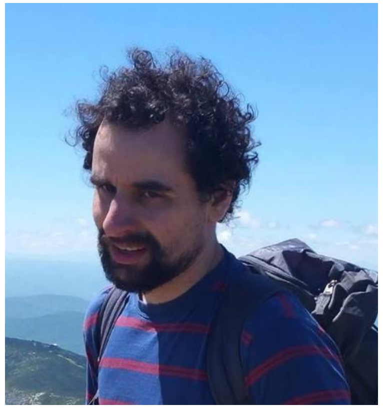
# Interventional Causal Discovery in a Mixture of DAGs

Burak Varıcı

Carnegie Mellon University



Dmitriy A. Katz

IBM Research

Dennis Wei

IBM Research

Prasanna Sattigeri

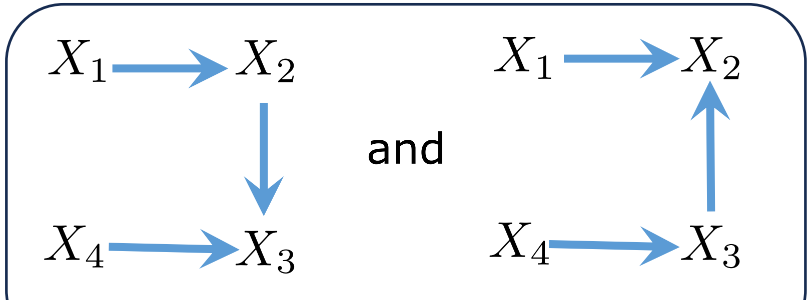
IBM Research

Ali Tajer

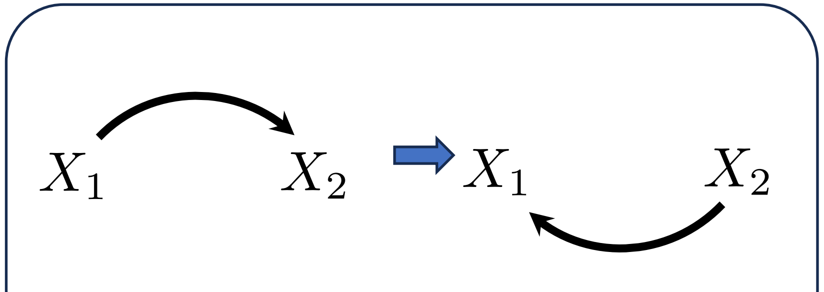
Rensselaer Polytechnic Institute

# Motivation

- **Time-varying systems**: causal relationships can change over time.
- **Cyclic relationships**: causal effects can form a cycle, e.g., feedback loops.
- **Modeling subpopulations**: e.g., subtypes of cancers do not share the same exact biological pathways



Aggregate of subpopulations



Causal system constantly evolves

These complex models are better modeled by a mixture of DAGs!

# Mixture model

- A mixture of (unknown) $K$ DAGs over $n$ nodes: $\{\mathcal{G}_1, \ldots, \mathcal{G}_K\}$

$$p_{\mathrm{m}} = \sum_{\ell \in [K]} w_\ell \cdot p_\ell \ , \qquad \text{where} \quad \sum_{\ell \in [K]} w_\ell = 1$$

- True edges: exist in at least one component DAG, define mixture parents

$$\mathrm{pa}_{\mathrm{m}}(i) = \bigcup_{\ell \in [K]} \mathrm{pa}_\ell(i)$$

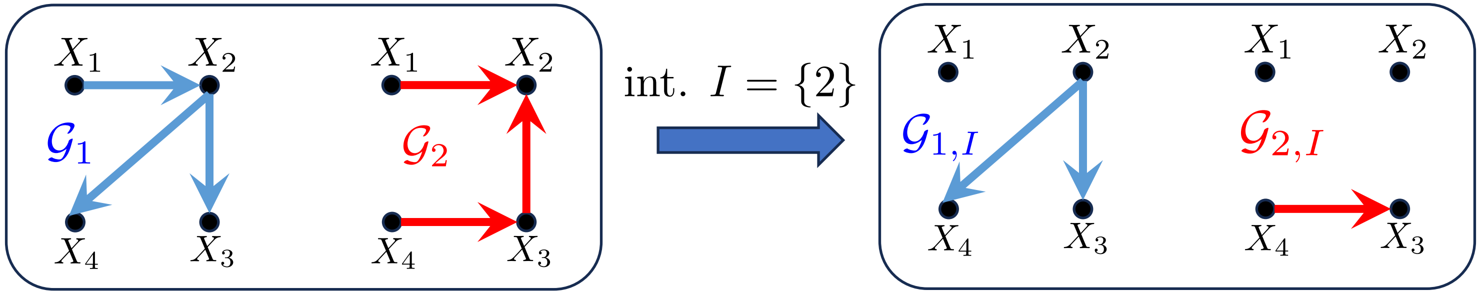
- Prior work: observational data, restrictive assumptions and/or results

# Interventional data

- Challenge for mixture: CI tests on observ. data is not sufficient for learning true edges
- **Intervention model:** Target a subset of nodes $I \subset [n]$, cut off parents across the mixture

$$p_\ell(x_i \mid x_{\mathrm{pa}(i)}) \to q_i(x_i) \ , \quad \forall i \in I, \ \forall \ell \in [K]$$

$\mathrm{pa_m}(3) = \{2, 4\}$




**Objective:** Identify *true edges* via interventions on the mixture

# Necessary & sufficient intervention sizes

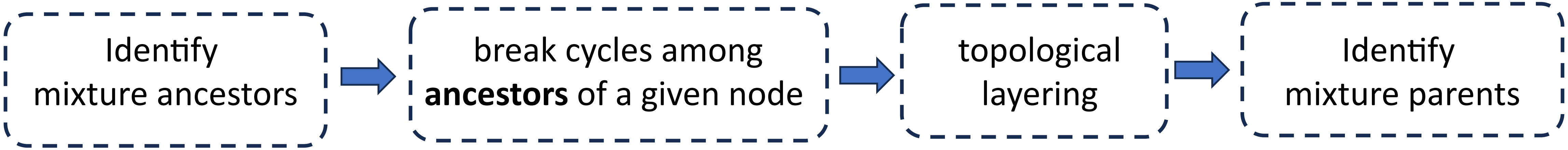
**Theorem (int. size - general):** To find mixture parents $\text{pa}_{\text{m}}(i)$ of node $i$ via CI tests,

1. interventions with size $|I| \leq |\text{pa}_{\text{m}}(i)| + 1$ are **sufficient** ,
2. at the worst-case, interventions with size $|I| = |\text{pa}_{\text{m}}(i)| + 1$ are **necessary**.

**Theorem (int. size - trees):** To find mixture parents $\text{pa}_{\text{m}}(i)$ of node $i$ in a mixture of trees

1. interventions with size $|I| \leq K + 1$ are **sufficient** ,
2. at the worst-case, interventions with size $|I| = K + 1$ are **necessary**.

# Learning algorithm

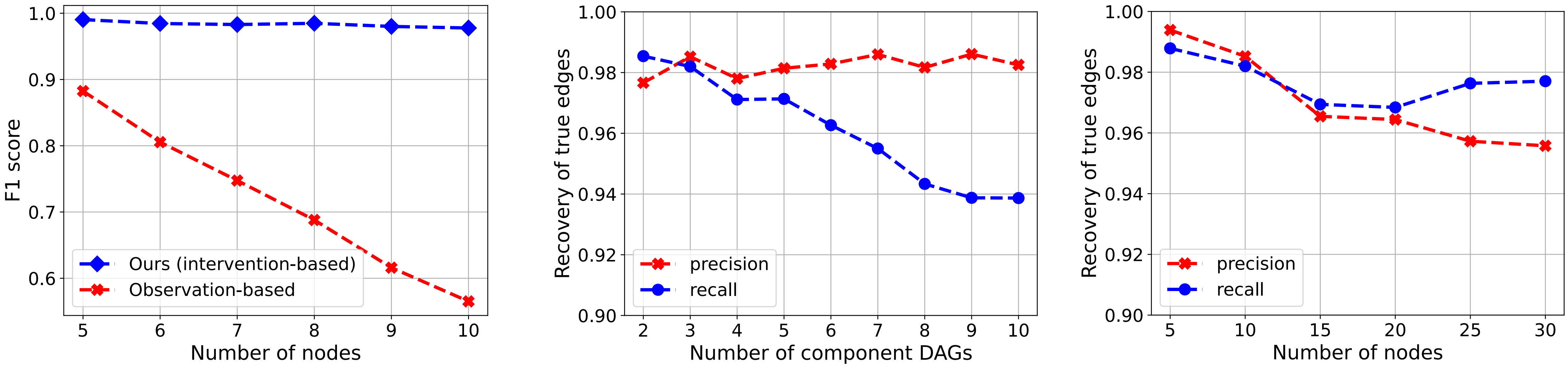



**cyclic complexity** $\tau_i$: minimal int. size to break such cycles (empirically small)

**Theorem:** Algorithm identifies all true edges using $\mathcal{O}(n^2)$ interventions, with size at most $|\mathrm{pa}_{\mathrm{m}}(i)| + \tau_i + 1$ for each node.

# Experiments

- Evaluate for recovering true edges on a mixture of Gaussians
- Verifying the need for interventions
- Strong performance for varying number of nodes & mixture components



# Interventional Causal Discovery in a Mixture of DAGs

Paper: https://arxiv.org/abs/2406.08666

Code: https://github.com/bvarici/intervention-mixture-DAG

Conference: Poster session 4, Wednesday December 12, 4.30-7.30pm

Contact: bvarici@andrew.cmu.edu

Project page