **Propose an Unfair Edge Prioritization** algorithm to prioritize unfair edges based on their potential to mitigate cumulative unfairness and edge unfairness. Using these priorities, agencies can address the real-world issues underlying the unfair edge with the top priority.
6. **Propose a Discrimination Removal algorithm** to de-bias data distribution by eliminating exponentially growing constraints and subjectively chosen threshold of discrimination.

Contents: We discuss the preliminaries in Section 2; quantify edge flow, edge unfairness, its impact on cumulative unfairness, and prove that discrimination is absent when edge unfairness is eliminated in Section 3; propose unfair edge prioritization and discrimination removal algorithm in Section 4; discuss experiments in Section 6; discuss related work in Section 3; discuss conclusion in Section 7.

2. PRELIMINARIES

Throughout the paper, we use boldfaced capital letters $\mathbf{X}$ to denote a set of nodes; italicized capital letter X to denote a single node; boldfaced small letters $\mathbf{x}$ to denote a specific value taken by $\mathbf{X}$; italicized small letter x to denote a specific value taken by its corresponding node X. $\mathbf{x_A}$ restricts the values of $\mathbf{x}$ to the node-set $\mathbf{A}$. $Pa(X)$ denotes parents of X and $pa(X)$ the specific values taken by them. Each node is associated with the conditional probability table $\mathbb{P}(X|Pa(X))$.

Assumptions: We assume the following are given: (1) A *Markovian causal model M* consisting of a causal graph $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ with variables V and edges E, (2) Conditional probability tables (CPTs) $\mathbb{P}(V|Pa(V))$ where $\forall V \in \mathbf{V}$ and $Pa(V)$ are the parents of V, (3) A list of *sensitive nodes*, whose emanating edges are potential sources of unfairness, (4) The variables in the causal graph are *discrete and observed*. Even though the theorem

results extend to the continuous variable setting, Assumption (4) is made to not digress into inference and identifiability challenges (Avin et al., 2005). This paper does not make any assumptions about the deterministic functions in the causal model.

Definition 1 *Node interventional distribution* *denoted by* $\mathbb{P}(\mathbf{Y}|do(\mathbf{X} = \mathbf{x}))$ *is the distribution of* $\mathbf{Y}$ *after forcibly setting* $\mathbf{X}$ *to* $\mathbf{x}$ *irrespective of the values taken by the parents of* X *(Pearl, 2009).*

Definition 2 *A* *causal model* *is formally defined as a triple* $M =< \mathbf{U}, \mathbf{V}, \mathbb{F} >$ *where,*

1. $\mathbf{U}$ *is a set of unobserved random variables also known as exogenous variables that are determined by factors outside the model. A joint probability distribution* $\mathbb{P}(\mathbf{u})$ *is defined over the variables in* $\mathbf{U}$.
2. $\mathbf{V}$ *is a set of observed random variables also known as endogenous that are determined by variables in the model, namely, variables in* $\mathbf{U} \cup \mathbf{V}$.
3. $\mathbb{F}$ *is a set of deterministic functions* $\{f_1, ..., f_i, ...\}$ *where each* f_i *is a mapping from* $\mathbf{U} \times (\mathbf{V} \backslash X_i)$ *to* X_i *written as,*

$$x_i = f_i(pa(X_i), \mathbf{u}_i) \tag{1}$$

where $X_i \in \mathbf{V}$, $pa(X_i)$ *are the specific values taken by the observed set of parents of* X_i *and* $\mathbf{u}_i$ *are the specific values taken by unobserved set of parents of* X_i.

Each causal model is associated with a causal graph $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ *where* $\mathbf{V}$ *are the observed nodes and* $\mathbf{E}$ *are the directed edges. We assume that the causal model is Markovian which means that all exogenous variables* $\mathbf{U}$ *are mutually independent and each node is independent of its non-descendants conditional on all its parents. For a markovian model joint distribution* $\mathbb{P}(\mathbf{V})$ *is given by,*

$$\mathbb{P}(\mathbf{V}) = \prod_{V \in \mathbf{V}} \mathbb{P}(v|pa(V)) \tag{2}$$

where $\mathbb{P}(v|pa(V))$ *is the conditional probability table CPT associated with* V.

Definition 3 *Identifiability:* *Let* $\mathbb{G}$ *be a causal graph. A node interventional distribution* $\mathbb{P}(\mathbf{Y}|do(\mathbf{X} = \mathbf{x}))$, *i.e., probability of* $\mathbf{Y}$ *when* $\mathbf{X}$ *is forcibly set to* $\mathbf{x}$ *is said to be identifiable if it can be expressed using the observational probability* $\mathbb{P}(\mathbf{V})$. *When* $\mathbb{G}$ *comprises of only observed variables as in our work,*

$$\mathbb{P}(\mathbf{y}|do(\mathbf{x})) = \sum_{\mathbf{v} \in \mathbf{V} \backslash \{\mathbf{X}, \mathbf{Y}\}} \prod_{V \in \mathbf{V} \backslash \{\mathbf{X}, \mathbf{Y}\}, \mathbf{Y} = \mathbf{y}} \mathbb{P}(v|pa(V))|_{\mathbf{x}} \tag{3}$$

Definition 4 *Total Causal Effect* $TE_{\mathbf{y}}(\mathbf{x}_2, \mathbf{x}_1)$ *measures causal effect of variables* $\mathbf{X}$ *on decision variables* $\mathbf{Y} = \mathbf{y}$ *when it is changed from* $\mathbf{x}_1$ *to* $\mathbf{x}_2$ *written as,*

$$TE_{\mathbf{y}}(\mathbf{x}_2, \mathbf{x}_1) = \mathbb{P}(\mathbf{y}|do(\mathbf{x}_2)) - \mathbb{P}(\mathbf{y}|do(\mathbf{x}_1)) \tag{4}$$

Definition 5 *Path-specific effect* $SE_{\pi, \mathbf{y}}(\mathbf{x}_2, \mathbf{x}_1)$ *measures effect of node* $\mathbf{X}$ *on decision* $\mathbf{Y} = \mathbf{y}$ *when it is changed from* $\mathbf{x}_1$ *to* $\mathbf{x}_2$ *along the directed paths* $\boldsymbol{\pi}$, *while retaining* $\mathbf{x}_1$ *for the directed paths not in* $\boldsymbol{\pi}$ *i.e.* $\tilde{\boldsymbol{\pi}}$ *written as,*

$$SE_{\pi, \mathbf{Y} = \mathbf{y}}(\mathbf{x}_2, \mathbf{x}_1) = \mathbb{P}(\mathbf{y}|do(\mathbf{x}_2|_{\pi}, \mathbf{x}_1|_{\tilde{\boldsymbol{\pi}}})) - \mathbb{P}(\mathbf{y}|do(\mathbf{x}_1)) \tag{5}$$

Definition 6 *A trail $V_1 \rightleftharpoons ..... \rightleftharpoons V_n$ is said to be an **active trail** given a set of nodes $\mathbf{X}$ in $\mathbb{G}$ if for every v-structure $V_i \rightarrow V_j \leftarrow V_k$ along the trail, V_j or any descendent of V_j is in $\mathbf{X}$ and no other node in the trail belongs to $\mathbf{X}$.*

Definition 7 $\mathbf{A}$ *is said to be **d-separated** from $\mathbf{B}$ given $\mathbf{C}$ in a graph $\mathbb{G}$ ($d\text{-}sep_{\mathbb{G}}(\mathbf{A};\mathbf{B}|\mathbf{C})$) if there is no active trail from any $A \in \mathbf{A}$ to any $B \in \mathbf{B}$ given $\mathbf{C}$ as discussed in* Pearl (2009), *and* Koller and Friedman (2009) . *If there is atleast one active trail from any $A \in \mathbf{A}$ to any $B \in \mathbf{B}$ given $\mathbf{C}$, then $\mathbf{A}$ is said to be **d-connected** from $\mathbf{B}$ given $\mathbf{C}$ in a graph $\mathbb{G}$ ($d\text{-}conn_{\mathbb{G}}(\mathbf{A};\mathbf{B}|\mathbf{C})$) as shown in Fig. 1.3 in* Pearl (2009).

Theorem 8 *If sets $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$ in a DAG $\mathbb{G}(\mathbf{E}, \mathbf{V})$, then $\mathbf{X}$ is independent of $\mathbf{Y}$ conditional on $\mathbf{Z}$ in every distribution $\mathbb{P}$ that factorizes over $\mathbb{G}$. Conversely, if $\mathbf{X}$ and $\mathbf{Y}$ are d-connected by $\mathbf{Z}$ in a DAG $\mathbb{G}$, then $\mathbf{X}$ and $\mathbf{Y}$ are dependent conditional on $\mathbf{Z}$ in at least one distribution $\mathbb{P}$ that factorizes over $\mathbb{G}$ as shown in Theorem 1.2.4 in* Pearl (2009).

Unfair edge $S \rightarrow X$**:** Unfair edge is a directed edge $S \rightarrow X$ with S being a sensitive node like race. Set of unfair edges in $\mathbb{G}$ is denoted by $\mathbf{E}_{\mathbb{G}}^{\text{unfair}}$. Unfair edge is a potential source of unfairness. For instance, in Fig. 1, $G \rightarrow E$ is unfair if the accused is denied admission to co-ed institutions based on gender. On the other hand, $G \rightarrow E$ is fair, if only gender-specific institutions existed in the locality as discussed in Chiappa and Isaac (2018). Hence, the usage of the term *potential*.

Unfair paths $\boldsymbol{\pi}_{\mathbf{S},Y,\mathbb{G}}^{\text{unfair}}$**:** Unfair paths are the set of directed paths from sensitive node $S \in \mathbf{S}$ to the decision node Y in graph $\mathbb{G}$. Unfair paths capture how unfairness propagates from the sensitive nodes onto a destination node. For instance, in Fig. 1, $\boldsymbol{\pi}_{G,J,\mathbb{G}}^{\text{unfair}}$ consists of $G \rightarrow E \rightarrow J$ that captures how unfairness in the edge $G \rightarrow E$ propagates to J. Non-causal paths do not propagate unfairness from sensitive nodes. Suppose there is another node, say religious belief R, and another non-causal path, say $R \leftarrow E \rightarrow J$. Still, $R \leftarrow E \rightarrow J$ is fair because bail decision J is taken based on employment E and not on religious belief R as discussed in Chiappa and Isaac (2018).

3. Edge Unfairness

In this section, we quantify *edge flow*, *edge unfairness*, prove that eliminating edge unfairness eliminates cumulative unfairness, and quantify the *potential to mitigate cumulative unfairness* when edge unfairness is reduced.

Edge flow along any edge, say $R \rightarrow J$ is the belief of observing a specific value of bail decision J due to the influence of a specific value of race R along $R \rightarrow J$. This can be extended to multiple direct edges from $\mathbf{M} = \{R, G\}$ to X.

Definition 9 *Edge flow $\mathbb{P}_{flow}^{\mathbf{M}=\mathbf{m}}(X = x)$ is defined as,*

$$\mathbb{P}_{flow}^{\mathbf{m}}(x) = \frac{e^{\mathbb{E}_{\mathbf{m}' \sim \mathbb{P}(\mathbf{m}')} SE_{\boldsymbol{\pi},x}(\mathbf{m},\mathbf{m}')}}{\sum\limits_{x} e^{\mathbb{E}_{\mathbf{m}' \sim \mathbb{P}(\mathbf{m}')} SE_{\boldsymbol{\pi},x}(\mathbf{m},\mathbf{m}')}}, \quad where \; \boldsymbol{\pi} = \{M \rightarrow X | M \in \mathbf{M}\} \tag{6}$$

Edge flow is formalized using direct effect $SE_{\boldsymbol{\pi},x}(\mathbf{m}, \mathbf{m}')$ (Avin et al., 2005). $SE_{\boldsymbol{\pi},x}(\mathbf{m}, \mathbf{m}')$ is the effect of $\mathbf{M} = \mathbf{m}$ on $X = x$ along the direct edges from $\mathbf{M}$ to X irrespective of the value $\mathbf{m}'$ set along the indirect paths ensured by averaging. $SE_{\boldsymbol{\pi},x}(\mathbf{m}, \mathbf{m}')$ is identifiable because there is no recanting witness (see Definition 3, Theorem 1, and Theorem 2 in Zhang et al. (2017)). A positive scaling like softmax ensures that the edge flow is a positive quantity which we use to prove Theorem 12.

To quantify edge unfairness along $R \to J$, we first decompose $\mathbb{P}(J|Pa(J))$ into edge flows along the direct edges $\{X \to J | X \in Pa(J)\}$. The rationale is that the active trails $\{X \to J | X \in Pa(J)\}$ (Definition 6) resulting from the dependencies $\{(J \not\perp X|Pa(J)\backslash X)|X \in Pa(J)\}$, which influence J's value in $\mathbb{P}(J|Pa(J))$ (Theorem 8), are same as the edges along which the edge flows from $Pa(J)$ to J propagate. The theorem below formalizes this concept.

Theorem 10 *The conditional probability distribution $\mathbb{P}(X = x|Pa(X) = pa(X))$ is a function $f^{\mathbf{w}}$ of FairFlow$_{x,pa(X)}$ and UnfairFlow$_{x,pa(X)}$ given by,*

$$\mathbb{P}(x|pa(X)) = f^{\mathbf{w}}(\text{FairFlow}_{x,pa(X)}, \text{UnfairFlow}_{x,pa(X)}) \tag{7}$$

$$\text{where,} \quad f^{\mathbf{w}} : \mathbb{W}^{|\mathbf{U}_X|+|\mathbf{F}_X|+1} \to [0,1], \tag{8}$$

$$\text{subject to,} \sum_x f^{\mathbf{w}}(\text{FairFlow}_{x,pa(X)}, \text{UnfairFlow}_{x,pa(X)}) = 1, \tag{9}$$

$$f^{\mathbf{w}}(\text{FairFlow}_{x,pa(X)}, \text{UnfairFlow}_{x,pa(X)}) \geq 0 \tag{10}$$

where $\mathbf{w}$ are the weights, $\mathbf{U}_X$ are the parents of X along an unfair edge, $\mathbf{F}_X$ are the parents of X along a fair edge, FairFlow$_{x,pa(X)} = \mathbb{P}_{flow}^{\mathbf{F}_X=pa(X)\mathbf{F}_X}(X = x)$ and UnfairFlow$_{x,pa(X)} = \{\mathbb{P}_{flow}^{A=pa(X)_A}(X = x); \ A \in \mathbf{U}_X\}$.

Proof: FairFlow$_{x,pa(X)} = \mathbb{P}_{flow}^{pa(X)\mathbf{F}_X}(x)$ measures the effect of parents of X along the fair edges and UnfairFlow$_{x,pa(X)} = \{\mathbb{P}_{flow}^{A=pa(X)_A}(X = x); \ A \in \mathbf{U}_X\}$ measures the effect of parents of X along the unfair edges [Definition 9]. Thus, $\bigcup_{A\in\mathbf{U}_X} \mathbb{P}_{flow}^{pa(X)_A}(x)$ and $\mathbb{P}_{flow}^{pa(X)\mathbf{F}_X}(x)$ measure the effects along the direct edges $\{M \to X | M \in Pa(X)\}$. Further, the set of active trails resulting from the dependencies $\{(X \not\perp M|Pa(X)\backslash M)|M \in Pa(X)\}$, which influence X's value in $\mathbb{P}(x|pa(X))$, is also $\{M \to X | M \in Pa(X)\}$ [Theorem 8]. Hence, $\mathbb{P}(x|pa(X))$ can be formulated as a function of $\bigcup_{A\in\mathbf{U}_X} \mathbb{P}_{flow}^{pa(X)_A}(x)$ and $\mathbb{P}_{flow}^{pa(X)\mathbf{F}_X}(x)$ provided the function satisfies the axioms of probability ∎.

This theorem aids in the formulation of edge unfairness in an unfair edge, say in $R \to J$ as the difference in $\mathbb{P}(J|Pa(J))$ with and without the edge flow in $R \to J$. *Edge unfairness is formalized below.*

Definition 11 *Edge unfairness μ_e of an unfair edge $e = K \to X$ is,*

$$\mu_e = \mathbb{E}_{(x,pa(X))\sim\mathbb{P}(x,pa(X))} \left[\frac{D_{x,pa(X)}^K}{\mathbb{P}_{flow}^{pa(X)_K}(x)} \right] \tag{11}$$
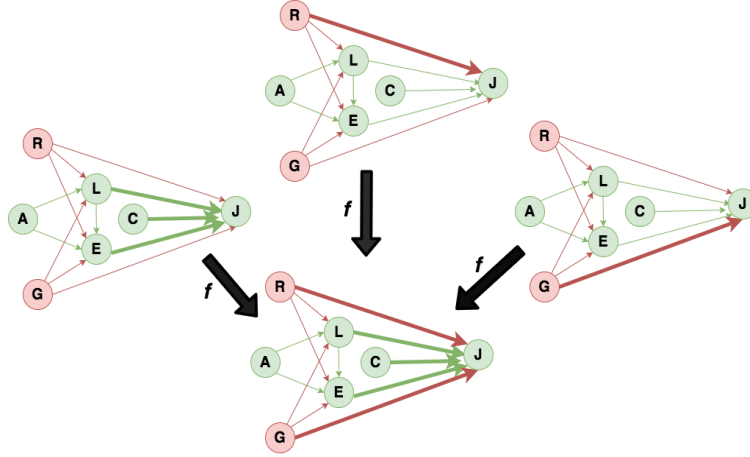
Figure 2: Edge flows $\mathbb{P}_{\text{flow}}^{\{L,C,E\}}(J)$, $\mathbb{P}_{\text{flow}}^{R}(J)$ and $\mathbb{P}_{\text{flow}}^{G}(J)$ interact via function f to generate $\mathbb{P}(J|Pa(J))$. $\mathbb{P}_{\text{flow}}^{\{L,C,E\}}(J)$ result in effects along the fair edges $\{L \rightarrow J, E \rightarrow J, C \rightarrow J\}$ as shown in middle-left. $\mathbb{P}_{\text{flow}}^{R}(J)$ result in the effect along the unfair edge $\{R \rightarrow J\}$ as shown in top-center. $\mathbb{P}_{\text{flow}}^{G}(J)$ result in the effect along the unfair edge $\{G \rightarrow J\}$ as shown in the middle-right. The active trails resulting as a consequence of the dependencies in $\mathbb{P}(J|Pa(J))$ are $\{R \rightarrow J, L \rightarrow J, E \rightarrow J, C \rightarrow J, G \rightarrow J\}$ as shown in bottom.

$$where, \quad D_{x,pa(X)}^{K} = \left| CPT_{x,pa(X)}^{K} - \widetilde{CPT}_{x,pa(X)}^{K} \right| \tag{12}$$

$$CPT_{x,pa(X)}^{K} = f_X\left(FairFlow_{x,pa(X)}, \bigcup_{A \in \mathbf{U}_X} \mathbb{P}_{flow}^{pa(X)_A}(x)\right) \tag{13}$$

$$\widetilde{CPT}_{x,pa(X)}^{K} = f_X\left(FairFlow_{x,pa(X)}, \bigcup_{A \in \mathbf{U}_X \backslash K} \mathbb{P}_{flow}^{pa(X)_A}(x), \mathbb{P}_{flow}^{pa(X)_K}(x) = 0\right) \tag{14}$$

Edge unfairness μ_e is the unit contribution of edge flow $\mathbb{P}_{\text{flow}}^{K}(X)$ to $\mathbb{P}(X|Pa(X))$. μ_e measures the difference in $\mathbb{P}(X|Pa(X))$ with the edge flow along e, given by $\text{CPT}_{x,pa(X)}^{K}$, and without the edge flow along e, given by $\widetilde{\text{CPT}}_{x,pa(X)}^{K}$. We measure per unit edge flow to capture that a large $D_{X,K}$ compared to $\mathbb{P}_{\text{flow}}^{K}(X)$ still results in large μ_e even though $D_{X,Pa(X)}^{K}$ is small.

Now, we quantify cumulative unfairness. The objective for introducing cumulative (overall) unfairness is twofold: (1)To prove that eliminating edge (local) unfairness eliminates cumulative unfairness (2)To formulate the potential to mitigate cumulative unfairness. We combine direct and indirect discrimination, discussed in Section 3 of Zhang et al. (2017), to define cumulative unfairness towards sensitive nodes $\mathbf{S} = \mathbf{s}$ in decision $Y = y$. Cumulative unfairness $C_{\mathbf{S}=\mathbf{s},Y=y}$ is,

$$C_{\mathbf{S}=\mathbf{s},Y=y} = \mathbb{E}_{\mathbf{s}' \sim \mathbb{P}(\mathbf{s}')} TE_{Y=y}(\mathbf{s}, \mathbf{s}') \tag{15}$$

$C_{\mathbf{S}=\mathbf{s},Y=y}$ measures the impact on outcome $Y = y$ when $\mathbf{S}$ is forcibly set to $\mathbf{s}$ along the unfair paths from $\mathbf{S}$ to Y irrespective of the value set along other paths. Since all edges

emanating from a sensitive node are potential sources of unfairness, the total causal effect TE (see Definition 4) is used to formulate cumulative unfairness. Proving the result that eliminating edge unfairness μ_e in all unfair edges e eliminates cumulative unfairness $C_{\mathbf{S}=\mathbf{s},Y=y}$ is not straightforward as $C_{\mathbf{S}=\mathbf{s},Y=y}$ cannot be expressed in terms μ_e. We first upper bound $|C_{\mathbf{S}=\mathbf{s},Y=y}|$ by $C^{\text{upper}}_{\mathbf{S}=\mathbf{s},Y=y}$ that can be expressed in terms of μ_e. Then, the result follows from the theorem.

Theorem 12 *The magnitude of cumulative unfairness in decision $Y = y$ towards sensitive nodes $S = s$, $C_{\mathbf{S}=\mathbf{s},Y=y}$, is upper bounded by $C^{upper}_{\mathbf{S}=\mathbf{s},Y=y}$ as shown below,*

$$|C_{\mathbf{S}=\mathbf{s},Y=y}| \leq C^{upper}_{\mathbf{S}=\mathbf{s},Y=y} \tag{16}$$

where,

$$C^{upper}_{\mathbf{S}=\mathbf{s},Y=y} = \sum_{\mathbf{s}'\in\mathbf{S}\backslash\mathbf{s}} \mathbb{P}(\mathbf{s}') \left[\sum_{v\in\mathbf{V}\backslash\{\mathbf{S},Y\}} \prod_{V\in\mathbf{V}\backslash\{\mathbf{S},Y\},Y=y} \sum_{A\in\mathbf{U}_V} \left[\frac{\mathbb{P}^{\mathbf{s}_A\vee pa(V)_A}_{flow}(v)}{\mathbb{P}(v,pa(V))|_{\mathbf{s}}} + \frac{\mathbb{P}^{\mathbf{s}'_A\vee pa(V)_A}_{flow}(v)}{\mathbb{P}(v,pa(V))|_{\mathbf{s}'}} \right] \mu_{A\to V} \right] \tag{17}$$

$$\mathbf{s}_A \vee pa(V)_A \equiv A = \begin{cases} \mathbf{s}_A, & \text{if } A \in \mathbf{S} \\ pa(V)_A & \text{otherwise} \end{cases} \tag{18}$$

Proof Sketch of Theorem:
Cumulative unfairness $C_{\mathbf{S}=\mathbf{s},Y=y}$ cannot be expressed in terms of edge unfairness when the conditional probability is modeled by a non-parametric model f. Therefore, we write $C_{\mathbf{S}=\mathbf{s},Y=y}$ in terms of conditional probabilities CPTs. Each CPT is substituted by its functional model f (see Theorem 10), because the edge unfairness is expressed in terms of f (see Definition 12). To bring edge unfairness μ_e into the formulation, we upper bound each f of a node present along an unfair edge e with the following quantities: edge unfairness μ_e and f having no edge flow along e. The rationale of this step comes from the definition of edge unfairness μ_e (see Definition 12) and the fact that the modulus operation is a non-negative quantity. This proves the result $C_{\mathbf{S}=\mathbf{s},Y=y} \leq C^{\text{upper}}_{\mathbf{S}=\mathbf{s},Y=y}$. By following similar steps and using modulus operation to lower bound f, we arrive at $C_{\mathbf{S}=\mathbf{s},Y=y} \geq -C^{\text{upper}}_{\mathbf{S}=\mathbf{s},Y=y}$. *(see Supplementary for full proof)* ∎

Corollary 13 *The cumulative unfairness in decision $Y = y$ towards sensitive nodes $S = s$, $C_{\mathbf{S}=\mathbf{s},Y=y}$, is non-existent when edge unfairness μ_e in all unfair edges is eliminated.*

We now measure the potential to mitigate cumulative unfairness when edge unfairness is reduced. Using the *potential* measure and *edge unfairness*, agencies can then prioritize the unfair edges before taking measures to mitigate them.

Sensitivity measures the variation in $C^{\text{upper}}_{\mathbf{S}=\mathbf{s},Y=y}$ when edge unfairness μ_e in unfair edge e is varied. Since $C^{\text{upper}}_{\mathbf{S}=\mathbf{s},Y=y}$ is a linear function in edge unfairness, higher order derivatives (≥ 2) of $C^{\text{upper}}_{\mathbf{S}=\mathbf{s},Y=y}$ with respect to μ_e are 0.

Definition 14 *Sensitivity* of $C_{\mathbf{S}=\mathbf{s},Y=y}^{upper}$ *w.r.t edge unfairness in edge* e μ_e *is,*

$$S_e^{\mathbf{S}=\mathbf{s},Y=y} = \left.\frac{\partial C_{\mathbf{S}=\mathbf{s},Y=y}^{upper}}{\partial \mu_e}\right|_{\boldsymbol{\mu}^*} \tag{19}$$

where, $\boldsymbol{\mu}^*$ *are the current edge unfairness obtained from observational distribution* $\mathbb{P}(\mathbf{V})$.

The following quantity measures the potential to mitigate $C_{\mathbf{S}=\mathbf{s},Y=y}^{\text{upper}}$ when edge unfairness μ_e in unfair edge e is decreased. $C_{\mathbf{S}=\mathbf{s},Y=y}^{\text{upper}}$ is used to measure the potential contrary to $C_{\mathbf{S}=\mathbf{s},Y=y}$ because $C_{\mathbf{S}=\mathbf{s},Y=y}$ cannot be expressed in terms of edge unfairness. Experiment 5.3 validates that decreasing $C_{\mathbf{S}=\mathbf{s},Y=y}^{\text{upper}}$ decreases $C_{\mathbf{S}=\mathbf{s},Y=y}$.

Definition 15 *Potential to Mitigate Cumulative Unfairness* *when* μ_e *is decreased.*

$$P_e^{\mathbf{S}=\mathbf{s},Y=y} = \begin{cases} -\left|S_e^{\mathbf{S}=\mathbf{s},Y=y}\right| & \text{if } C_{\mathbf{S}=\mathbf{s},Y=y}^{upper} = 0 \\[2ex] S_e^{\mathbf{S}=\mathbf{s},Y=y} & \text{if } C_{\mathbf{S}=\mathbf{s},Y=y}^{upper} > 0 \end{cases} \tag{20}$$

$P_e^{\mathbf{S}=\mathbf{s},Y=y}$ states that if $C_{\mathbf{S}=\mathbf{s},Y=y}^{\text{upper}} = 0$, then $C_{\mathbf{S}=\mathbf{s},Y=y}^{\text{upper}}$ deviates from 0 (indicative of no-discrimination) as edge unfairness is decreased. The potential of $C_{\mathbf{S}=\mathbf{s},Y=y}^{\text{upper}}$ to move towards 0 or to get mitigated is then quantified by $-|S_e^{\mathbf{S}=\mathbf{s},Y=y}|$ wherein negative is due to $C_{\mathbf{S}=\mathbf{s},Y=y}^{\text{upper}}$ deviating from 0. Similarly, one can analyze the other case.

4. Unfair Edge Prioritization & Discrimination Removal

Based on the theorems and the definitions, we present pseudo-codes for fitting the $CPTs$ in Algorithm 1, computing priority of the unfair edges in Algorithm 2, and removing discrimination in Algorithm 3. Algorithm 2 aids the agencies to mitigate unfairness underlying the unfair edges in the real-world during the data generation phase. Algorithm 3 de-biases data distribution after the data generation phase. Algorithm 2 calls Algorithm 1. Algorithm 1 does not call Algorithm 2 and Algorithm 3. Algorithm 3 does not call Algorithm 2 and Algorithm 1.

(1)**fitCPT() Algorithm 1**: It takes the causal model $(\mathbb{G}, \mathbb{P})$, the set of unfair edges $\mathbf{E}_{\mathbb{G}}^{\text{unfair}}$, and the attribute X as inputs and approximates the $CPTs$ $\mathbb{P}(X|Pa(X))$ by the model $f^{\mathbf{w}}$ using the least-squares loss.

Algorithm 1 fitCPT($\mathbb{G}$,$\mathbb{P}$,$\mathbf{E}_{\mathbb{G}}^{\text{unfair}}$,X)

1: Initialize $\mathbf{w}$ randomly
2: $\mathbf{Y} \leftarrow \mathbb{P}(X = x | Pa(X) = pa(X))$
3: Compute $\text{FairFlow}_{x,pa(X)}$ and $\text{UnfairFlow}_{x,pa(X)}$ (Theorem 10)
4: $\hat{\mathbf{Y}}_X(\mathbf{w}) \leftarrow f^{\mathbf{w}}(\text{FairFlow}_{x,pa(X)}, \text{UnfairFlow}_{x,pa(X)})$ (Theorem 10)
5: $\mathbf{w}^* \leftarrow \arg\min_{\mathbf{w}} ||\mathbf{Y} - \hat{\mathbf{Y}}_X(\mathbf{w})||^2$ subject to Eq. 9 & Eq. 10
6: **Output: $\mathbf{w}^*$**

---

(2)**computePriority() Algorithm 2**: It computes priorities of the unfair edges based on the edge unfairness and the potential to mitigate the cumulative unfairness. The priorities can be used to address unfairness in the real world.

Algorithm 2 computePriority($\mathbb{G}, \mathbb{P}, \mathbf{E}_{\mathbb{G}}^{\text{unfair}}, \mathbf{s}, y, w_u, w_p$)

1: $\mathbf{w}^* = \{\}$ ▷ Optimal weights of the approximated CPTs
2: for V in $\mathbf{V}$ do
3:     $\mathbf{w}_V^* \leftarrow \text{fitCPT}(\mathbb{G}, \mathbb{P}, \mathbf{E}_{\mathbb{G}}^{\text{unfair}}, V)$ ▷ See Algorithm 1
4:     $\mathbf{w}^* \leftarrow \mathbf{w}^* \cup \{\mathbf{w}_V^*\}$
5: priorityList = $\{\}$
6: for $e = S \rightarrow V$ in $\mathbf{E}_{\mathbb{G}}^{\text{unfair}}$
7:     Compute μ_e using $f^{\mathbf{w}_V^*}$ from Eq. 12
8:     Compute $P_e^{\mathbf{S}=\mathbf{s}, Y=y}$ ▷ See Definition 15
9:     priority = $w_u \mu_e + w_p P_e^{\mathbf{S}=\mathbf{s}, Y=y}$ ▷ w_u and w_p are weights
10:     priorityList = priorityList $\cup \{(e, \text{priority})\}$
11: **Output:** priorityList
---

(3)**removeDiscrimination() Algorithm 3**: It removes discrimination by regenerating new CPTs for the causal model $(\mathbb{G}, \mathbb{P})$ with unfair edges $\mathbf{E}_{\mathbb{G}}^{\text{unfair}}$. These CPTs are approximated by solving an optimization problem of minimizing the overall edge unfairness subject to the axioms of probability as constraints. A *data utility* term, which is the Mean Squared Error (MSE) between $\mathbb{P}(\mathbf{V})$ and the new joint distribution computed from the product of approximated CPTs, is added to the objective function to ensure that the influences from other insensitive nodes are preserved. For instance, a sensitive node like religious belief R can have insensitive nodes like literacy L as a parent. By minimizing only the edge unfairness in the objective function, indirect influences like $L \rightarrow R \rightarrow J$ can get altered, thereby not preserving data utility. Also, this algorithm gets away with the subjectively chosen threshold of discrimination in the constraints, unlike previous works. This circumvents the problem of the regenerated data distribution being unfair had a smaller threshold been chosen.

Algorithm 3 removeDiscrimination($\mathbb{G}, \mathbb{P}, \mathbf{E}_{\mathbb{G}}^{\text{unfair}}$)

1: $\mathbf{w}^* \leftarrow \text{argmin}_{\mathbf{w}} \sum_{e \in \mathbf{E}_{\mathbb{G}}^{\text{unfair}}} \mu_e + \left\| \mathbb{P}(\mathbf{V}) - \prod_{Z \in \mathbf{V}} f^{\mathbf{w}}(\text{FairFlow}_{z, pa(Z)}, \text{UnfairFlow}_{z, pa(Z)}) \right\|^2$

         subject to Eq. 9 and Eq. 10 (Theorem 10)
2: $\mathbb{P}_{\text{new}}(\mathbf{V}) \leftarrow \prod_{Z \in \mathbf{V}} f^{\mathbf{w}^*}(\text{FairFlow}_{z, pa(Z)}, \text{UnfairFlow}_{z, pa(Z)})$ (Theorem 10)
3: **Output:** $\mathbb{P}_{\text{new}}(\mathbf{V})$
---

5. EXPERIMENTS

In this section, we perform experiments to validate the model and input specifications for approximating CPTs using the criminal recidivism graph as shown in Fig. 1. The effectiveness and efficiency of algorithms depend upon their building blocks: (1) Edge

Unfairness, (2) Theorem 3, (3) Algorithm 1. Hence, we focus the experimental section on *Edge Unfairness* formulation, Theorem 12 and Algorithm 1 as they are used in Algorithm 2 and Algorithm 3. In particular, we analyze the relationship between *Cumulative Unfairness* and the upper bound of *Cumulative Unfairness*, and the applicability of our method to realistic scenarios where the causal model and the *CPT*s are unavailable. We first define the causal model by constructing the *CPT*s.

5.1. Causal Model

This paper uses the causal graph shown in Fig. 1 for experiments. This graph is similar to the one constructed in VanderWeele and Staudt (2011); the difference is in the usage of the defendant's attributes like race as nodes contrary to the judge's attributes. The values taken by the nodes are discrete and specified in Supplementary material. For each attribute V in the graph, the conditional probability distribution $\mathbb{P}(V|Pa(V))$ is generated by the following quantities:

- *Parameters*: $\theta_{A \to V} \in [0,1] \ \forall A \in Pa(V)$ where $\theta_{A \to V}$ quantifies the direct influence of parent A on V that is independent of the specific values taken by A and V. $\theta_{A \to V}$ is a property of the edge $A \to V$.
- *Scores*: $\lambda_{A=a \to V=v} \in [0,1] \ \forall A \in Pa(V)$ where $\lambda_{A=a \to V=v}$ quantifies the direct influence of parent A on V. It is dependent on the specific values of A and V.

CPT of node V is computed as the weighted sum of $\lambda_{A=a \to V=v}$ with $\theta_{A \to V}$ being the weights,

$$\mathbb{P}(v|pa(V)) = \sum_{A \in Pa(V)} \theta_{A \to V} \lambda_{A=pa(V)_A \to V=v} \tag{21}$$

To ensure that the *CPT*s satisfy marginality conditions, the following constraints are defined over the parameters and scores: $\sum_{A \in Pa(V)} \theta_{A \to V} = 1$ and $\sum_v \lambda_{A=a \to V=v} = 1, \forall A \in Pa(V)$. We generate 625 models with different combinations of $\{\theta_{A \to J}, \theta_{B \to E} | A \in Pa(J), B \in Pa(E)\}$ that are used to generate *CPT*s while keeping $\lambda_{A=a \to V=v}$ fixed.

5.2. Approximating the CPTs

We implement Algorithm 1 for each *CPT* and solve the constrained least-squares problem (CLSP) to find the optimal solution $\mathbf{w}^*$ (Algorithm 1: Step 5). CLSP is a well-known optimization problem for the linear model. In the case of a non-linear model, we implement a neural network and apply Adam optimizer (Kingma and Ba, 2014) with default hyperparameters to minimize the *MSE* loss. These two models were implemented using SCIKIT-LEARN library and PyTorch library respectively (Paszke et al., 2019).

5.3. Experiment 1: Relationship between $|C_{\mathbf{S}=\mathbf{s},Y=y}|$ and $C_{\mathbf{S}=\mathbf{s},Y=y}^{\mathbf{upper}}$

Utility: We know from Theorem 12 that $C_{\mathbf{S}=\mathbf{s},Y=y}^{\text{upper}} = 0$ and $C_{\mathbf{S}=\mathbf{s},Y=y} = 0$ when edge unfairness $\mu_e = 0$ in all the unfair edges e. Here, we investigate whether decreasing $C_{\mathbf{S}=\mathbf{s},Y=y}^{\text{upper}}$ decreases $|C_{\mathbf{S}=\mathbf{s},Y=y}|$. This investigation provides utility for the formulation of *Potential to Mitigate Cumulative Unfairness* quantity in terms of $C_{\mathbf{S}=\mathbf{s},Y=y}^{\text{upper}}$ as $C_{\mathbf{S}=\mathbf{s},Y=y}$ cannot be expressed in terms of edge unfairness.

Setting: We set $\theta_{A \to V} = 0$ for all A along an unfair edge to 0 except the racial parent R. $\theta_{A \to V}$ is indicative of edge unfairness $\mu_{A \to V}$ from Experiment 5.2. Next, we plot $C^{\text{upper}}_{R=0, J=1}$ and $|C_{R=0, J=1}|$ with varying $\theta_{R \to J}$ for different values of $\theta_{R \to E}$ (Fig. 3(a)) and $\theta_{G \to J}$ (Fig. 3(b)) respectively.

Inference: We observe that for small edge unfairness, decreasing $C^{\text{upper}}_{R=0, J=1}$ decreases $|C_{R=0, J=1}|$. Both converge to 0 when all the edge unfairness are eliminated. This inference helps the policy makers to mitigate cumulative unfairness $|C_{R=0, J=1}|$ by mitigating the upper bound $C^{\text{upper}}_{R=0, J=1}$. On the other hand, as we increase the edge unfairness in the edges other than $R \to J$, this linear trend diminishes as observed for $\theta_{R \to E} = 0.33$ (Fig. 3(a)) and $\theta_{G \to J} = 0.24$ (Fig. 3(b)).

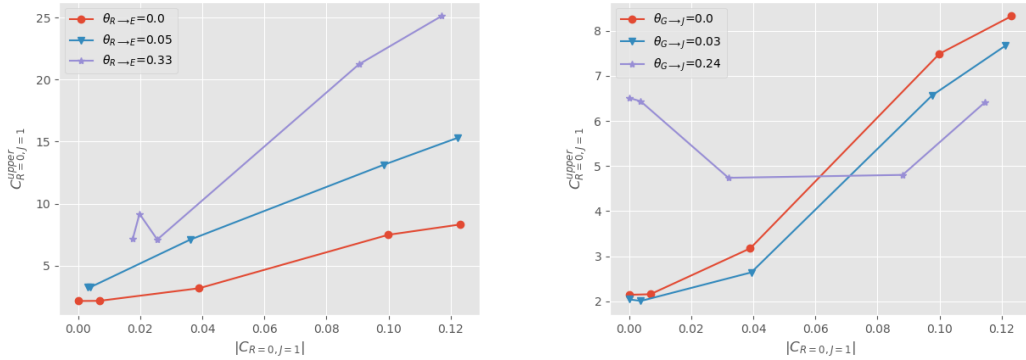Figure 3: Decreasing $C^{\text{upper}}_{\mathbf{S}=\mathbf{s}, Y=y}$ decreases $|C_{\mathbf{S}=\mathbf{s}, Y=y}|$ when edge unfairness μ_e in all unfair edges e are small. (a) $\theta_{R \to J}$, indicative of edge unfairness $\mu_{R \to J}$, is varied for different $\theta_{R \to E}$. (b) $\theta_{R \to J}$ is varied for different $\theta_{G \to J}$. When $\theta_{R \to J}$ is large, decreasing $C^{\text{upper}}_{\mathbf{S}=\mathbf{s}, Y=y}$ does not decrease $|C_{\mathbf{S}=\mathbf{s}, Y=y}|$ as seen from the purple graph.

5.4. Experiment 2: Edge Unfairness with Finite data

Utility: We investigate the applicability of our approach to realistic scenarios where the causal model and the CPTs are unavailable. We do not dwell on discovering causal structures using finite data.[2] Instead, we focus on estimating CPTs using a finite amount of data and compare the edge unfairness calculated using original CPTs, $\mathbb{P}$, with the one by estimated CPTs, $\mathbb{P}^m$, where m is the number of samples drawn randomly from $\mathbb{P}$ for estimation. Intuitively, the distance should decrease as m increases because a large number of i.i.d. samples produce a better approximation of the original distribution $\mathbb{P}$, thereby reducing the euclidean distance.

Setting: In Fig. 4, we plot the euclidean distance $D^{\mathbb{P}}_L(m)$ between $\mathbf{w}^*(\mathbb{P})$ and $\mathbf{w}^*(\mathbb{P}^m)$ by varying m. Here, CPTs are approximated using the linear model for different distributions $\mathbb{P}$ that are randomly generated as shown in different colors. Similarly, the euclidean distance $D^{\mathbb{P}}_{NL}(m)$ between $\mu(\mathbb{P})$ and $\mu(\mathbb{P}^m)$ assuming the non-linear model is shown in Fig. 4(b).

Inference: We observe that $\mathbf{w}^*(\mathbb{P}^m)$ moves closer to $\mathbf{w}^*(\mathbb{P})$ as m increases. Moreover, since $\mathbb{P}$ was randomly generated, we also observe that there exists an empirical bound

---

2. TETRAD software discussed in Ramsey et al. (2018) can be used for this purpose.
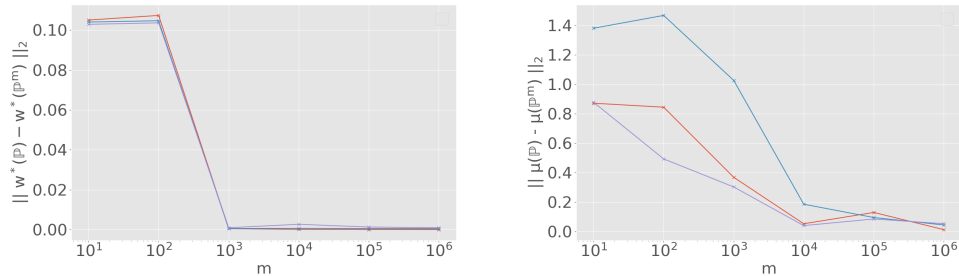
Figure 4: Edge unfairness decreases as the number of samples increases. (a) $||\mathbf{w}^*(\mathbb{P}) - \mathbf{w}^*(\mathbb{P}^m)||_2 = D_L^{\mathbb{P}}(m)$ vs. number of samples m. (b) $||\mu(\mathbb{P}) - \mu(\mathbb{P}^m)||_2 = D_{NL}^{\mathbb{P}}(m)$ vs. number of samples m.

over Euclidean distance for a given m. For instance, in Fig. 4(a), $D_L^{\mathbb{P}}(m)$ is less than 0.01 for m greater than 10^3. A similar observation can be made in Fig. 4(b). Further, more samples are required to make $D_L^{\mathbb{P}}(m)$ and $D_{NL}^{\mathbb{P}}(m)$ comparable. For instance, around 10^3 samples are required to observe $D_L^{\mathbb{P}}(m) = 0.01$, while at least 10^4 samples are required to observe $D_{NL}^{\mathbb{P}}(m) \approx 0.2$. The presence of an empirical bound motivates one to investigate the possibility of a theoretical bound over the Euclidean distance. In addition to the above experiments, we empirically show that the *Edge Unfairness* is a property of an edge and discuss the benefits of using a non-linear model in the Supplementary Material.

6. Related Work

Mitigating Unfairness in the Data Generation Phase: Gebru et al. (2018) suggests documenting the dataset by recording the motivation and creation procedure. However, it does not attempt to provide a solution for mitigation with limited resources.

Assumptions: Zhang et al. (2017) assumes that the sensitive variable S has no parents as it is an inherent nature of the individual. We follow Zhang et al. (2019) that relaxes this assumption because sensitive nodes such as religious belief can have parents like literacy L. Nabi and Shpitser (2018) and Chiappa (2019) propose discrimination removal procedures in the continuous node setting by handling the non-identifiability issues. We do not discuss the continuous variable setting to avoid digressing into the intractability issues. Wu et al. (2019) formulates cumulative unfairness as a solution to the optimization problem for the semi-markovian setting. We restrict our discussion to the markovian setting to avoid digressing into the challenges of formulating cumulative unfairness in terms of edge unfairness. Ravishankar et al. (2020) solves the problem for the trivial linear case when the cumulative unfairness can be expressed in terms of edge unfairness.

Edge Flow: Decomposing direct parental dependencies of a child into independent contributions from each of its parents helps in quantifying the edge flow. Srinivas (1993), Kim and Pearl (1983), and Henrion (2013) separate the independent contributions by using unobserved nodes in the representation of causal independence. To overcome the issues of intractability in unobserved nodes, Heckerman (1993) proposed a temporal definition of causal independence. It states that if and only cause c transitions from time t to $t + 1$, then

the effect's distribution at time $t + 1$ depends only on the effect and the cause at time t, and the cause at time $t + 1$. Based on this definition, a belief network representation is constructed with the observed nodes that make the probability assessment and inference tractable. Heckerman and Breese (1994) proposes a temporal equivalent of the temporal definition of Heckerman (1993). The aforementioned works do not quantify the direct dependencies from the parents onto the child as in our work.

Edge Unfairness: Multiple statistical criteria have been proposed to identify discrimination (Berk et al., 2018) but it is mathematically incompatible to satisfy them all when base rates of the dependent variable differ across groups (Chouldechova, 2017; Kleinberg et al., 2016). Consequently, there is an additional task of selecting which criterion has to be achieved. Moreover, statistical criteria caution about discrimination but do not help in identifying the sources of unfairness. Zhang et al. (2017) uses path-specific effects to identify direct and indirect discrimination after data is generated but does not address the problem of mitigating unfairness in the data generation phase. Unlike Zhang et al. (2017) that uses the presence of a redlining attribute in an indirect path and the presence of a sensitive node on a direct path to determine the unfairness of a path, our work uses the notion of an unfair edge as the potential source of unfairness akin to Chiappa and Isaac (2018).

Discrimination Removal Procedure: Zhang et al. (2017) and Kusner et al. (2017) remove discrimination by altering the data distribution. *Firstly*, the optimization technique in Zhang et al. (2017) and the sampling procedure in Kusner et al. (2017), scale exponentially in the number of nodes (and values taken by the sensitive nodes) that eventually increases the time to solve the quadratic programming problem. *Secondly*, the constraints in Zhang et al. (2017) depend on a subjectively chosen threshold of discrimination that is disadvantageous because the regenerated data distribution would remain unfair had a smaller threshold been chosen. Our paper formulates a discrimination removal procedure without exponentially growing constraints and a threshold of discrimination.

7. CONCLUSION

We introduce the problem of quantifying edge unfairness in an unfair edge. We give a novel formulation that models $CPTs$ in terms of edge flows to quantify edge unfairness. We prove a result that eliminating edge unfairness eliminates cumulative unfairness. Proving this result is not straightforward because cumulative unfairness cannot be expressed in terms of edge unfairness when $CPTs$ are modeled as a non-parametric function of the edge flows. Hence, we prove the result via an intermediate theorem that upper bounds the magnitude of cumulative unfairness by a quantity that can be expressed in terms of edge unfairness. To analyze the impact of edge unfairness on cumulative unfairness, we quantify the potential to mitigate cumulative unfairness when edge unfairness is decreased. This formulation uses the upper bound of cumulative unfairness as it can be expressed in terms of edge unfairness. Experimental results validate that mitigating cumulative unfairness mitigates its upper bound as well, thereby establishing the rationale for using the upper bound of cumulative unfairness in the formulation. Using the theorem result and measures, we present an unfair edge prioritization algorithm and a discrimination removal algorithm. The unfair edge prioritization algorithm gives tangible directions to agencies to mitigate unfairness in the real world while the data is being generated. There is no utility in making cautionary

claims of potential discrimination when it is not complemented with information that aids in mitigating unfairness causing discrimination. On the other hand, the discrimination removal algorithm de-biases data after the data is generated. In the future, we aim to evaluate the impact of edge unfairness on subsequent stages of the machine learning pipeline such as selection, classification, etc. We also plan to extend to the semi-Markovian causal model (Wu et al., 2019) and continuous nodes settings (Nabi and Shpitser, 2018).

References

Civil Rights Act. Civil rights act of 1964. *Title VII, Equal Employment Opportunities*, 1964.

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

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533, 2018.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

Silvia Chiappa and William S Isaac. A causal bayesian networks viewpoint on fairness. In *IFIP International Summer School on Privacy and Identity Management*, pages 3–20. Springer, 2018.

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

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

David Heckerman. Causal independence for knowledge acquisition and inference. In *Uncertainty in Artificial Intelligence*, pages 122–127. Elsevier, 1993.

David Heckerman and John S Breese. A new look at causal independence. In *Uncertainty Proceedings 1994*, pages 286–292. Elsevier, 1994.

Max Henrion. Practical issues in constructing a bayes' belief network. *arXiv preprint arXiv:1304.2725*, 2013.

JinHyung Kim and Judea Pearl. A computational model for causal and diagnostic reasoning in inference systems. In *International Joint Conference on Artificial Intelligence*, pages 0–0, 1983.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Judea Pearl. *Causality.* Cambridge university press, 2009.

Joseph D Ramsey, Kun Zhang, Madelyn Glymour, Ruben Sanchez Romero, Biwei Huang, Imme Ebert-Uphoff, Savini Samarasinghe, Elizabeth A Barnes, and Clark Glymour. Tetrad—a toolbox for causal discovery. In *8th International Workshop on Climate Informatics*, 2018.

Pavan Ravishankar, Pranshu Malviya, and Balaraman Ravindran. A causal linear model to quantify edge unfairness for unfair edge prioritization and discrimination removal. *arXiv e-prints*, pages arXiv–2007, 2020.

Sampath Srinivas. A generalization of the noisy-or model. In *Uncertainty in artificial intelligence*, pages 208–215. Elsevier, 1993.

Tyler J VanderWeele and Nancy Staudt. Causal diagrams for empirical legal research: a methodology for identifying causation, avoiding bias and interpreting results. *Law, Probability & Risk*, 10(4):329–354, 2011.

Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. Pc-fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*, pages 3404–3414, 2019.

L. Zhang, Y. Wu, and X. Wu. Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 31(11):2035–2050, 2019.

Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017.