Title: A Causal Approach for Unfair Edge Prioritization and Discrimination Removal

Supplementary Material

1. Choices for $f^{\mathbf{w}}$

We present two instances for $f^{\mathbf{w}}$. The list is not limited to these and can be extended as long as $f^{\mathbf{w}}$ satisfies the constraints of the conditional probability (Eq. 9, 10).

1. $f^{\mathbf{w}}$ is a linear combination in the inputs where,

$$f^{\mathbf{w}}(\mathbb{P}_{\text{flow}}^{pa(X)_{\mathbf{F}_X}}(X), \bigcup_{pa(X)_A \in \mathbf{U}_X} \mathbb{P}_{\text{flow}}^{pa(X)_A}(X)) = w_{\mathbf{F}_X \to X}\mathbb{P}_{\text{flow}}^{pa(X)_{\mathbf{F}_X}}(X) + \sum_{A\in\mathbf{U}_X} w_{A\to X}\mathbb{P}_{\text{flow}}^{pa(X)_A}(X) \quad (22)$$

$$\text{subject to, } 0 \le w_{\mathbf{F}_X\to X}, w_{A\to X} \le 1, \forall A \in \mathbf{U}_X \quad (23)$$

$$w_{\mathbf{F}_X\to X} + \sum_{A\in\mathbf{U}_X} w_{A\to X} = 1 \quad (24)$$

$w_{\mathbf{F}_X\to X}$ and $w_{A\to X}$ are constrained between 0 and 1 since the objective of the mapper $f^{\mathbf{w}}$ is to capture the interaction between the fraction of the beliefs given by $w_{\mathbf{F}_X\to X}\mathbb{P}_{\text{flow}}^{pa(X)_{\mathbf{F}_X}}(X)$ and $\bigcup_{A\in\mathbf{U}_X} w_{A\to X}\mathbb{P}_{\text{flow}}^{pa(X)_A}(X)$ and approximate $P(x|pa(X))$. Eq. 23 and Eq. 24 ensure that the conditional probability axioms of $f^{\mathbf{w}}$ are satisfied.

2. $f^{\mathbf{w}} = f_N^{\mathbf{w}_N} \circ ... \circ f_1^{\mathbf{w}_1}$ is composite function representing a N-layer neural network with i^{th} layer having M_i neurons and weights $\mathbf{w}_i$ capturing the non-linear combination of the inputs where,

$$f^{\mathbf{w}}(\mathbb{P}_{\text{flow}}^{pa(X)_{\mathbf{F}_X}}(x), \bigcup_{A\in\mathbf{U}_X} \mathbb{P}_{\text{flow}}^{pa(X)_A}(x)) = f_N(...f_1(\mathbb{P}_{\text{flow}}^{pa(X)_{\mathbf{F}_X}}(x), \bigcup_{A\in\mathbf{U}_X} \mathbb{P}_{\text{flow}}^{pa(X)_A}(x))) \quad (25)$$

$$\text{subject to, } f_i : \mathbb{R}^{M_i} \to [0,1]^{|X|}, \quad (26)$$

$$\sum_x f^{\mathbf{W}}(\mathbb{P}_{\text{flow}}^{pa(X)_{\mathbf{F}_X}}(x), \bigcup_{A\in\mathbf{U}_X} \mathbb{P}_{\text{flow}}^{pa(X)_A}(x)) = 1 \quad (27)$$

$f^{\mathbf{W}}$ captures the interaction between $\mathbb{P}_{\text{flow}}^{pa(X)_{\mathbf{F}_X}}(x)$ and $\bigcup_{A\in\mathbf{U}_X} \mathbb{P}_{\text{flow}}^{pa(X)_A}(x)$ and models $P(x|pa(X))$. Eq. 26 and Eq. 27 ensure that the conditional probability axioms of $f^{\mathbf{W}}$ are satisfied. One possibility is to use a softmax function for f_N to ensure that the outputs of $f^{\mathbf{W}}$ satisfy probability axioms.

2. Proof of Theorem 12 & Corollary 13

Proof of Theorem 12

$$C_{\mathbf{S}=\mathbf{s},Y=y} \quad (28)$$

$$= \sum_{\mathbf{s}'\in\mathbf{S}\setminus\mathbf{s}} TE_{Y=y}(\mathbf{s},\mathbf{s}')\mathbb{P}(\mathbf{s}') \quad (29)$$

$$= \sum_{\mathbf{s}' \in \mathbf{S} \backslash \mathbf{s}} [\mathbb{P}(y|do(\mathbf{s})) - \mathbb{P}(y|do(\mathbf{s}'))]\mathbb{P}(\mathbf{s}') \tag{30}$$

$$= \sum_{\mathbf{s}' \in \mathbf{S} \backslash \mathbf{s}} \mathbb{P}(\mathbf{s}')[\sum_{\mathbf{v}_1 \in \mathbf{V} \backslash Y} \mathbb{P}(\mathbf{v}_1, y|do(\mathbf{s})) - \sum_{\mathbf{v}_2 \in \mathbf{V} \backslash Y} \mathbb{P}(\mathbf{v}_2, y|do(\mathbf{s}'))]$$

[$\mathbf{v}_1$ is consistent with $\mathbf{s}$ and $\mathbf{v}_2$ is consistent with $\mathbf{s}'$.] (31)

$$= \sum_{\mathbf{s}' \in \mathbf{S} \backslash \mathbf{s}} \mathbb{P}(\mathbf{s}')[\sum_{\mathbf{v} \in \mathbf{V} \backslash \{\mathbf{S}, Y\}} \prod_{V \in \mathbf{V} \backslash \{\mathbf{S}, Y\}, Y=y} \mathbb{P}(v|pa(V))|_{\mathbf{s}} - \sum_{\mathbf{v} \in \mathbf{V} \backslash \{\mathbf{S}, Y\}} \prod_{V \in \mathbf{V} \backslash \{\mathbf{S}, Y\}, Y=y} \mathbb{P}(v|pa(V))|_{\mathbf{s}'}]$$

[Definition 3] (32)

$$= \sum_{\mathbf{s}' \in \mathbf{S} \backslash \mathbf{s}} \mathbb{P}(\mathbf{s}')[\sum_{\mathbf{v} \in \mathbf{V} \backslash \{\mathbf{S}, Y\}} \prod_{V \in \mathbf{V} \backslash \{\mathbf{S}, Y\}, Y=y} f_V(\mathbb{P}_{\text{flow}}^{pa(V)_{\mathbf{F}_V}}(v), \bigcup_{A \in \mathbf{U}_V} \mathbb{P}_{\text{flow}}^{\mathbf{s}_A \vee pa(V)_A}(v)) -$$
$$\sum_{\mathbf{v} \in \mathbf{V} \backslash \{\mathbf{S}, Y\}} \prod_{V \in \mathbf{V} \backslash \{\mathbf{S}, Y\}, Y=y} f_V(\mathbb{P}_{\text{flow}}^{pa(V)_{\mathbf{F}_V}}(v), \bigcup_{A \in \mathbf{U}_V} \mathbb{P}_{\text{flow}}^{\mathbf{s}'_A \vee pa(V)_A}(v))]$$

[Theorem 10 and Notation 18] (33)

$$\leq \sum_{\mathbf{s}' \in \mathbf{S} \backslash \mathbf{s}} \mathbb{P}(\mathbf{s}')[\sum_{\mathbf{v} \in \mathbf{V} \backslash \{\mathbf{S}, Y\}} \prod_{V \in \mathbf{V} \backslash \{\mathbf{S}, Y\}, Y=y} [f_V(\mathbb{P}_{\text{flow}}^{pa(V)_{\mathbf{F}_V}}(v), \bigcup_{A \in \mathbf{U}_V \backslash B} \mathbb{P}_{\text{flow}}^{\mathbf{s}_A \vee pa(V)_A}(v),$$
$$\mathbb{P}_{\text{flow}}^{\mathbf{s}_B \vee pa(V)_B}(v) = 0) + \frac{\mathbb{P}_{\text{flow}}^{\mathbf{s}_B \vee pa(V)_B}(v)}{\mathbb{P}(v, pa(V))|_{\mathbf{s}}} \mu_{B \to V}] - \sum_{\mathbf{v} \in \mathbf{V} \backslash \{\mathbf{S}, Y\}} \prod_{V \in \mathbf{V} \backslash \{\mathbf{S}, Y\}, Y=y}$$
$$[f_V(\mathbb{P}_{\text{flow}}^{pa(V)_{\mathbf{F}_V}}(v), \bigcup_{A \in \mathbf{U}_V \backslash B} \mathbb{P}_{\text{flow}}^{\mathbf{s}'_A \vee pa(V)_A}(v), \mathbb{P}_{\text{flow}}^{\mathbf{s}'_B \vee pa(V)_B}(v) = 0) - \frac{\mathbb{P}_{\text{flow}}^{\mathbf{s}'_B \vee pa(V)_B}(v)}{\mathbb{P}(v, pa(V))|_{\mathbf{s}'}} \mu_{B \to V}]]$$

[Definition 12, property that $|.| \geq 0,$ and Notation 18] (34)

$$\leq \sum_{\mathbf{s}' \in \mathbf{S} \backslash \mathbf{s}} \mathbb{P}(\mathbf{s}')[\sum_{\mathbf{v} \in \mathbf{V} \backslash \{\mathbf{S}, Y\}} \prod_{V \in \mathbf{V} \backslash \{\mathbf{S}, Y\}, Y=y} [f_V(\mathbb{P}_{\text{flow}}^{pa(V)_{\mathbf{F}_V}}(v), \bigcup_{A \in \mathbf{U}_V} \mathbb{P}_{\text{flow}}^{\mathbf{s}_A \vee pa(V)_A}(v) = 0)$$
$$+ \sum_{A \in \mathbf{U}_V} \frac{\mathbb{P}_{\text{flow}}^{\mathbf{s}_A \vee pa(V)_A}(v)}{\mathbb{P}(v, pa(V))|_{\mathbf{s}}} \mu_{A \to V}] - \sum_{\mathbf{v}_2 \in \mathbf{V} \backslash \{\mathbf{S}, Y\}} \prod_{V \in \mathbf{V} \backslash \{\mathbf{S}, Y\}, Y=y} [f_V(\mathbb{P}_{\text{flow}}^{pa(V)_{\mathbf{F}_V}}(v),$$
$$\bigcup_{A \in \mathbf{U}_V} \mathbb{P}_{\text{flow}}^{\mathbf{s}'_A \vee pa(V)_A}(v) = 0) - \sum_{A \in \mathbf{U}_\mathbf{V}} \frac{\mathbb{P}_{\text{flow}}^{\mathbf{s}'_A \vee pa(V)_A}(v)}{\mathbb{P}(v, pa(V))|_{\mathbf{s}'}} \mu_{A \to V}]]$$

[Recursively apply previous step for every A $\in \mathbf{U}_V$ and Notation 18] (35)

$$\leq \sum_{\mathbf{s}' \in \mathbf{S} \setminus \mathbf{s}} \mathbb{P}(\mathbf{s}')[\sum_{\mathbf{v} \in \mathbf{V} \setminus \{\mathbf{S}, Y\}} \prod_{V \in \mathbf{V} \setminus \{\mathbf{S}, Y\}, Y=y} \sum_{A \in \mathbf{U}_V} \frac{\mathbb{P}_{\text{flow}}^{\mathbf{s}_A \vee pa(V)_A}(v)}{\mathbb{P}(v, pa(V))|_{\mathbf{s}}} \mu_{A \to V} +$$

$$\sum_{\mathbf{v} \in \mathbf{V} \setminus \{\mathbf{S}, Y\}} \prod_{V \in \mathbf{V} \setminus \{\mathbf{S}, Y\}, Y=y} \sum_{A \in \mathbf{U}_{\mathbf{V}}} \frac{\mathbb{P}_{\text{flow}}^{\mathbf{s}'_A \vee pa(V)_A}(v)}{\mathbb{P}(v, pa(V))|_{\mathbf{s}'}} \mu_{A \to V}] \tag{36}$$

$$\leq \sum_{\mathbf{s}' \in \mathbf{S} \setminus \mathbf{s}} \mathbb{P}(\mathbf{s}')[\sum_{v \in \mathbf{V} \setminus \{\mathbf{S}, Y\}} \prod_{V \in \mathbf{V} \setminus \{\mathbf{S}, Y\}, Y=y} \sum_{A \in \mathbf{U}_V} \left[\frac{\mathbb{P}_{\text{flow}}^{\mathbf{s}_A \vee pa(V)_A}(v)}{\mathbb{P}(v, pa(V))|_{\mathbf{s}}} + \frac{\mathbb{P}_{\text{flow}}^{\mathbf{s}'_A \vee pa(V)_A}(v)}{\mathbb{P}(v, pa(V))|_{\mathbf{s}'}} \right] \mu_{A \to V}] \tag{37}$$

Thus,

$$C_{\mathbf{S}=\mathbf{s}, Y=y} \leq C^{\text{upper}}_{\mathbf{S}=\mathbf{s}, Y=y} \quad \text{[From Eq. 37 and Eq. 17]} \tag{38}$$

$$C_{\mathbf{S}=\mathbf{s}, Y=y} \geq -C^{\text{upper}}_{\mathbf{S}=\mathbf{s}, Y=y} \quad \text{[Similar proof]} \tag{39}$$

$$\therefore |C_{\mathbf{S}=\mathbf{s}, Y=y}| \leq C^{\text{upper}}_{\mathbf{S}=\mathbf{s}, Y=y} \quad \blacksquare \tag{40}$$

Proof of Corollary 13

When edge unfairness $\mu_{e,\mathbb{G}} = 0,\ \forall e$ from $\mathbf{S}$,

$$C^{\text{upper}}_{\mathbf{S}=\mathbf{s}, Y=y} = 0$$

$$\text{[Edge unfairness } \mu_{e,\mathbb{G}} = 0\ \forall e \text{ from } \mathbf{S} \text{ and Eq. 17]} \tag{41}$$

$$C_{\mathbf{S}=\mathbf{s}, Y=y} = 0$$

$$\text{[Eq. 40 and Eq. 41]} \quad \blacksquare \tag{42}$$

3. Experiments - Additional Details

The values taken by each of the node in the causal graph 1 are shown in Table 1.

3.1 Edge Unfairness is an Edge Property

We investigate that the edge unfairness depends on the parameters of the edge and not on the specific values of the attributes.

Inference: When a linear model is used, $\mathbf{w}^*$ is observed to be insensitive to the specific values taken by the nodes as there is minimal variation in $\mathbf{w}^*_e$ for any fixed θ_e as shown in Fig. 5(a). $w^*_{R \to J}$ was observed to be in the range $[0.2, 0.3]$ for different $\theta_{R \to J}$. A small

Table 1: Nodes and their Values.

Node	Values
Race R	African American(0), Hispanic(1) and White(2)
Gender G	Male(0), Female(1) and Others(2)
Age A	Old (0)(>35y) and Young (1) ($\leq$ 35y)
Literacy L	Literate (0) and Illiterate (1)
Employment E	Not Employed (0) and Employed (1)
Bail Decision J	Bail granted (0) and Bail rejected (1)
Case History C	Strong (0) and Weak criminal history (1)

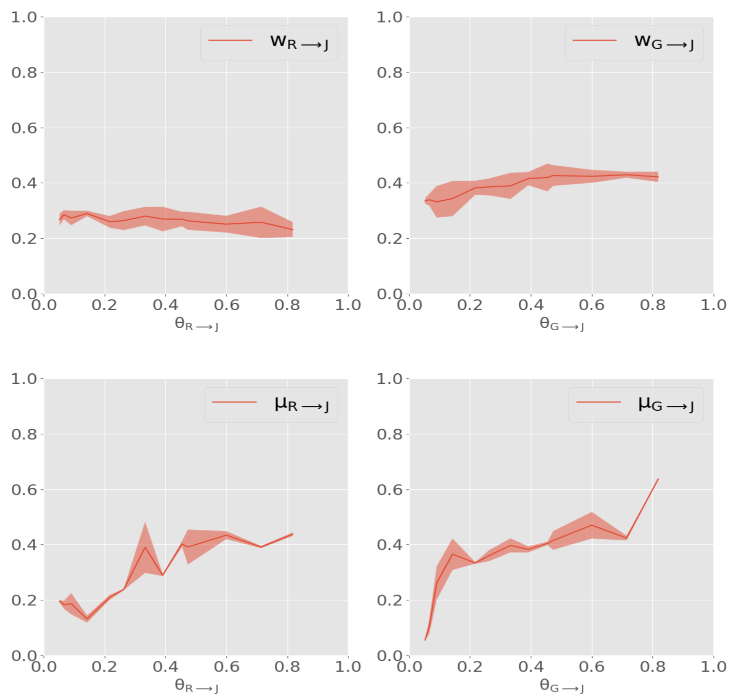

Figure 5: Edge unfairness is a property of the edge because there is minimal variation in edge unfairness for a specific θ_e. (a) $w^*_{R\to J}$ vs. $\theta_{R\to J}$ and $w^*_{G\to J}$ vs. $\theta_{G\to J}$ for linear model. (b) $\mu_{R\to J}$ vs. $\theta_{R\to J}$ and $\mu_{G\to J}$ vs. $\theta_{G\to J}$ for non-linear model.

deviation in $w^*_{R\to J}$ shows that $w^*_{R\to J}$ depends only on $\theta_{R\to J}$ and not on the specific values taken by the nodes. Since edge unfairness in an edge, say $R \to J$, is $\mu_{R\to J} = |Pa(J)|w_{R\to J}$ in the linear model setting, it indicates that edge unfairness is also insensitive to the specific values taken by nodes and hence is a property of the edge. Similarly for the non-linear model, edge unfairness μ_e is insensitive to the specific values taken by the nodes as there is minimal variation in μ_e for any fixed θ_e as observed from Fig. 5(b). For instance, $\mu_{R\to J}$ obtained in the models with $\theta_{R\to J} = 0.5$ are in the range $[0.35, 0.43]$. A similar observation can be made for $w^*_{G\to J}$ and $\mu_{G\to J}$ in Fig. 5(a) and 5(b) respectively. We also analyze the MSE for both the linear and non-linear settings in Supplementary material.

3.2 Linear and Non-linear model comparison

To validate the benefits of a non-linear model, the MSEs between the CPTs for bail decision $\mathbb{P}(J|Pa(J))$ and its functional approximation $f^{\mathbf{w}}$ were recorded for these settings:

1. MSEs e^L_J calculated when $f^{\mathbf{w}}$ is approximated using a linear model (Eq. 22)
2. MSEs e^{NL}_J calculated when $f^{\mathbf{w}}$ is approximated using a non-linear model (Eq. 25)

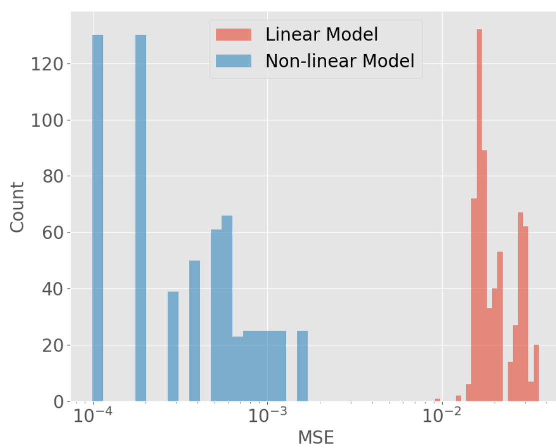

Figure 6: Histogram for MSE by using a linear model shown in red and using a non-linear model shown in blue for 625 different models (discussed in Section 5.1).

Inference: Distributions of e^L_J and e^{NL}_J are plotted in Fig. 6. Here, the maximum value of e^L_J shown in the red bar is obtained above 0.01 and its values mostly lie in the range $(0.01, 0.02)$. On the other hand, e^{NL}_J shown in blue bars is distributed in the range $(0.0001, 0.001)$ with the maximum value of e^{NL}_J obtained around 0.002. Hence, a non-linear model like a neural network to approximate $f^{\mathbf{w}}$ is a better choice because the MSEs distribution lies in the lower error range.