LOCAL CAUSAL DISCOVERY FOR STRUCTURAL EVIDENCE OF DIRECT DISCRIMINATION

JACQUELINE MAASCH[1], KYRA GAN[1], VIOLET CHEN[2], AGNI ORFANOUDAKI[3], NIL-JANA AKPINAR[4*], FEI WANG[5] // [1]*Cornell Tech*, [2]*Stevens Institute of Technology*, [3]*University of Oxford*, [4]*Amazon AWS*, [5]*Weill Cornell* ([*]Work done outside of Amazon)

Efficient graph learning enables

causal fairness analysis

in complex decision systems.

DETECTING DIRECT DISCRIMINATION == CAUSAL PARENT DISCOVERY

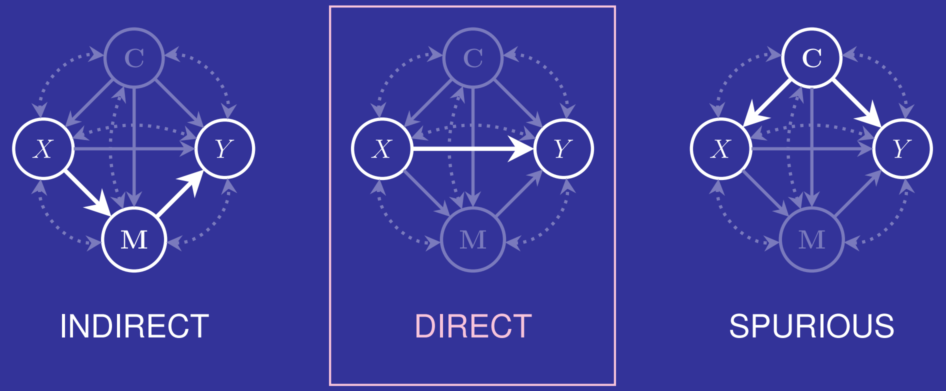

Fig. 1: The *standard fairness model* (SFM) with protected attribute X, outcome Y, confounders $\mathbf{C}$, and mediators $\mathbb{M}$ [1]. Directed edges denote active paths. Bidirected edges denote latent confounding. **This work identifies direct mechanisms of unfairness in a data-driven way.**

LD3: CAUSAL PARENT DISCOVERY FOR FAIRNESS ANALYSIS

- **APPROACH.** We introduce LD3, a constraint-based discovery method that leverages the **causal partition taxonomy** proposed in [2] to label variables by their causal relation to the protected attribute X and outcome Y (Fig. 2), rather than learning the full graph. We assume that Y has no observed descendants and no unobserved parents (other latent variables are permitted).

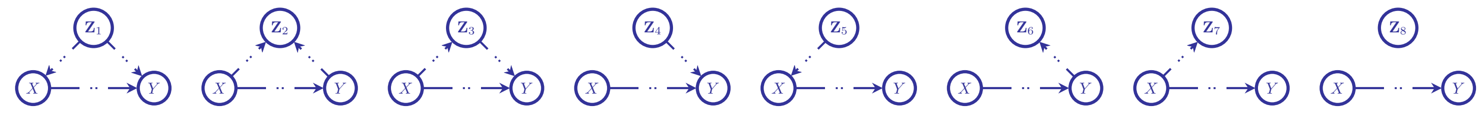

Fig. 2: The nodes of any DAG can be uniquely partitioned into 8 disjoint subsets defined by the paths shared with a pair $\{X, Y\}$ [2]. This applies to DAGs of any size; triple DAGs are for illustration only. Partition $\mathbf{Z}_1$ generalizes the *confounder*, $\mathbf{Z}_2$ the *collider*, $\mathbf{Z}_3$ the *mediator*, etc.

- **COMPLEXITY.** LD3 discovers $parents(Y)$ in a **linear number of conditional independence tests** w.r.t. variable set size.
- **FAIRNESS CRITERIA.** LD3 results directly evaluate the **SDC** and can be used as a valid adjustment set for the **WCDE**:

Definition 1 (Structural direct criterion (SDC), Plečko and Bareinboim 2024)**.** A structural causal model is fair w.r.t. direct discrimination if and only if the following evaluates to 0:

$$\mathrm{SDC} = \mathbf{1}(X \in parents(Y)). \qquad (1)$$

Definition 2 (Weighted controlled direct effect (WCDE), Pearl 2000)**.** Let $\mathbf{M}' \subseteq \mathbf{M}$ denote mediators that are parents of Y. WCDE is nonzero if and only if $X \in parents(Y)$ (i.e., SDC = 1):

$$\mathrm{WCDE} = \sum_{\mathbf{m}'} \left(\mathbb{E}[Y \mid do(x, \mathbf{m}')] - \mathbb{E}[Y \mid do(x^*, \mathbf{m}')]\right) P(\mathbf{m}'). \qquad (2)$$

RESULTS

- **FASTER.** LD3 ran 46–5870$\times$ faster than baselines on real-world data.
- **MORE PLAUSIBLE RESULTS.** Parent sets predicted from real-world data aligned with expert knowledge better than baselines.
- **ENABLES EFFECT ESTIMATION.** LD3 returns a valid adjustment set for the WCDE under a new graphical criterion.

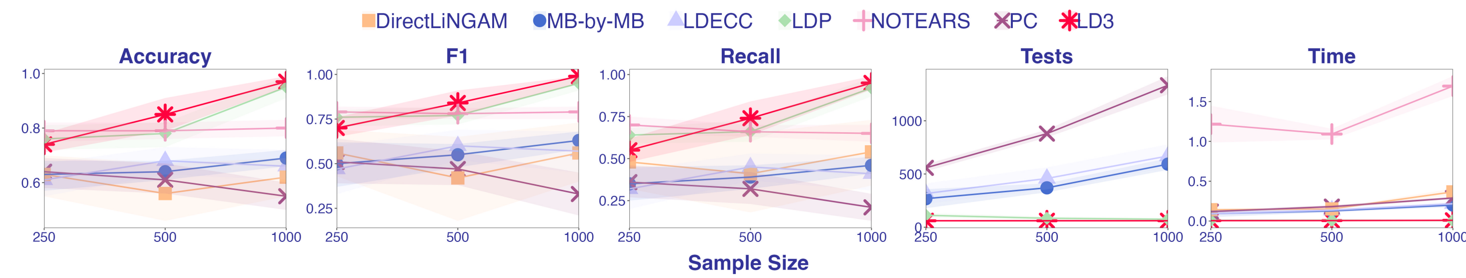

Fig. 3: Baseline results for parent discovery on the SANGIOVESE benchmark (`bnlearn`). Independence test count (Tests) is reported for constraint-based methods. Time is in seconds. Shaded regions denote 95% confidence intervals over ten replicates.

CASE STUDY: PATIENT SEX & LIVER TRANSPLANT ALLOCATION

Fairness query: Are sex-based disparities due to direct discrimination? $\Rightarrow$ **Graphical query:** Is sex (S) a parent of liver allocation (L)?

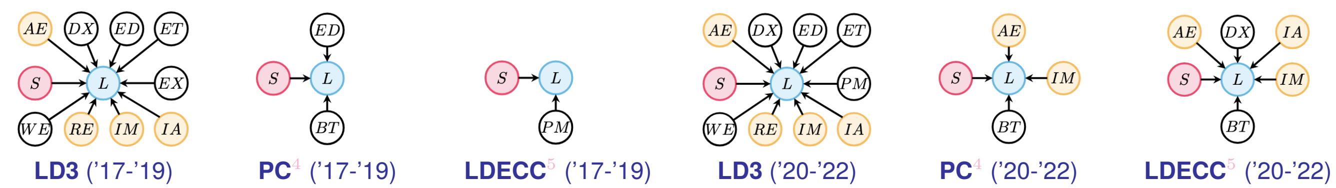

Fig. 4: Predicted parent sets for OPTN STAR datasets. **Known parents of L are in yellow.** Exposure = patient sex (S; red), outcome = receiving a liver (L; blue). *AE* = active exception case; *BT* = blood type; *DX* = diagnosis; *ED* = education; *ET* = ethnicity; *EX* = exception type; *IA* = initial age; *IM* = initial MELD; *PM* = payment method; *RE* = region; *WE* = weight. For all methods, SDC = 1 and WCDE p-value = 0.000.

REFERENCES

[1] Plečko, D., and Bareinboim, E. 2024. Causal Fairness Analysis: A Causal Toolkit for Fair Machine Learning. Foundations and Trends in Machine Learning. [2] Maasch, J.; Pan, W.; Gupta, S.; Kuleshov, V.; Gan, K.; Wang, F. 2024. Local Discovery by Partitioning: Polynomial-Time Causal Discovery Around Exposure-Outcome Pairs. UAI. [3] Pearl, J. 2000. Causality: Models, Reasoning and Inference. Cambridge University Press. ISBN 978-0-521-77362-1. [4] Spirtes, P.; Glymour, C.; Scheines, R. Causation, Prediction, and Search. Springer. [5] Gupta, S.; Childers, D.; and Lipton, Z. 2023. Local Causal Discovery for Estimating Causal Effects. CLeaR.

AAAI-25 / 25 FEBRUARY – 4 MARCH 2025 / PHILADELPHIA, PA, USA