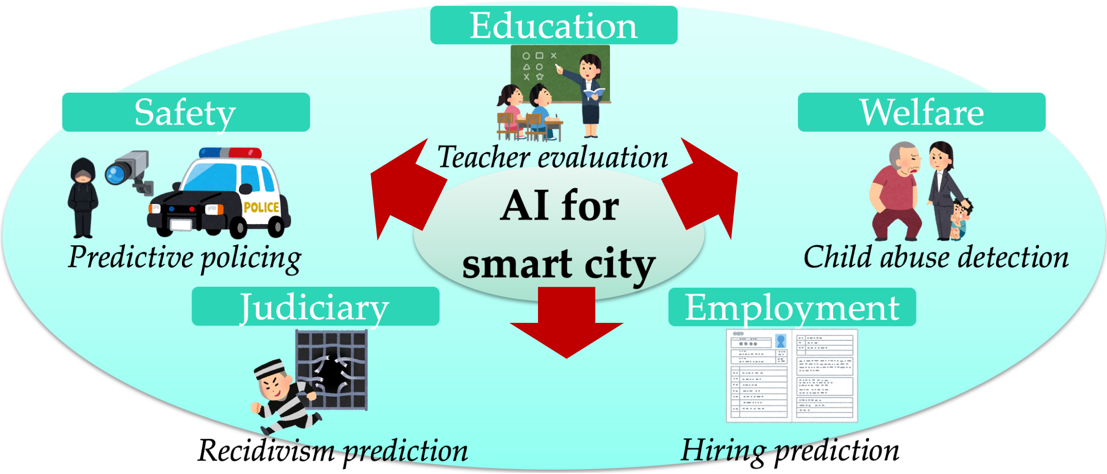
Toward Interventionally Fair Decision-Making under Partial Causal Graph Knowledge

NTT Communication Science Laboratories
Yoichi Chikahara

2026/08/08

Algorithmic Decisions via ML Prediction

NTT



Input Training data about individuals

A Gender (Sensitive)	Q Skill	D Department	M Physical strength	Y Decision
Female	A	Science	B	Accept
Male	B	Economics	B	Accept

⋮

Output Machine learning (ML) model

$$\hat{Y} = h_{\theta}(A, Q, D, M)$$

Prediction

ML model
with parameter θ

Legal Requirements: *Fairness*

Toward Society Shaped by Algorithmic Decision-Making

- ML predictions must be **fair** with respect to *sensitive attributes* (e.g., race, gender)
- Discrimination is often legally assessed based on **causality**:

An Age Discrimination Case [U.S. Supreme Court (2009)]

Plaintiff must show that **age was the *but-for cause* of the employment decision**

Are there any *causal effects*?

Age → ? → Promotion

Today's topics

1. Achieve fairness using a cluster causal graph [Chikahara. **UAI2026**]

Fairness under Graph Uncertainty: Achieving Interventional Fairness with Partially Known Causal Graphs over Clusters of Variables

Yoichi Chikahara[1]

2. Latent discrimination discovery via mean and variance causal graph discovery [Chikahara. **KDD2026**] *(if time permits)*

Moment Matters: Mean and Variance Causal Graph Discovery from Heteroscedastic Observational Data

Yoichi Chikahara

Today's topics

1. Achieve fairness using a cluster causal graph [Chikahara. UAI2026]

Fairness under Graph Uncertainty: Achieving Interventional Fairness with Partially Known Causal Graphs over Clusters of Variables

Yoichi Chikahara[1]

2. Latent discrimination discovery via mean and variance causal graph discovery [Chikahara. **KDD2026**] *(if time permits)*

Moment Matters: Mean and Variance Causal Graph Discovery from Heteroscedastic Observational Data

Yoichi Chikahara

Fair & Accurate ML Prediction via Causal Inference

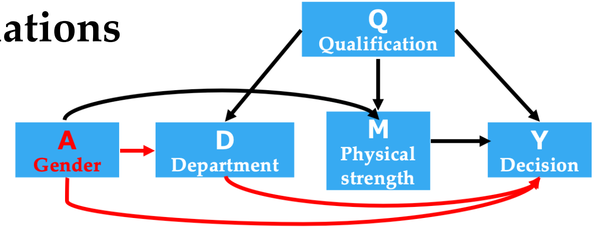
Input *Incomplete* training data

Sensitive feature

A Gender	Q Qualification	D Department	M Physical strength	Y Decision
Female	A	Economics	B	Accept
Male	B	Literature	B	Accept
Male	C	Science	C	Reject

Infer causal relations

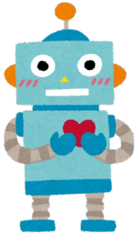
Causal graph



Learn a fair & accurate ML model

Output Fair and accurate ML model

$\hat{Y} = f_\theta(A, Q, D, M)$



Learn model parameters θ

$$\min_\theta \underbrace{\frac{1}{n}\sum_{i=1}^{n}\ell(\boldsymbol{y}_i, h_\theta(\boldsymbol{x}_i))}_{\text{prediction loss}} + \underbrace{\lambda g_\theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)}_{\text{penalty for unfairness}},$$

prediction loss **penalty for unfairness**

Fair & Accurate ML Prediction via Causal Inference

NTT

Accurately inferring a causal graph is challenging ☹

- We are NOT interested in the causal graph itself!
- We just want to ensure fairness

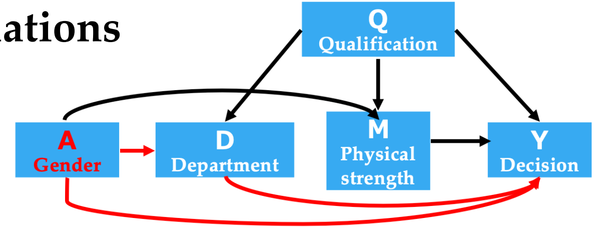
Input *Incomplete* training data

Sensitive feature

A Gender	Q Qualification	D Department	M Physical strength	Y Decision
Female	A	Economics	B	Accept
Male	B	Literature	B	Accept
Male	C	Science	C	Reject

Infer causal relations

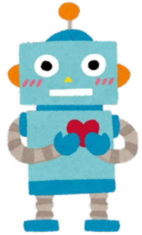
Causal graph



Learn a fair & accurate ML model

Output Fair and accurate ML model

$$\hat{Y} = f_\theta(A, Q, D, M)$$



Learn model parameters θ

$$\min_\theta \quad \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{y}_i, h_\theta(\boldsymbol{x}_i)) + \lambda g_\theta(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n),$$

prediction loss **penalty for unfairness**

Q: Can we ensure fairness using a cluster causal graph?

NTT

Input Training data

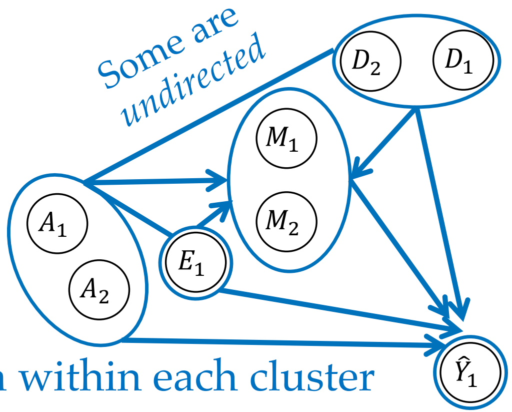
Observational data

User-specified cluster partition

$\boldsymbol{A} = [A_1, A_2], \boldsymbol{E} = [E_1]$
$\boldsymbol{M} = [M_1, M_2], \dots, \boldsymbol{\hat{Y}} = [\hat{Y}_1]$

Conditional Independence (CI) tests over clusters of variables

Cluster CPDAG
[Anand+; NeurIPS2025]



Graph within each cluster is **unknown**

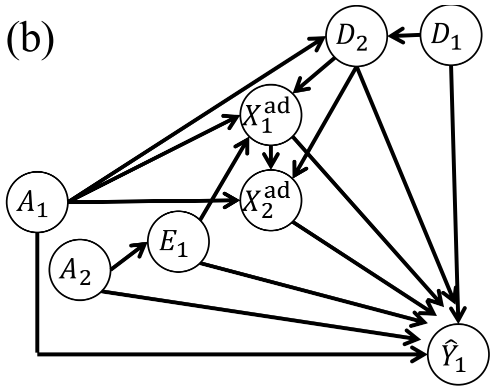
Cluster CPDAG inference is **much easier** than variable-level one ☺

- Graph size is reduced, so the number of required CI tests is much smaller

Example scenario [Chikahara+; AISTATS2021, DMKD2023]:

Hiring decisions for physically demanding jobs

Causal graph

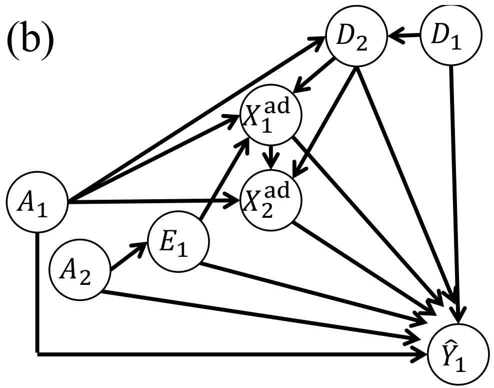


Indirect effects are **not** discriminatory

A: gender & race (*sensitive features*)
$\boldsymbol{X}^{\mathrm{ad}}$: physical test score (*admissible features*)
D: medical test results
E: educational background
$\hat{\boldsymbol{Y}}$: prediction

Example scenario [Chikahara+; AISTATS2021, DMKD2023]:
Hiring decisions for physically demanding jobs

Causal graph



A: gender & race (*sensitive features*)
$\boldsymbol{X}^{\text{ad}}$: physical test score (*admissible features*)
D: medical test results
E: educational background
$\hat{\boldsymbol{Y}}$: prediction

A structural causal model (SCM)

describes data-generating processes

E.g.,

$$X_1^{\text{ad}} = \underbrace{f_{X_1^{\text{ad}}}}_{\text{deterministic (unknown)}}(A_1, D_2, E_1, \underbrace{U_{X_1^{\text{ad}}}}_{\text{exogenous noise}});$$

$$\hat{\boldsymbol{Y}} = \underbrace{h_\theta}_{\text{predictive model (known)}}(\boldsymbol{X}, \underbrace{\mathrm{U}_{\hat{\boldsymbol{Y}}}}_{\text{randomness in probabilistic model}});$$

Interventional fairness [Salimi+; SDM2019]

1. **Interventions**: Replacing structural equations

E.g.,

$$X_1^{\mathrm{ad}} = \underline{f_{X_1^{\mathrm{ad}}}}(A_1, D_2, E_1, \underline{U_{X_1^{\mathrm{ad}}}})$$

deterministic (unknown)

exogenous noise

Intervention $do(X_1^{\mathrm{ad}} = x_1^{\mathrm{ad}})$

$$X_1^{\mathrm{ad}} = x_1^{\mathrm{ad}}$$

Values are forced to be constant x_1^{ad}

Interventional fairness [Salimi+; SDM2019]

2. Interventional fairness:

Definition 2.1 (Salimi et al. [2019]). Prediction $\hat{Y}$ is *interventionally fair* with respect to sensitive features $\boldsymbol{A}$, if

$$\begin{aligned} &\mathrm{P}(\hat{\boldsymbol{Y}} = \hat{\boldsymbol{y}} \mid do(\boldsymbol{A} = \boldsymbol{a}), do(\boldsymbol{X}^{\mathrm{ad}} = \boldsymbol{x}^{\mathrm{ad}})) \\ =&\mathrm{P}(\hat{\boldsymbol{Y}} = \hat{\boldsymbol{y}} \mid do(\boldsymbol{A} = \boldsymbol{a}'), do(\boldsymbol{X}^{\mathrm{ad}} = \boldsymbol{x}^{\mathrm{ad}})), \end{aligned} \tag{4}$$

holds for all possible prediction values $\hat{\boldsymbol{y}}$, for any discrete values $\boldsymbol{a}, \boldsymbol{a}'$ of sensitive features $\boldsymbol{A}$ and for any discrete values $\boldsymbol{x}^{\mathrm{ad}}$ of admissible features $\boldsymbol{X}^{\mathrm{ad}}$.

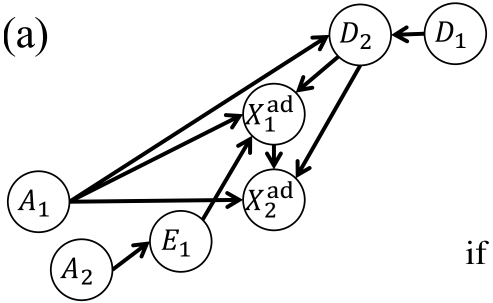
Changing sensitive attributes
do not change the distributions of predictions

Cluster DAGs [Anand+; AAAI2023]

- Group variables according to a user-defined partition

Variable-level DAG

(a)



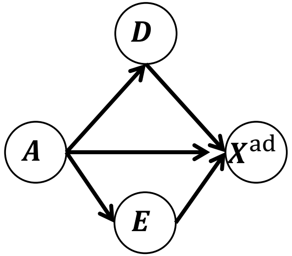
Definition

$$\boldsymbol{C}_i \rightarrow \boldsymbol{C}_j$$

if some $V_i \in \boldsymbol{C}_i$ and $V_j \in \boldsymbol{C}_j$ have edge $V_i \rightarrow V_j$ in G_{v}.

Cluster DAG

(d)

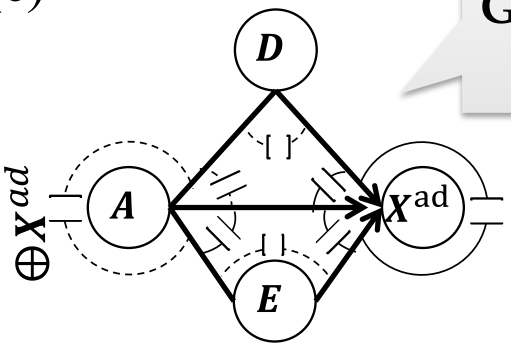


Graph structure **within** cluster is ***unknown***

Cluster CPDAGs [Anand+; NeurIPS2025]

- A cluster CPDAG represents a **cluster Markov Equivalence Class (MEC)**
- **Cluster MEC**: a class of cluster DAGs with identical CI relations among clusters

(e)

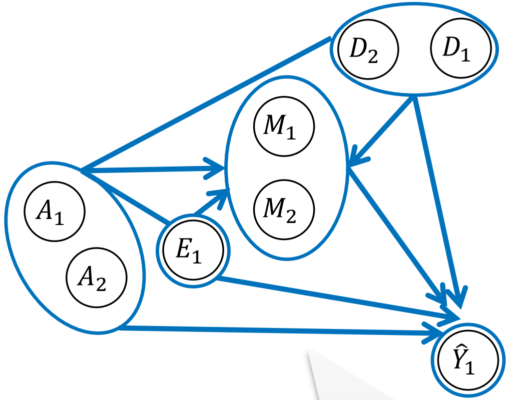


CI relations vary depending on inner graph structures.

Graphical objects (arcs over each triplet & marks) **are added to explicitly represent CI relations.**

Q: Can we estimate the distance between interventional distributions using cluster CPDAGs?

Cluster CPDAG
[Anand+; NeurIPS2025]



Graph structure **within** cluster is *unknown*

Distance between interventional distributions

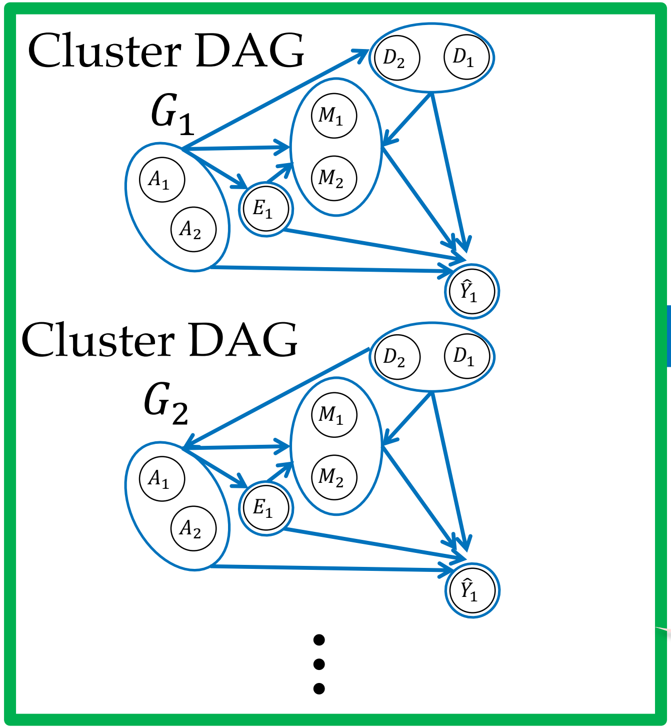
$$\mathrm{P}(\hat{\boldsymbol{Y}} = \hat{\boldsymbol{y}} \mid do(\boldsymbol{A} = \boldsymbol{a}), do(\boldsymbol{X}^{\mathrm{ad}} = \boldsymbol{x}^{\mathrm{ad}}))$$

Unfairness of predictions $\hat{Y}$

$$\mathrm{P}(\hat{\boldsymbol{Y}} = \hat{\boldsymbol{y}} \mid do(\boldsymbol{A} = \boldsymbol{a}'), do(\boldsymbol{X}^{\mathrm{ad}} = \boldsymbol{x}^{\mathrm{ad}}))$$

A: We can estimate the *maximum* distance over possible cluster DAG structures in the cluster MEC!

Cluster MEC



Distributional distances

Distance estimated with G_1

Distance estimated with G_2

⋮

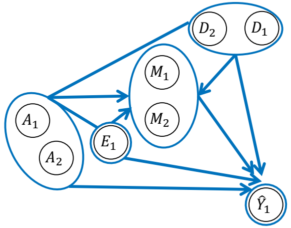
Maximum

Worst-case unfairness under graph uncertainty

Q: Can we obtain the maximum without performing DAG enumeration?

Idea: Back-door adjustment set enumeration

NTT



Cluster MEC

Adjustment candidate sets

$\boldsymbol{Z}^1$

$\boldsymbol{Z}^2$

Distributional distances

Distance estimated with $\boldsymbol{Z}^1$

Distance estimated with $\boldsymbol{Z}^2$

⋮

Maximum

Worst-case unfairness

Back-door adjustment

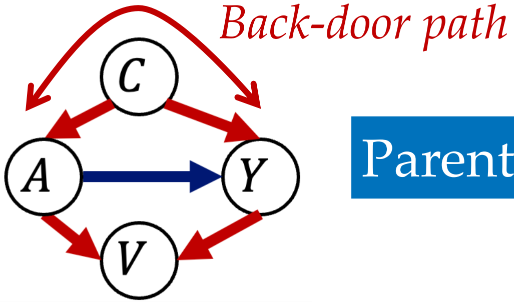
- Adjustment set $\boldsymbol{Z}$: Identifying an interventional distribution via marginalization

$$
\begin{aligned}
&\mathrm{P}(\hat{\boldsymbol{Y}} = \hat{\boldsymbol{y}} \mid do(\boldsymbol{A} = \boldsymbol{a}), do(\boldsymbol{X}^{\mathrm{ad}} = \boldsymbol{x}^{\mathrm{ad}})) \\
&= \mathbb{E}_{\boldsymbol{Z}, \boldsymbol{X}^{\mathrm{re}} | \boldsymbol{a}, \boldsymbol{x}^{\mathrm{ad}}} \left[\mathrm{P}(\hat{\boldsymbol{Y}} = \hat{\boldsymbol{y}} | \boldsymbol{A} = \boldsymbol{a}, \boldsymbol{X}^{\mathrm{ad}} = \boldsymbol{x}^{\mathrm{ad}}, \boldsymbol{Z}, \boldsymbol{X}^{\mathrm{re}}) \right]
\end{aligned}
$$

※ $\boldsymbol{X}^{\mathrm{re}} = \boldsymbol{X} \backslash \{\boldsymbol{A}, \boldsymbol{X}^{\mathrm{ad}}, \boldsymbol{Z}\}$: remaining variables

- Parent set is a common choice of adjustment set:

Example:



Parent of A → $\boldsymbol{Z} = \{C\}$

Proposed penalty function

Adjustment candidate sets

Kernel MMDs

Parent enumeration + Cluster addition for cluster CPDAGs

Cluster MEC

$\boldsymbol{Z}^1$

$\boldsymbol{Z}^2$

⋮

MMD estimation via IPW

MMD estimated with $\boldsymbol{Z}^1$

MMD estimated with $\boldsymbol{Z}^2$

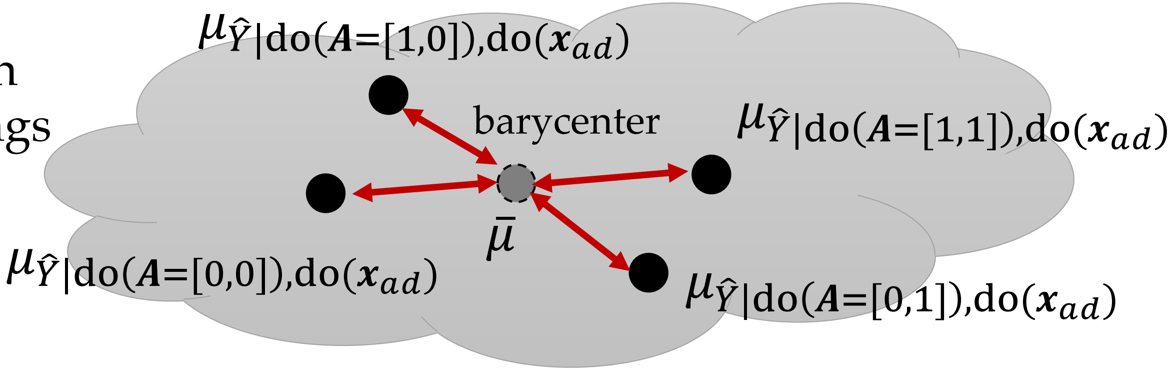
⋮

Maximum

Worst-case unfairness

Proposed penalty function

Kernel MMD:
Distance between kernel embeddings



$$g_\theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \max_{m=1,\ldots,M} \sum_{\boldsymbol{x}^{\mathrm{ad}}} \sum_{\boldsymbol{a}} \| \hat{\mu}_{\hat{\boldsymbol{Y}}|do(\boldsymbol{a}),do(\boldsymbol{x}^{\mathrm{ad}})} - \underbrace{\hat{\bar{\mu}}_{\hat{\boldsymbol{Y}}|do(\boldsymbol{x}^{\mathrm{ad}})}}_{\text{barycenter}} \|_2^2,$$

$$\hat{\mu}_{\hat{\boldsymbol{Y}}|do(\boldsymbol{a}),do(\boldsymbol{x}^{\mathrm{ad}})} = \frac{1}{n}\sum_{i=1}^{n} w_i \phi_{\mathrm{RFF}}(\hat{\boldsymbol{y}}_i) \quad \text{where} \quad w_i = \frac{\mathbf{I}(\boldsymbol{A}_i = \boldsymbol{a}, \boldsymbol{X}_i^{\mathrm{ad}} = \boldsymbol{x}^{\mathrm{ad}})}{\pi_{\boldsymbol{a},\boldsymbol{x}^{\mathrm{ad}}}(\boldsymbol{z}_i^m)}, \text{ (IPW weight)}$$

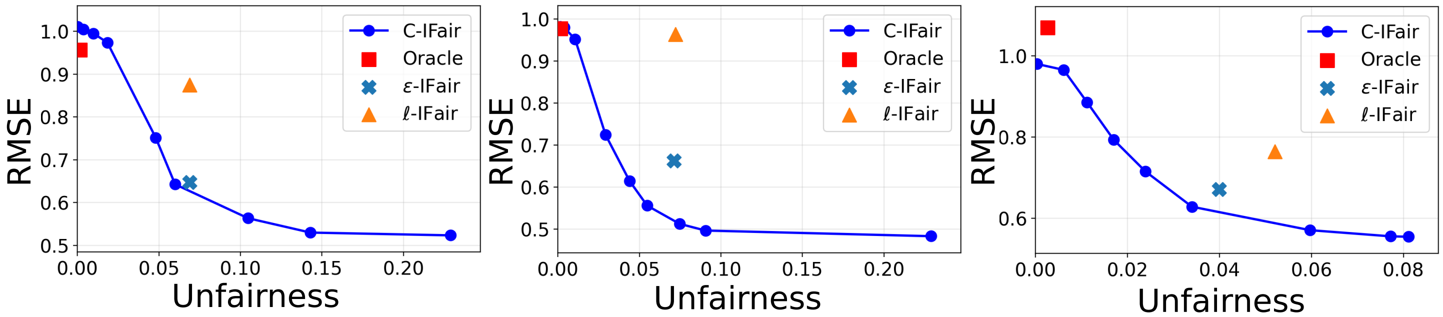
Weight depends on each set $\boldsymbol{Z}^m$

Empirical Results

Empirically, the proposed method performs best ☺

	ADULT		GERMAN		OULAD	
	AUC↑	UNFAIRNESS↓	AUC↑	UNFAIRNESS↓	AUC↑	UNFAIRNESS↓
ORACLE	0.709 ± 0.009	0.000 ± 0.000	0.582 ± 0.004	0.000 ± 0.000	0.628 ± 0.015	0.000 ± 0.000
FULL	0.874 ± 0.004	0.030 ± 0.024	0.762 ± 0.060	0.173 ± 0.011	0.697 ± 0.024	0.061 ± 0.001
UNAWARE	0.805 ± 0.049	0.018 ± 0.014	0.695 ± 0.061	0.101 ± 0.013	0.654 ± 0.019	0.004 ± 0.000
NO-DESCS	0.815 ± 0.007	0.026 ± 0.022	0.726 ± 0.059	0.085 ± 0.008	0.654 ± 0.020	0.004 ± 0.000
ϵ-IFAIR	0.812 ± 0.006	0.015 ± 0.012	0.756 ± 0.008	0.082 ± 0.000	0.641 ± 0.005	0.005 ± 0.002
ℓ-IFAIR	0.764 ± 0.002	0.015 ± 0.008	0.750 ± 0.035	0.151 ± 0.011	0.639 ± 0.011	0.003 ± 0.001
C-IFAIR	$\mathbf{0.829 \pm 0.021}$	$\mathbf{0.014 \pm 0.010}$	$\mathbf{0.760 \pm 0.061}$	$\mathbf{0.065 \pm 0.008}$	$\mathbf{0.660 \pm 0.018}$	$\mathbf{0.001 \pm 0.000}$

(a) $d = 5$ $(d_v = 15)$ (b) $d = 10$ $(d_v = 30)$ (c) $d = 15$ $(d_v = 45)$



https://arxiv.org/abs/2602.23611

This work was supported by JST ACT-X

Today’s topics

1. Achieve fairness using a cluster causal graph [Chikahara. **UAI2026**]

Fairness under Graph Uncertainty: Achieving Interventional Fairness with Partially Known Causal Graphs over Clusters of Variables

Yoichi Chikahara[1]

2. Latent discrimination discovery via mean and variance causal graph discovery [Chikahara. KDD2026] ***(if time permits)***

Moment Matters: Mean and Variance Causal Graph Discovery from Heteroscedastic Observational Data

Yoichi Chikahara

Q: Can we help discover *latent* sensitive attributes?

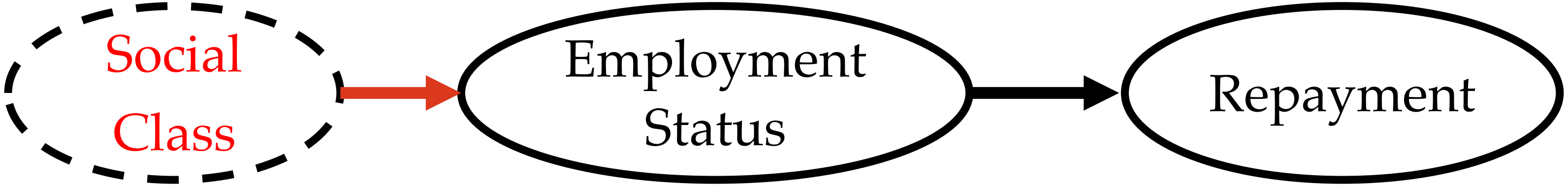
NTT

***latent* sensitive attributes**

- <u>Key idea</u>: Features that affect the <u>**(conditional) variance**</u> may cause *latent* statistical discrimination

<u>**Example:**</u>

Affects the variance
(hard to detect)

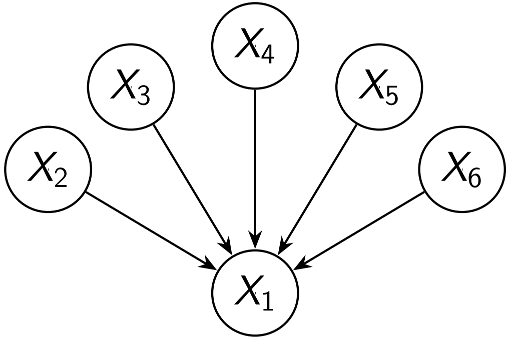


Unobserved

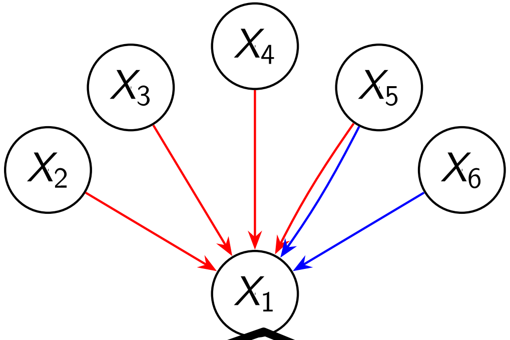
Legally acceptable for now,
but potentially discriminatory in the future.

Moment-driven causal discovery

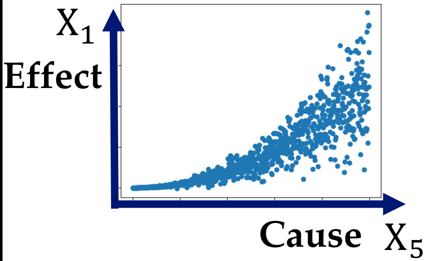
Standard causal graph



Mean/Variance Causal Graphs



Heteroscedastic data:



Defined by extending identifiable SCM class called heteroscedastic noise models (HNMs)

- **Mean-Variance HNMs** [Chikahara; KDD2026]:

$$X = f_X\left(\mathrm{pa}^{\mathrm{M}}(X)\right) + g_X(\mathrm{pa}^{\mathrm{V}}(X))\, U_X$$

Take-home messages

1. Cluster causal graphs enable fair and accurate decision-making under graph uncertainty.

2. Mean–variance causal discovery can help identify hidden sources of discrimination through interpretable causal graph representations.

Thank you!