Proceedings of Machine Learning Research 275:1–19, 2025 4th Conference on Causal Learning and Reasoning

Fair Clustering: A Causal Perspective

Fritz Bayer FRBAYER@ETHZ.CH
D-BSSE, ETH Zurich, Basel, Switzerland

Drago Plečko DP3144@COLUMBIA.EDU
Department of Computer Science, Columbia University, New York, USA

Niko Beerenwinkel NIKO.BEERENWINKEL@BSSE.ETHZ.CH
D-BSSE, ETH Zurich, Basel, Switzerland

Jack Kuipers JACK.KUIPERS@BSSE.ETHZ.CH
D-BSSE, ETH Zurich, Basel, Switzerland

Editors: Biwei Huang and Mathias Drton

Abstract

Clustering algorithms may unintentionally propagate or intensify existing disparities, leading to unfair representations or biased decision-making. Current fair clustering methods rely on notions of fairness that do not capture any information on the underlying causal mechanisms. We show that optimising for non-causal fairness notions can paradoxically induce direct discriminatory effects from a causal standpoint. We present a clustering approach that incorporates causal fairness metrics to provide a more nuanced approach to fairness in unsupervised learning. Our approach enables the specification of the causal fairness metrics that should be minimised. We demonstrate the efficacy of our methodology using datasets known to harbour unfair biases.

Keywords: Algorithmic fairness, causality, clustering

1. Introduction

Clustering is a fundamental unsupervised learning technique which is used in various domains to uncover hidden structures and patterns in data (Kanungo et al., 2002; Bayer et al., 2023). However, clustering algorithms may inadvertently propagate or even exacerbate existing disparities related to protected attributes such as gender, race, or religion (Chierichetti et al., 2017; Barocas et al., 2017). These disparities can lead to biased representations or decision-making processes that may disproportionately affect marginalized communities (see Example 1). In recent years, there has been a growing interest in developing clustering algorithms that strive to balance the trade-off between quality and fairness of the clustering (Chierichetti et al., 2017; Backurs et al., 2019; Bera et al., 2019; Chhabra et al., 2021).

Despite recent advancements in fair clustering, existing methods predominantly rely on non-causal notions of fairness, neglecting the underlying causal structures that may inadvertently contribute to unfairness within the data. This oversight can be particularly significant, as considering these causal structures might be not only ethically pertinent but also a legal requirement (Barocas and Selbst, 2016). Achieving causal fairness involves comprehending and modelling the causal relationships between variables using adequate causal semantics. Remarkably, while the causal approach to fairness has been thoroughly explored in the context of supervised learning (Kusner et al., 2017; Zhang and Bareinboim, 2018b,a; Chiappa, 2019; Wu et al., 2019; Nabi and Shpitser, 2018; Plečko and Bareinboim, 2024), its incorporation into unsupervised methods, such as clustering, remains uncharted.

© 2025 F. Bayer, D. Plečko, N. Beerenwinkel & J. Kuipers.

In this work, we introduce a novel causally fair clustering approach that integrates the causal structure of the data. Our approach allows one to specify which causal fairness metrics should be optimised, offering a more targeted perspective on fair clustering. We demonstrate how to perform causally fair clustering to mitigate direct, indirect, and spurious sources of unfairness.

Our key contributions are as follows:

- We show that fair clustering algorithms that are based on non-causal notions of fairness can induce direct discriminatory effects from a causal standpoint (see Example 2 and Figure 3).
- We propose an alternative clustering approach that incorporates causal fairness notions, allowing for a more detailed understanding of fairness in unsupervised learning (see Theorem 7 and Algorithm 1).
- We demonstrate the effectiveness of our causally fair clustering approach on two real-world datasets, showing improvements in causal fairness metrics compared to previous clustering methods (see Figure 3).

1.1. Related Work

Clustering methods with fairness constraints can generally be grouped into those that aim for group fairness and those that aim for individual fairness (Chhabra et al., 2021).

In the domain of group fairness, Chierichetti et al. (2017) proposed a method of fair clustering that seeks to create balanced clusters using a technique known as fairlet decomposition. This concept of balanced clusters has been further extended to account for various lp-norm cost functions, multiple groups, relaxed balance requirements, and scalability issues (Chierichetti et al., 2017; Rösner and Schmidt, 2018; Bera et al., 2019; Bercea et al., 2019; Ahmadian et al., 2019; Kleindessner et al., 2019b).

Kleindessner et al. (2019a) introduced a novel approach that focuses on fair representation within selected cluster centres. Their formulation blends k-center objectives with a partition matroid constraint. Building on this, Jones et al. (2020) put forth an approximation algorithm for this paradigm. The recent advent of the socially fair clustering paradigm, as posited by Abbasi et al. (2021) and Ghadiri et al. (2021), seeks to balance the incurred costs across groups. Makarychev and Vakilian (2021) and Goyal and Jaiswal (2023) proposed approximations in this context.

At the intersection of group and individual fairness, recent work has also focused on the fair allocation of public resources in clustering contexts (Chen et al., 2019; Micha and Shah, 2020). These approaches propose that a fair clustering is one where no subset of points has an incentive to assign themselves to a centre outside their designated cluster.

On the other hand, the notion of individual fairness has also received considerable attention. The key is the idea of stability in clusters, where each point in a cluster should not have an average distance to its own cluster larger than to any other cluster (Kleindessner et al., 2020). Anderson et al. (2020) extended the concept of individual fairness to incorporate distributional assignments, ensuring similar points receive analogous assignments, and Brubach et al. (2021) introduced pairwise fairness, leveraging point distance to guide separations.

The concept of priority k-center, which deals with usage weights and metric embedding, has also been considered in the context of fair clustering (Jung et al., 2019; Mahabadi and Vakilian, 2020). Negahbani and Chakrabarty (2021) proposed a bi-criteria approximation algorithm for individually fair k-clustering, which has been extended by Vakilian and Yalciner (2022).

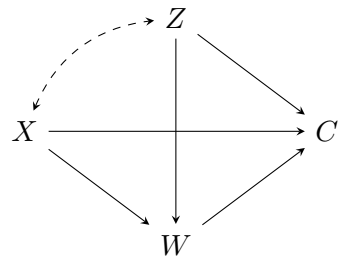

Figure 1: Standard fairness model in clustering.

In conclusion, the landscape of fair clustering is evidently marked by a plethora of non-causal strategies, with various methods prioritising different non-causal notions of fairness and optimising for different types of constraints. In contrast to this prevalent trend, our contribution pivots to a causal perspective on fair clustering, bringing forth a novel approach that optimises causal notions of fairness.

Example 1 (Job advertisement) *Consider the application of clustering in the field of job advertisement, where it is used for target audience segmentation. The clusters are learned based on demographic variables W (e.g., interests, behaviour, income, education, and occupation) and geographic variables Z (e.g., country and ZIP code). The clusters leading to specialised advertisements should be fair with respect to the protected attributes X (e.g., sex, gender or race). For instance, we want to avoid that ads for high-paying jobs might be unfairly clustered towards a certain gender or ethnic group, perpetuating systemic biases.*

2. Preliminaries

We use the language of structural causal models (SCMs) as our semantic framework (Pearl, 2000). A SCM is a 4-tuple $\mathcal{M} := \langle V, U, \mathcal{F}, P(u) \rangle$ that encapsulates the following elements:

- V and U represent endogenous (observable) and exogenous (unobserved) variable sets, respectively.
- $\mathcal{F}$ encompasses a collection of functions f_{V_i} with $V_i \in V$. Each individual function is defined as $V_i \leftarrow f_{V_i}(\mathrm{pa}(V_i), U_{V_i})$, with $\mathrm{pa}(V_i) \subseteq V$ and $U_{V_i} \subseteq U$.
- $P(u)$ represents a probability distribution over the exogenous variables U.

A SCM is associated with a causal diagram, $\mathcal{G}$, built over the node set V. Within this diagram, a directed edge $V_i \rightarrow V_j$ emerges if V_i serves as an argument for f_{V_j}, while an undirected edge $V_i \dashleftarrow\dashrightarrow V_j$ appears if the related U_{V_i}, U_{V_j} are interdependent. Throughout, an instantiation of the exogenous variables $U = u$ is termed a *unit*. $Y_x(u)$ captures the potential outcome of Y upon imposing the condition $X = x$ for the given unit u. Specifically, $Y_x(u)$ represents the solution for $Y(u)$ in the submodel $\mathcal{M}_x$, in which the equations linked to X within $\mathcal{F}$ are substituted with the condition $X = x$.

2.1. Defining Causal Fairness in Clustering

To define fairness in unsupervised learning, we adapt the formalism introduced for supervised learning by Plečko and Bareinboim (2024). Consider the causal diagram displayed in Figure 1, called the standard fairness model (SFM). In the case of clustering, the outcome variable is the cluster assignment C, which can either be a discrete assignment c_k to one of the K clusters with $k \in \{1, \ldots, K\}$, or a continuous K-dimensional cluster membership probability.

Definition 1 (Standard Fairness Model, Plečko and Bareinboim, 2024) *The SFM is represented by the causal diagram G_{SFM} with endogenous variables $\{X, Z, W, C\}$, which are defined by*

- *The protected attribute, denoted by X (e.g., gender, race, religion),*
- *The set of confounding variables Z, which are not causally influenced by the attribute X (e.g., demographic information),*
- *The set of mediator variables W that could be causally influenced by the attribute X (e.g., educational level or other job-related information),*
- *The cluster assignment, denoted by C.*

Nodes Z and W are possibly multi-dimensional or empty. Furthermore, for a causal diagram G, the projection of G onto the SFM is defined as the mapping of the endogenous variables V appearing in G into four groups X, Z, W, C, as described above. The projection is denoted by $\Pi_{SFM}(G)$ and is constructed by choosing the protected attribute and grouping the confounders Z and mediators W.

The standard fairness model allows us to define fair cluster associations according to common fairness measures. Note that the variables suggested for each category of the standard fairness model should not be generalised and have to be carefully selected for each dataset. In the clustering setting, the predicted attributes are the cluster associations (or cluster membership probabilities) C of the K clusters. For the sake of simplicity, we take X to be binary, while Z and W can be categorical or continuous.

A range of possible causal fairness constraints can be added to this model. First, following the approach of (Pearl, 2001), we use the natural direct effect (NDE) as a notion of a direct effect along the edge $X \to C$

$$\text{NDE}_{x_0,x_1}(c_k) = P\big((c_k)_{x_1,W_{x_0}}\big) - P\big((c_k)_{x_0}\big). \tag{1}$$

The counterfactual distribution $P((c_k)_{x_0})$ denotes the causal effect of an intervention $do(X = x_0)$ (Pearl, 2000, Ch. 3) on c_k, where the counterfactual variable $(c_k)_{x_0}$ denotes the response of c_k to the intervention. Following the same notation, the counterfactual distribution $P\big((c_k)_{x_1,W_{x_0}}\big)$ describes how c_k changes when setting X to x_1, but keeping the mediators W at the value it would have taken had X been x_0.

To achieve that the constructed cluster assignments c_k are fair with respect to the NDE, one may require that

$$\text{NDE}_{x_0,x_1}(c_k) = \text{NDE}_{x_1,x_0}(c_k) = 0, \tag{2}$$

$\forall k \in \{1, \ldots, K\}$, that is, for every cluster there is a separate constraint. Under the assumptions of the SFM introduced above, the NDE can be identified (i.e., uniquely computed) from the observational data. Thus, the constraint $\text{NDE}_{x_0,x_1}(c_k) = 0$ can be written as

$$\sum_{z,w} [P(c_k \mid x_1, z, w) - P(c_k \mid x_0, z, w)] \cdot P(w \mid x_0, z)P(z) = 0. \tag{3}$$

This constraint ensures that the direct effect, when averaged across different values of Z, W, equals 0 and thus guarantees absence of any population direct effect in the clustering assignment.

Turing around the definition of the NDE, we can obtain a notion for the natural indirect effect (NIE)

$$\text{NIE}_{x_0,x_1}(c_k) = P\big((c_k)_{x_0,W_{x_1}}\big) - P\big((c_k)_{W_{x_0}}\big). \tag{4}$$

which quantifies the difference of the mediator values W_{x_0} and W_{x_1} on c_k had X been x_0. Under the assumptions of the SFM, the NIE is identifiable from observational data, and hence the requirement $\text{NIE}_{x_0,x_1}(c_k) = 0$ can be written as

$$\sum_{z,w} P(c_k \mid x_0, z, w)[P(w \mid x_1, z) - P(w \mid x_0, z)]P(z) = 0. \tag{5}$$

Further, we define the total variation (TV) as

$$\text{TV}_{x_0,x_1}(c_k) = P(c_k \mid x_1) - P(c_k \mid x_0), \tag{6}$$

which we will relate to the natural direct and indirect effects over the next sections.

3. Causal Perspective on Fair Clustering

3.1. Fairness Through Unawareness in Clustering

We start with an informal recap of fairness through unawareness (FTU), which means neglecting the protected attribute throughout the analysis in supervised learning before discussing the unsupervised learning case. We exclude Z during this discussion.

Supervised Learning (Classification). Figure 2a shows the connections among the variables in a supervised learning scenario. The predicted attribute in this scenario is Y. An example could be that we want to predict the salary Y based on the occupation W, while accounting for the protected attribute gender X. The measured data could be subject to the following unfair biases: first, a direct bias δ_1 along $X \to Y$, and second, an indirect bias δ_2 along $X \to W$. Both biases are intrinsic to the data and hence have to be adjusted for in order to be removed. Since the indirect bias δ_2 is propagated via $W \to Y$, we denote the resulting bias along $X \to W \to Y$ as $\hat{\delta}_2$.

Including the protected attribute X in the classification of Y means that we will learn both the direct and indirect effects in the classifier. Hence, the total variation is a combination of the direct and indirect effect

$$\text{TV}_{x_0,x_1}(y) = \underbrace{\delta_1}_{X \to Y} + \underbrace{\hat{\delta}_2}_{X \to W \to Y}. \tag{7}$$

(a) Supervised learning of Y.

(b) Unsupervised learning of C.

Figure 2: Neglecting the protected attribute in supervised and unsupervised learning from a causal perspective.

Unfortunately, simply ignoring the protected attribute X during the classification process (fairness through unawareness) does not remove the direct or indirect biases. This is due to the correlation between X and W, which allows the classifier to gain knowledge about X via W. Thus, the direct bias δ_1 may still be propagated via the indirect pathway through the correlation with W. We denote the bias that is propagated via this correlation $\hat{\delta}_1$, since in practice, this indirect propagation reduces the direct bias δ_1. The total variation in this scenario is

$$\text{TV}_{x_0,x_1}(y) = \underbrace{0}_{X\to Y} + \underbrace{\hat{\delta}_1 + \hat{\delta}_2}_{X\to W\to Y} . \tag{8}$$

Hence, even though the protected attribute is not used in the classification, it could be learned indirectly via a correlation with the mediator W and hence will propagate the direct discriminatory bias.

Unsupervised Learning (Clustering). In supervised learning, all edges $X \to W$, $X \to Y$, and $W \to Y$ reflect the generating process of the data. This is in contrast to the unsupervised learning case (Figure 2b), where only the edge $X \to W$ depicts the generating process of the data and contains the unfair bias δ_2. Learning the cluster attribute C based on the input W introduces a new assignment mechanism f_C in the SCM that is under our control. Note that if the protected attribute is not used as input to f_C, there is no direct edge between X and C, which implies the conditional independence $X \perp\!\!\!\perp C \mid W$. This conditional independence guarantees the absence of the natural direct effect.

Lemma 2 (Fairness through unawareness in clustering) *If the protected attribute X is not an input of the clustering mechanism f_C, the natural direct effect is zero, i.e.,*

$$\text{NDE}_{x_0,x_1}(c_k) = 0. \tag{9}$$

Hence, in contrast to the supervised case, fairness through unawareness allows the removal of the natural direct effect in clustering. However, an unfair bias can still persist via indirect effects. In the next section, we will discuss the implications of fair clustering and show how to remove indirect discriminatory effects.

3.2. Fairness Through Balanced Clusters

Balanced clusters exist when each protected class has equal representation across the clusters (Chierichetti et al., 2017). This implies that the total variation of a balanced cluster c_k equals 0,

i.e.,

$$\text{TV}_{x_0,x_1}(c_k) = P(c_k \mid x_1) - P(c_k \mid x_0) = 0. \tag{10}$$

However, previous work has shown that optimizing the TV measure to be zero does not necessarily reduce causal measures (Nilforoshan et al., 2022). From a causal perspective, we can further decompose the total variation into the natural direct, natural indirect and experimental spurious effect, for which we reintroduce Z into the discussion.

Proposition 3 (Theorem 4.2 in Plečko and Bareinboim, 2024) *The total variation measure can be decomposed as*

$$\text{TV}_{x_0,x_1}(c_k) = \text{NDE}_{x_0,x_1}(c_k) - \text{NIE}_{x_0,x_1}(c_k) + \text{Exp-SE}_{x_0,x_1}(c_k), \tag{11}$$

where $\text{Exp-SE}_{x_0,x_1}(c_k)$ *is the experimental spurious effect, defined as*

$$\text{Exp-SE}_{x_0,x_1}(c_k) = \text{Exp-SE}_{x_1}(c_k) - \text{Exp-SE}_{x_0}(c_k) \tag{12}$$

with

$$\text{Exp-SE}_{x}(c_k) = P(c_k \mid x) - P\big((c_k)_x\big). \tag{13}$$

The experimental spurious effect measures the disparity in c_k when setting $X = x$ by intervention, compared to observing that $X = x$. It captures the extent to which spurious correlations, such as those between ZIP codes and race due to historical inequalities, distort clustering outcomes.

A direct consequence of Proposition 3 is that if the clusters are balanced, i.e., $\text{TV}_{x_0,x_1}(c_k) = 0$, then

$$\text{NDE}_{x_0,x_1}(c_k) = \text{NIE}_{x_0,x_1}(c_k) - \text{Exp-SE}_{x_0,x_1}(c_k). \tag{14}$$

Hence, enforcing the clusters to be balanced can induce a natural direct effect unless $\text{NIE}_{x_0,x_1}(c_k) - \text{Exp-SE}_{x_0,x_1}(c_k) = 0$, as illustrated in the following example.

Example 2 *Consider the case where we want to cluster data which includes the country Z, race X, and the browsing preferences W of internet users. Assume the browsing preferences are identical across X, i.e., there is no natural indirect effect* $\text{NIE}_{x_0,x_1}(c_k) = 0$*. If there is a spurious effect* $\text{Exp-SE}_{x}(c_k) \neq 0$*, then enforcing balanced clusters with* $TV_{x_0,x_1}(c_k) = 0$ *would induce a direct effect from race to cluster in order to counterbalance the spurious effect.*

As a special case, consider a scenario in which we have no spurious effect and neglect the protected attribute in the clustering. This is an interesting scenario, as it bridges the gap between causal and non-causal fairness optimizations.

Corollary 4 *If there is no natural direct effect* $\text{NDE}_{x_0,x_1}(c_k) = 0$ *because the protected attribute is neglected, and there is no spurious effect* $\text{Exp-SE}_{x_0,x_1}(c_k) = 0$*, then*

$$\text{TV}_{x_0,x_1}(c_k) = -\text{NIE}_{x_0,x_1}(c_k). \tag{15}$$

The proof is a direct consequence of Proposition 3, when $\text{NDE}_{x_0,x_1}(c_k) = \text{Exp-SE}_{x_0,x_1}(c_k) = 0$. Hence, in this particular setting, optimizing for balanced clusters is identical to minimising the natural indirect effect. This implies that under the assumption of the SFM, minimising the natural indirect effect could be an alternative to balanced clustering algorithms if the aim is to minimise the total variation.

Algorithm 1 Targeted Learning of Causally Fair Clusters

Input: A matrix of variables that we wish to cluster, an SFM mapping (X, Z, W), and a binary vector of length three $(\text{NDE}, \text{NIE}, \text{SE})$ specifying which effects should be minimised
Output: Causally fair cluster associations $\phi(X, Z, W)$

Optimal transport:
if $\text{SE} = 1$ **then**
  Transport $Z \mid x_1$ onto $Z \mid x_0$
  Denote the optimal transport map with τ^z
else
  Let τ^z be the identity map $\tau^z(Z) = Z$
end if
if $\text{NIE} = 1$ **then**
  Transport $W \mid x_1, \tau^z(Z)$ onto $W \mid x_0, Z$
  Denote the transport map with τ^w
else
  Transport $W \mid x, \tau^z(Z)$ onto $W \mid x, Z$ for $x \in \{x_0, x_1\}$
  Denote the transport map with τ^w
end if
if $\text{NDE} = 1$ **then**
  Fairness through unawareness (neglecting X in input):
  $\phi(X, Z, W) \leftarrow f_C(W, Z)$
else
  $\phi(X, Z, W) \leftarrow f_C(X, W, Z)$
end if

4. Learning Causally Fair Clusters

We define causally fair clusters as cluster assignments that minimise the following causal fairness notions: natural direct, natural indirect, and experimental spurious effect. Note that minimisation in this context refers to reducing the absolute value of these effects. In particular, we propose Algorithm 1, which allows to minimise each of the causal fairness notions separately. As part of the input, one needs to specify which of the causal fairness notions should be optimised.

4.1. Natural Direct Effect

By excluding the protected attribute from the clustering process as described in Algorithm 1, we can mitigate the natural direct effect, i.e., $\text{NDE}_{x_0,x_1}(c_k) = 0$, as shown in Lemma 2. Nevertheless, the protected attribute plays a crucial role in neutralising the natural indirect and experimental spurious effect, as we will show in the subsequent sections.

4.2. Natural Indirect Effect

The natural indirect effect from $X \to W \to C$ can be separated in the two pathways $X \to W$ and $W \to C$. The pathway $X \to W$ displays a bias that is intrinsic to the data. In contrast, the pathway $W \to C$ is learned throughout the clustering mechanism f_C and may propagate an unfair bias.

One option is to use optimal transport (Peyré and Cuturi, 2019) to adjust W such that the following condition holds (this approach was used in the context of fairness by Plečko and Meinshausen, 2020)

$$P(w \mid x_1) - P(w \mid x_0) = 0, \ \forall \ w \in W. \tag{16}$$

From Equation (16), given the SFM, we infer that the natural indirect effect is null. Since, under the assumptions of the SFM, the natural indirect effect is identifiable from observational data, the constraint $\text{NIE}_{x_0,x_1}(c_k) = 0$ can be written as

$$\sum_{z,w} P(c_k \mid x_1, z, w)[P(w \mid x_1, z) - P(w \mid x_0, z)]P(z) = 0. \tag{17}$$

In practice, this optimization may be imperfect and only yield $\|P(W \mid x_0, z) - P(W \mid x_1, z)\|_1 \leq \delta$, such that the remaining difference can propagate to the natural indirect effect. In this case, we have the following upper bound.

Lemma 5 (Natural indirect effect) *Under the assumptions of the SFM and* $\|P(W \mid x_1, z) - P(W \mid x_0, z)\|_1 \leq \delta_w$*, we have that*

$$\text{NIE}_{x_0,x_1}(c_k) \leq \| \sup_w P(c_k \mid x_0, w, Z) \cdot P(Z) \cdot \delta_w \|_1 \tag{18}$$

4.3. Experimental Spurious Effect

Analogous to the natural indirect effect, we can also adapt Z such that

$$P(z \mid x_1) - P(z \mid x_0) = 0, \ \forall \ z \in Z. \tag{19}$$

If we neglect the protected attribute in the clustering and adapt both W and Z, we can provide the following upper bound.

Lemma 6 (Experimental spurious effect) *Assume that we perform fairness through unawareness and adapt* W *such that* $P(w \mid x_1) - P(w \mid x_0) = 0$*. For* $\|P(Z \mid x_1) - P(Z \mid x_0)\|_1 \leq \delta_z$ *and under the assumption of the SFM*

$$\text{Exp-SE}_{x_0,x_1}(c_k) \leq \sup_z P(c_k \mid z) \cdot \delta_z. \tag{20}$$

As an example of the experimental spurious effect, consider a scenario where Z represents ZIP codes. In regions where historic inequalities have made ZIP codes correlated with race, clustering decisions based on Z could propagate racial biases indirectly.

In summary, the natural direct effect can be removed by ignoring the protected attribute in the clustering, and the natural indirect and experimental spurious effect can be removed by adapting W and Z, respectively (see Theorem 7). Notably, pre-processing the data is sufficient for obtaining causally fair clusters, and hence, there is no need for more involved in-processing algorithms.

Theorem 7 (Soundness of Algorithm 1) *Under the assumption of the SFM, the clustering assignments resulting from Algorithm 1 satisfy the following conditions:*

- ***NDE**: If the protected attribute X is not part of the clustering mechanism f_C, then in the infinite sample case*

$$\text{NDE}_{x_0,x_1}(c_k) = 0 \tag{21}$$

- ***NIE**: Algorithm 1 restricts the NIE as*

$$\text{NIE}_{x_0,x_1}(c_k) \leq \| \sup_w P(c_k \mid x_0, w, Z) \cdot P(Z) \cdot \delta_w \|_1, \tag{22}$$

where the variable adaption is subject to the limit $\|P(W \mid x_1, z) - P(W \mid x_0, z)\|_1 \leq \delta_w$.

- ***Exp-SE**: The Exp-SE is bounded by*

$$\text{Exp-SE}_{x_0,x_1}(c_k) \leq \sup_z P(c_k \mid z) \cdot \delta_z, \tag{23}$$

where the variable adaption is subject to the limit $\|P(z \mid x_1) - P(z \mid x_0)\|_1 \leq \delta_z$.

If δ_w and δ_z collapse to zero in the optimal transport procedure in the infinite sample case, then $\text{NDE}_{x_0,x_1}(c_k) = \text{NIE}_{x_0,x_1}(c_k) = \text{Exp-SE}_{x_0,x_1}(c_k) = 0$.

5. Experiments

To evaluate the efficiency of our proposed approach, we benchmarked it against several existing clustering methods: balanced clustering, fairness through unawareness clustering, and clustering without any adjustments for fairness, which we refer to as unadjusted clustering. Since Algorithm 1 allows one to specify which causal fairness notions are minimised, we implemented two different versions of our approach. One version minimises the NDE, NIE and Exp-SE, which we denote *causally fair (NDE+NIE+SE)*, and the other minimises the NDE and NIE only, which we denote *causally fair (NDE+NIE)*.

Our experiments were conducted on the Adult dataset from the UCI machine learning repository (Lichman et al., 2013) and the COMPAS dataset (Larson et al., 2016). The UCI Adult dataset contains demographic data from the 1994 U.S. census and is a common benchmark for income prediction. The COMPAS dataset includes defendant profiles used to assess recidivism risks. Discriminatory biases in the UCI Adult and COMPAS datasets have been well-documented in literature, making them standard benchmarks in both unsupervised and supervised learning (Chierichetti et al., 2017; Nabi and Shpitser, 2018; Chiappa, 2019; Larson et al., 2016). These datasets contain sensitive attributes such as gender and race, which were utilised to measure the fairness of the clustering algorithms. The pre-processed UCI Adult and COMPAS datasets were downloaded from the **fairadapt** package (Plečko et al., 2024) and the pre-processing procedure is described in detail in (Plečko and Meinshausen, 2020). The datasets were pre-processed by excluding features such as relationship, final weight, education, capital gain, and capital loss. Furthermore, work class, marital status, and native country were re-categorised into simplified levels, resulting in fewer distinct groups for each variable.

The SFM model assignments were chosen analogous to the work of Plečko and Bareinboim (2024), which is based on previous suggestions for the causal structure underlying these datasets (Nabi and Shpitser, 2018; Chiappa, 2019; Plečko and Meinshausen, 2020). Nevertheless, we acknowledge that alternative causal associations could be possible, and inaccuracies in the causal

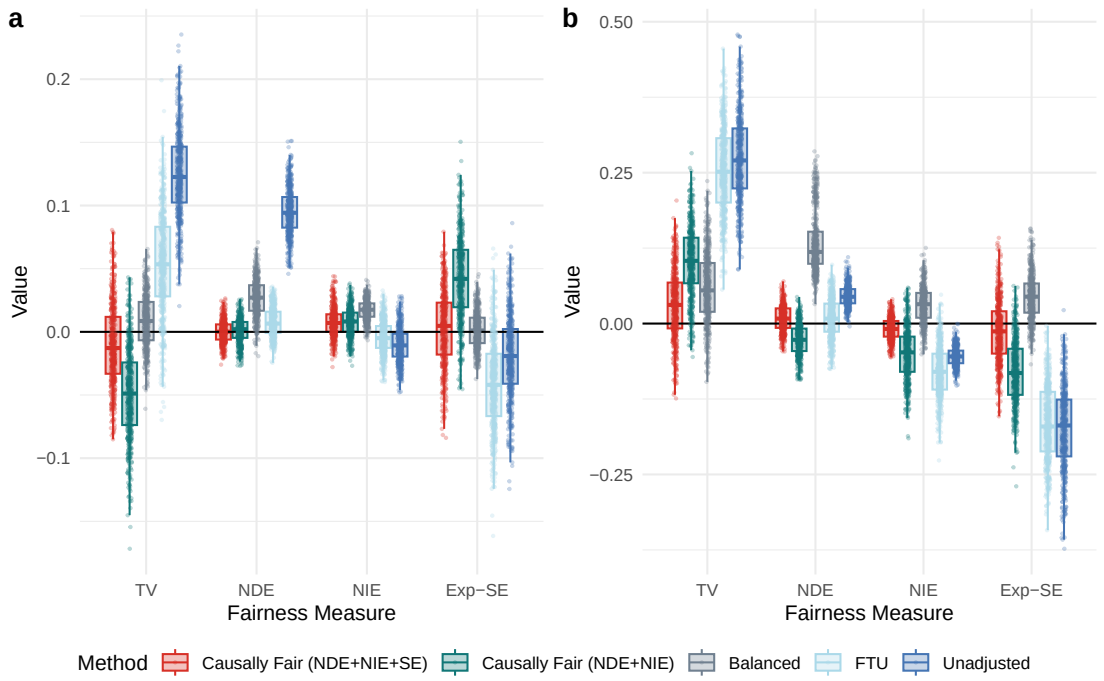

Figure 3: Comparison of fair clustering approaches on the (a) UCI Adult and (b) the COMPAS dataset. Causal Fair: Causally fair clustering, FTU: Fairness through unawareness by neglecting the protected attribute, Unadjusted: Naive clustering without adjustments; TV: total variation, NDE: natural direct effect, NIE: natural indirect effect, and Exp-SE: experimental spurious effect.

structure could affect the efficiency of our approach. For the Adult dataset within the SFM framework, we chose the confounding variables in Z to include a range of demographic variables, including age, citizenship, and economic region. The mediator variables in W were chosen to encompass employment and socio-economic factors, including salary, marital status, family size, children, education level, English level, hours worked, weeks worked, occupation, and industry. In the COMPAS dataset within the SFM framework, the confounders in Z include age. The mediator variables in W include various components regarding criminal history and legal circumstances, such as juvenile felony count, juvenile misdemeanour count, other juvenile counts, number of prior offences, degree of charge, and two-year recidivism indicator. As protected attributes, we chose sex and race in the COMPAS and UCI Adult datasets, respectively. We discuss the limitations of our simplification of complex protected attributes in our ethical statement in the Appendix.

The performance of the algorithms was evaluated based on the following metrics of fairness: total variation (TV), which is a non-causal fairness notion, and the previously introduced causal notions of fairness, including the natural direct effect (NDE), natural indirect effect (NIE), and experimental spurious effect (Exp-SE).

The bias along $X \to W$ was corrected using optimal transport using the package **fairadapt**, where quantile regression was executed via random forest quantile regression. Subsequently, the processed data was clustered using the package **clustMixType** (Szepannek, 2018). The balanced clustering was performed using the package **FairMclus** and the fairness measures were calculated using the package **faircause** (Plečko and Bareinboim, 2024). Code implementing our approach and

reproducible benchmarks are open-source and publicly available at https://github.com/cbg-ethz/fairClust. All computations were performed in **R** on a machine with a quad-core Intel core i5 (2.4 GHz) CPU and an Intel iris plus graphics 655 (1536 MB) GPU. In the estimation of the causal fairness measures shown in Figure 3, we chose 100 inner and five outer bootstrap repetitions. The inner bootstrap repetitions specify the iterations of the fitting procedure, whereas the outer bootstrap repetitions determine the number of bootstrap samples that are taken after the potential outcomes have been obtained from the estimation. This choice represented a trade-off between accuracy and computation time. A higher number of repetitions might increase the accuracy but also result in a longer computation time. We chose a binary clustering, given the high computational cost of the balanced clustering algorithm.

Figure 3 summarises the benchmark results over both datasets. The unadjusted clustering approach exhibits large discriminatory biases across all fairness measures, including the NDE. In contrast, the fairness through unawareness approach manages to minimise the NDE effectively. Nonetheless, fairness through unawareness displays noticeable effects across the NIE and Exp-SE, which aligns with our expectation as indirect effects are not controlled. The balanced clustering method, despite having a TV near zero, exhibits minor biases in both the NDE and NIE. This illustrates that achieving low TV does not guarantee complete fairness across all measures.

The first version of our causally fair clustering approach, optimising the NDE, NIE and Exp-SE, exhibits minimal biases across all fairness measures. This aligns with our expectation that optimising for causal fairness notions minimises the total variation (see Proposition 3). In contrast, the second version of our causally fair clustering approach, which only optimises the NDE and NIE, shows minimal biases across the NDE and NIE, but a large Exp-SE.

Notably, the Silhouette scores decrease from 0.34 for unadjusted clustering to 0.33 when optimizing for NDE and NIE, and further to 0.31 when also minimizing Exp-SE in the Adult dataset, illustrating the tradeoff between fairness constraints and clustering accuracy.

These results demonstrate that our novel approach enables us to precisely control which fairness notions are minimised during the clustering process. In practice, this implies we can specify whether confounding variables, such as economic region, can inform the clustering through indirect effects about protected attributes. This may allow for a more nuanced alignment of the clustering process with ethical standards or legal requirements.

6. Conclusion

We introduced a novel causally fair clustering approach as an alternative to existing fair clustering algorithms, which mostly rely on non-causal notions of fairness. Through our experimental analysis of standard datasets, we demonstrated the robustness and efficacy of our approach with respect to causal fairness notions compared to conventional strategies. In particular, our approach allows us to specify which fairness metrics should be optimised, allowing for a more nuanced and targeted optimisation of fairness in unsupervised learning.

In settings where causal relationships are clearly defined, this can allow a more thoughtful alignment with ethical principles, legal requirements, and societal needs. Minimizing the natural direct effect allows to target direct discriminatory effects, ensuring that clustering decisions are not directly influenced by protected attributes such as gender or race. Addressing the natural indirect effect mitigates systemic biases transmitted through mediating variables, such as education. Ad-

ditionally, controlling for the experimental spurious effect counteracts systemic biases encoded in confounding variables, such as ZIP codes that may indirectly reflect race.

Nevertheless, it is important to recognise that our approach requires clear understanding of the causal relationships involved, which may not always be available in practice. Incorrect causal assumptions could inadvertently lead to new biases rather than mitigating existing ones. Therefore, non-causal fairness approaches may be preferable when causal relationships are poorly understood, data limitations prevent the construction of robust causal models, or non-causal fairness constraints suffice to meet regulatory or ethical guidelines.

In this work, we optimised for causal fairness notions that investigate both direct and indirect discriminatory effects under the assumptions of the standard fairness model. Our approach lays a practical foundation for causally fair unsupervised learning, which can be extended by exploring other causal fairness notions (Barocas et al., 2017; Castelnovo et al., 2022) in future research.

Acknowledgements

The authors are grateful to acknowledge funding support for this work from the two Cantons of Basel through project grant PMB-02-18 granted by the ETH Zurich (to JK).

References

Mohsen Abbasi, Aditya Bhaskara, and Suresh Venkatasubramanian. Fair clustering via equitable group representations. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 504–514, 2021.

Sara Ahmadian, Alessandro Epasto, Ravi Kumar, and Mohammad Mahdian. Clustering without over-representation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 267–275, 2019.

Nihesh Anderson, Suman K Bera, Syamantak Das, and Yang Liu. Distributional individual fairness in clustering. *arXiv preprint arXiv:2006.12589*, 2020.

Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. Scalable fair clustering. In *International Conference on Machine Learning*, pages 405–413. PMLR, 2019.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California law review*, pages 671–732, 2016.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NeurIPS tutorial*, 1:2017, 2017.

Fritz Bayer, Marco Roncador, Giusi Moffa, Kiyomi Morita, Koichi Takahashi, Niko Beerenwinkel, and Jack Kuipers. Network-based clustering unveils interconnected landscapes of genomic and clinical features across myeloid malignancies. *bioRxiv*, pages 2023–10, 2023.

Suman Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. Fair algorithms for clustering. *Advances in Neural Information Processing Systems*, 32, 2019.

Ioana O Bercea, Martin Groß, Samir Khuller, Aounon Kumar, Clemens Rösner, Daniel R Schmidt, and Melanie Schmidt. On the cost of essentially fair clusterings. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

Brian Brubach, Darshan Chakrabarti, John P Dickerson, Aravind Srinivasan, and Leonidas Tsepenekas. Fairness, semi-supervised learning, and more: A general framework for clustering with stochastic pairwise constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6822–6830, 2021.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):4209, 2022.

Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. Proportionally fair clustering. In *International Conference on Machine Learning*, pages 1032–1041. PMLR, 2019.

Anshuman Chhabra, Karina Masalkovaitė, and Prasant Mohapatra. An overview of fairness in clustering. *IEEE Access*, 9:130698–130720, 2021.

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

Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. *Advances in neural information processing systems*, 30, 2017.

Mehrdad Ghadiri, Samira Samadi, and Santosh Vempala. Socially fair k-means clustering. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 438–448, 2021.

Dishant Goyal and Ragesh Jaiswal. Tight fpt approximation for socially fair clustering. *Information Processing Letters*, 182:106383, 2023.

Matthew Jones, Huy Nguyen, and Thy Nguyen. Fair k-centers via maximum matching. In *International Conference on Machine Learning*, pages 4940–4949. PMLR, 2020.

Christopher Jung, Sampath Kannan, and Neil Lutz. A center in your neighborhood: Fairness in facility location. *arXiv preprint arXiv:1908.09041*, 2019.

Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):881–892, 2002.

Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. Fair k-center clustering for data summarization. In *International Conference on Machine Learning*, pages 3448–3457. PMLR, 2019a.

Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, and Jamie Morgenstern. Guarantees for spectral clustering with fairness constraints. In *International Conference on Machine Learning*, pages 3458–3467. PMLR, 2019b.

Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. A notion of individual fairness for clustering. *arXiv preprint arXiv:2006.04960*, 2020.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9(1):3–3, 2016.

Moshe Lichman et al. UCI machine learning repository, 2013.

Sepideh Mahabadi and Ali Vakilian. Individual fairness for k-clustering. In *International Conference on Machine Learning*, pages 6586–6596. PMLR, 2020.

Yury Makarychev and Ali Vakilian. Approximation algorithms for socially fair clustering. In *Conference on Learning Theory*, pages 3246–3264. PMLR, 2021.

Evi Micha and Nisarg Shah. Proportionally fair clustering revisited. In *47th International Colloquium on Automata, Languages, and Programming (ICALP 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Maryam Negahbani and Deeparnab Chakrabarty. Better algorithms for individually fair k-clustering. *Advances in Neural Information Processing Systems*, 34:13340–13351, 2021.

Hamed Nilforoshan, Johann D Gaebler, Ravi Shroff, and Sharad Goel. Causal conceptions of fairness and their consequences. In *International Conference on Machine Learning*, pages 16848–16887. PMLR, 2022.

Judea Pearl. *Causality*. Cambridge university press, 2000.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, page 411–420, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Drago Plečko and Elias Bareinboim. Causal fairness analysis: a causal toolkit for fair machine learning. *Foundations and Trends® in Machine Learning*, 17(3):304–589, 2024.

Drago Plečko and Nicolai Meinshausen. Fair data adaptation with quantile preservation. *The Journal of Machine Learning Research*, 21(1):9776–9819, 2020.

Drago Plečko, Nicolas Bennett, and Nicolai Meinshausen. fairadapt: Causal reasoning for fair data pre-processing. *Journal of Statistical Software*, 110:1–35, 2024.

Clemens Rösner and Melanie Schmidt. Privacy preserving clustering with constraints. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

Gero Szepannek. ClustMixType: User-friendly clustering of mixed-type data in R. *The R Journal*, 10(2):200, 2018.

Ali Vakilian and Mustafa Yalciner. Improved approximation algorithms for individually fair clustering. In *International Conference on Artificial Intelligence and Statistics*, pages 8758–8779. PMLR, 2022.

Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. Pc-fairness: A unified framework for measuring causality-based fairness. *Advances in neural information processing systems*, 32, 2019.

Junzhe Zhang and Elias Bareinboim. Equality of opportunity in classification: A causal approach. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3671–3681, Montreal, Canada, 2018a. Curran Associates, Inc.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b.

Appendix A. Proofs

Proof [Proof of Lemma 2] Neglecting the protected attribute in the clustering implies the conditional independence $X \perp\!\!\!\perp C \mid W$ and induces the equality

$$P(c_k \mid w) = P(c_k \mid x_1, w) = P(c_k \mid x_0, w).$$

Inserting this equality into the identification expression for the NDE under the assumptions of the SFM proves the proposition

$$\begin{aligned}
\text{NDE}_{x_0,x_1}(c_k) &= P\big((c_k)_{x_1,W_{x_0}}\big) - P\big((c_k)_{x_0}\big) \\
&= \sum_w [P(c_k \mid x_1, w) - P(c_k \mid x_0, w)] \\
&\quad \cdot P(w \mid x_1) \\
&= \sum_w [P(c_k \mid w) - P(c_k \mid w)] P(w \mid x_1) \\
&= 0.
\end{aligned}$$

■

Proof [Proof of Lemma 5] Under the condition that $\|P(w \mid x_1, z) - P(w \mid x_0, z)\|_1 \leq \delta_w$, we can proof Proposition 5 following Hölder's inequality.

$$\begin{aligned}
\text{NIE}_{x_0,x_1}(c_k) &= P\big((c_k)_{x_1,W_{x_0}}\big) - P\big((c_k)_{W_{x_1}}\big) \\
&= \sum_z P(z) \cdot \|P(c_k \mid x_0, z, W) \\
&\quad [P(W \mid x_1, z) - P(W \mid x_0, z)]\|_1 \\
&\leq \sum_z P(z) \cdot \sup_w P(c_k \mid x_0, w) \\
&\quad \cdot \|P(W \mid x_1, z) - P(W \mid x_0, z)\|_1 \\
&\leq \sum_z P(z) \cdot \sup_w P(c_k \mid x_0, w) \cdot \delta_w \\
&= \|\sup_w P(c_k \mid x_0, w, Z) \cdot P(Z) \cdot \delta_w\|_1
\end{aligned}$$

■

Proof [Proof of Lemma 6]

Under the SFM, the experimental spurious effect can be written as

$$\begin{aligned}
\text{Exp-SE}_{x_0,x_1}(c_k) =& \text{Exp-SE}_{x_1}(c_k) - \text{Exp-SE}_{x_0}(c_k) \\
=& P(c_k \mid x_1) - P\big((c_k)_{x_1}\big) \\
& - P(c_k \mid x_0) - P\big((c_k)_{x_0}\big) \\
=& \sum_z \big(P(c_k \mid x_1, z)[P(z) - P(z \mid x_1)] \\
& - P(c_k \mid x_0, z)[P(z) - P(z \mid x_0)]\big),
\end{aligned}$$

where we inserted

$$\text{Exp-SE}_x(c_k) = P(c_k \mid x) - P\big((c_k)_x\big).$$

Assuming that we perform fairness through unawareness and adapt W such that $P(w \mid x_1) - P(w \mid x_0) = 0$, we can conclude that $P(c_k \mid x, z) = P(c_k \mid z)$. Further, assuming $\|P(z \mid x_1) - P(z \mid x_0)\|_1 \leq \delta_z$, we can proof Proposition 6 following Hölder's inequality

$$\begin{aligned}
\text{Exp-SE}_{x_0,x_1}(c_k) =& \sum_z \big(P(c_k \mid x_1, z)[P(z) - P(z \mid x_1)] \\
&\quad - P(c_k \mid x_0, z)[P(z) - P(z \mid x_0)]\big) \\
=& \sum_z \big(P(c_k \mid z)[P(z) - P(z \mid x_1)] \\
&\quad - P(c_k \mid z)[P(z) - P(z \mid x_0)]\big) \\
=& \sum_z P(c_k \mid z)\big[P(z \mid x_0) - P(z \mid x_1)\big] \\
=& \|P(c_k \mid Z)\big[P(Z \mid x_0) - P(Z \mid x_1)\big]\|_1 \\
\leq& \sup_z P(c_k \mid x_0, z) \\
&\quad \cdot \|P(Z \mid x_1) - P(Z \mid x_0)\|_1 \\
\leq& \sup_z P(c_k \mid z) \cdot \delta_z
\end{aligned}$$

■

Proof [Proof of Theorem 7] Assuming the SFM, we proof Theorem 7 by applying Lemma 2, Lemma 5, and Lemma 6. First, Lemma 2 proofs that the $\text{NDE}_{x_0,x_1}(c_k) = 0$ in the infinite sample case. Further, Lemma 5 proofs that the upper bound on the NIE is

$$\text{NIE}_{x_0,x_1}(c_k) \leq \|\sup_w P(c_k \mid x_0, w, Z) \cdot P(Z) \cdot \delta_w\|_1, \tag{24}$$

where the variable adaption is subject to the limit $\|P(W \mid x_1, z) - P(W \mid x_0, z)\|_1 \leq \delta_w$. Finally, Lemma 6 proofs the upper bound on the Exp-SE is

$$\text{Exp-SE}_{x_0,x_1}(c_k) \leq \sup_z P(c_k \mid z) \cdot \delta_z, \tag{25}$$

where the variable adaption is subject to the limit $\|P(z \mid x_1) - P(z \mid x_0)\|_1 \leq \delta_z$.

If $\delta_w = \delta_z = 0$ in the optimal transport procedure in the infinite sample case, then $\text{NDE}_{x_0,x_1}(c_k) = \text{NIE}_{x_0,x_1}(c_k) = \text{Exp-SE}_{x_0,x_1}(c_k) = 0$. ■

Appendix B. Ethical Statement

We would like to point out the following ethical considerations:

Simplification of Complex Attributes: We recognize and affirm the complexity of sex and gender, which can encompass a broad spectrum of identities and expressions. However, due to the

constraints of the UCI Adult and COMPAS datasets, we are utilizing a binary representation of sex. This simplification is not intended to overlook or invalidate the diverse realities of sex and gender but is a practical limitation of the analysed data.

Algorithmic Fairness: Our research aims to enhance fairness in unsupervised learning by addressing both direct and indirect discriminatory effects, but we acknowledge the inherent challenges in capturing all facets of fairness and bias.

Causal Relationships: Our approach requires a clear understanding of the causal relationships involved. In practice, these relationships may not always be known, and erroneous assumptions could lead to unintended consequences.

Wider Societal Implications: We recognize that our work may influence decision-making processes in various domains, such as employment, healthcare, or finance.