Model-oriented Graph Distances via Partially Ordered Sets

F. Richard Guo

Mar 2nd, 2026

Causal identification and discovery (CIFW02)

Isaac Newton Institute, Cambridge UK

Department of Statistics, University of Michigan

Armeen Taeb
University of Washington, Seattle

Leonard Henckel
University College Dublin

☞ This talk is based on recent work:
Model-oriented Graph Distances via Partially Ordered Sets, arXiv: 2511.10625.

Motivation

Statistical graphs

A statistical graph is a graph over a vertex set V with one or more types of edges.
It is a compact (and intuitive!) form for encoding and reasoning about dependencies.

❶ **Probabilistic graphical models**.

$$\boxed{\text{absence of edge between } u, v \implies \text{some form of independence between } X_u \text{ and } X_v}$$

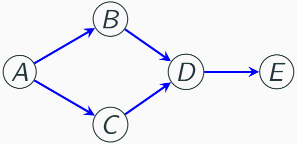
Statistical graphs

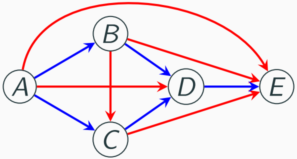
A statistical graph is a graph over a vertex set V with one or more types of edges.
It is a compact (and intuitive!) form for encoding and reasoning about dependencies.

❶ **Probabilistic graphical models**.

$$\boxed{\text{absence of edge between } u, v \implies \text{some form of independence between } X_u \text{ and } X_v}$$

☞ For example, in this Directed Acyclic Graph (DAG)





The missing '$B \to C$' posits

$$P(C \mid A, B) = P(C \mid A) \iff \boxed{B \perp\!\!\!\perp C \mid A}$$

Statistical graphs

A statistical graph is a graph over a vertex set V with one or more types of edges.
It is a compact (and intuitive!) form for encoding and reasoning about dependencies.

❶ **Probabilistic graphical models**.

$$\boxed{\text{absence of edge between } u, v \implies \text{some form of independence between } X_u \text{ and } X_v}$$

- Undirected Graph (UG)
- Directed Acyclic Graph / Bayesian Network (DAG)
- Completed Partially Directed Acyclic Graph / Essential Graph (CPDAG)
- Partial Ancestral Graph (PAG)
- Local Independence Graph

☞ They represent a set of distributions over $(X_v : v \in \mathrm{V})$.

Statistical graphs

A statistical graph is a graph over a vertex set V with one or more types of edges.
It is a compact (and intuitive!) form for encoding and reasoning about dependencies.

❷ **Causal graphical models**.

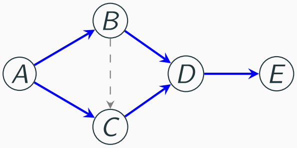
$$\boxed{\text{absence of } u \to v \implies \text{intervening on } X_u \text{ has no direct effect on } X_v}$$

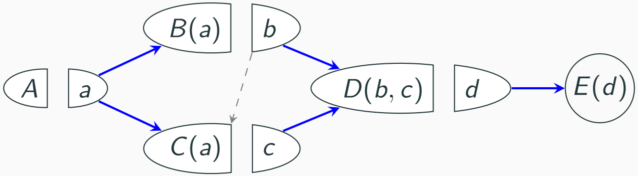
Statistical graphs

A statistical graph is a graph over a vertex set V with one or more types of edges.
It is a compact (and intuitive!) form for encoding and reasoning about dependencies.

❷ **Causal graphical models**.

$$\boxed{\text{absence of } u \to v \implies \text{intervening on } X_u \text{ has no direct effect on } X_v}$$





SWIG / FFRCISTG model (Richardson & Robins, 2013): The absence of '$B \to C$' encodes that $C(a, b) = C(a)$, i.e., C ignores any input from b.

Statistical graphs

A statistical graph is a graph over a vertex set V with one or more types of edges.
It is a compact (and intuitive!) form for encoding and reasoning about dependencies.

❷ **Causal graphical models**.

$$\boxed{\text{absence of } u \to v \implies \text{intervening on } X_u \text{ has no direct effect on } X_v}$$

- Directed Acyclic Graph / Bayesian Network (DAG)
- Acyclic Directed Mixed Graph (ADMG)
- Maximally Oriented Partially Directed Acyclic Graph (MPDAG)

☞ They represent a set of distributions over all the counterfactuals $X(\cdot)$.

$\hookrightarrow X(\cdot) = (X_v(x_I) : v \in \mathrm{V}, I \subseteq \mathrm{V}, x_I \in \mathcal{X}_I)$

$X(\cdot)$ contains $X(\emptyset)$, the naturally occurring random variables.

Semantics

A DAG can be either a probabilistic graphical model or a causal graphical model, depending on the semantics you use to interpret it.

A statistical graph has no intrinsic meaning until you interpret it with certain semantics!

- Consider a fixed vertex set V.

$$\begin{tikzcd} & & \text{Distributions of } (X_v : v \in \mathrm{V}) \\ \{\text{a class of graphs over } \mathrm{V}\} =: \mathfrak{G} \ni & \mathcal{G} \arrow[ur, "\mathcal{M}^o"] \arrow[dr, "\mathcal{M}^c"'] & \\ & & \text{Distributions of } X(\cdot) \end{tikzcd}$$

  - $\mathcal{M}^o(\mathcal{G})$ is a probabilistic model (aka Markov properties)
  - $\mathcal{M}^c(\mathcal{G})$ is a causal model

Distance between graphs

Given a class of statistical graphs $\mathfrak{G}$ over a common vertex set V, we want to have a distance metric $d : \mathfrak{G} \times \mathfrak{G} \to [0, \infty)$ that satisfies

1. Symmetry: $d(\mathcal{G}_s, \mathcal{G}_t) = d(\mathcal{G}_t, \mathcal{G}_s)$,
2. Anti-symmetry: $d(\mathcal{G}_s, \mathcal{G}_t) = 0$ if and only if $\mathcal{G}_s = \mathcal{G}_t$,
3. Triangle inequality: $d(\mathcal{G}_s, \mathcal{G}_t) \leq d(\mathcal{G}_s, \mathcal{G}) + d(\mathcal{G}_t, \mathcal{G})$ for any $\mathcal{G} \in \mathfrak{G}$.

Distance between graphs

Given a class of statistical graphs $\mathfrak{G}$ over a common vertex set V, we want to have a distance metric $d : \mathfrak{G} \times \mathfrak{G} \to [0, \infty)$ that satisfies

1. Symmetry: $d(\mathcal{G}_s, \mathcal{G}_t) = d(\mathcal{G}_t, \mathcal{G}_s)$,
2. Anti-symmetry: $d(\mathcal{G}_s, \mathcal{G}_t) = 0$ if and only if $\mathcal{G}_s = \mathcal{G}_t$,
3. Triangle inequality: $d(\mathcal{G}_s, \mathcal{G}_t) \leq d(\mathcal{G}_s, \mathcal{G}) + d(\mathcal{G}_t, \mathcal{G})$ for any $\mathcal{G} \in \mathfrak{G}$.

☞

- Benchmark causal discovery algorithms
- Quantify consistency and rate of convergence: $d(\widehat{\mathcal{G}}_n, \mathcal{G}) \leq \underbrace{d(\widehat{\mathcal{G}}_n, \mathcal{G}_\infty)}_{\text{variability}} + \underbrace{d(\mathcal{G}_\infty, \mathcal{G})}_{\text{approx. error}}$
- Uncertainty quantification and sensitivity analysis using $\{\mathcal{G}' : d(\mathcal{G}', \mathcal{G}) \leq \varepsilon\}$
- Measuring complexity of the model class: radius and covering number
- Finding a consensus graph through a Frechét mean (Ferguson & Meyer, 2023; Wang et al., 2025)

$$\widehat{\mathcal{G}} = \arg\min_{\mathcal{G}} \sum_i d^2(\widehat{\mathcal{G}}_i, \mathcal{G}).$$

Current standard: Structural Hamming distance

The Structural Hamming Distance

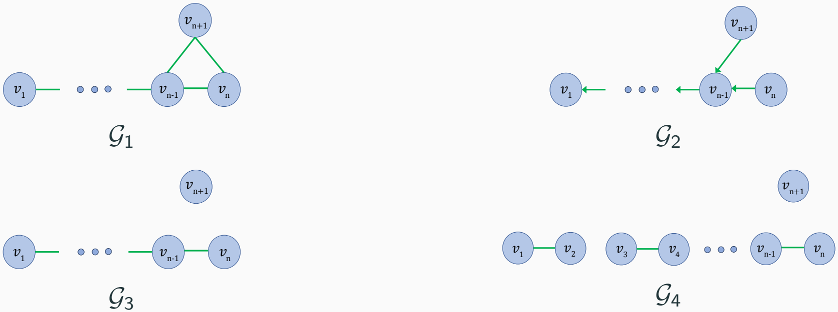
$$\boxed{\mathrm{SHD} = \#\{\text{edges that are different between two graphs}\}.}$$

Depending on how to count the difference between $\rightarrow$ and $-$, there are two common variants:

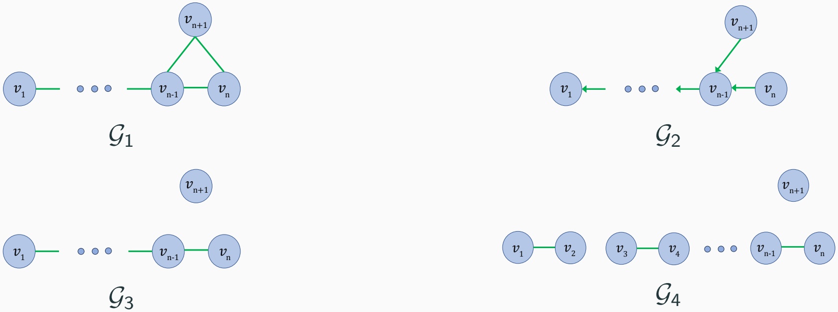
$$\mathrm{SHD}_1(\mathcal{G}_s, \mathcal{G}_t) := \sum_{i<j} \mathbb{I}\left[A^s_{i,j} \neq A^t_{i,j} \text{ or } A^s_{j,i} \neq A^t_{j,i}\right] \;\leq\; \mathrm{SHD}_2(\mathcal{G}_s, \mathcal{G}_t) := \sum_{i,j} \mathbb{I}\left[A^s_{i,j} \neq A^t_{i,j}\right].$$

☞ SHD measures the difference in the sets of edges — does not involve $\mathfrak{G}$ or $\mathcal{M}$!

Prob. CPDAGs over $n+1$ vertices: $\mathcal{G}_1 \succ \mathcal{G}_2 \succ \mathcal{G}_3 \succ \mathcal{G}_4$ in terms of the models they represent



Prob. CPDAGs over $n+1$ vertices: $\mathcal{G}_1 \succ \mathcal{G}_2 \succ \mathcal{G}_3 \succ \mathcal{G}_4$ in terms of the models they represent

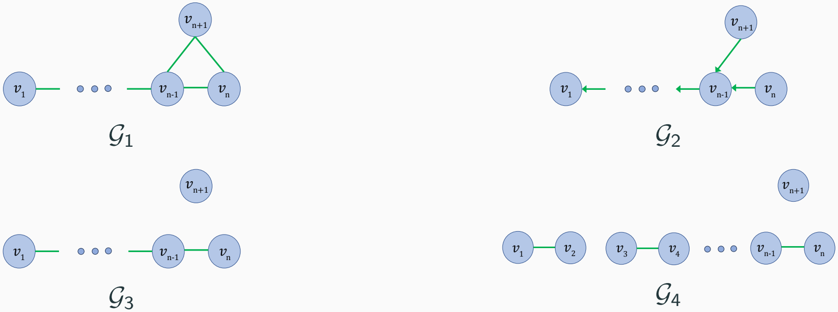


☞ Generate data under $\mathcal{G}_1$ (linear Gaussian SEM)

Distances and BIC ($\pm$ s.e.) relative to $\mathcal{G}_1$

Graph	BIC
$\mathcal{G}_2$	690 ± 5
$\mathcal{G}_3$	1084 ± 6
$\mathcal{G}_4$	10159 ± 20

Prob. CPDAGs over $n+1$ vertices: $\mathcal{G}_1 \succ \mathcal{G}_2 \succ \mathcal{G}_3 \succ \mathcal{G}_4$ in terms of the models they represent



☞ Generate data under $\mathcal{G}_1$ (linear Gaussian SEM)

Distances and BIC ($\pm$ s.e.) relative to $\mathcal{G}_1$

Graph	BIC	SHD_1	SHD_2	Model-oriented distance
$\mathcal{G}_2$	690 ± 5	$n+1$	$n+2$	1
$\mathcal{G}_3$	1084 ± 6	2	4	2
$\mathcal{G}_4$	10159 ± 20	$n/2+1$	$n+2$	$n/2+1$

Model-oriented poset and distance

Given a class of graphs $\mathfrak{G}$ and semantic $\mathcal{M}$ (probabilistic or causal), we define partial order

$$\mathcal{G}_1 \preceq \mathcal{G}_2 \quad \text{if} \quad \mathcal{M}(\mathcal{G}_1) \subseteq \mathcal{M}(\mathcal{G}_2).$$

If $\mathcal{M}(\cdot)$ is **injective** on $\mathfrak{G}$, then $\mathcal{L} := (\mathfrak{G}, \preceq)$ is the model-oriented poset (partially ordered set).

☞ satisfying reflexivity, transitivity and antisymmetry.

Model-oriented poset and distance

Given a class of graphs $\mathfrak{G}$ and semantic $\mathcal{M}$ (probabilistic or causal), we define partial order

$$\mathcal{G}_1 \preceq \mathcal{G}_2 \quad \text{if} \quad \mathcal{M}(\mathcal{G}_1) \subseteq \mathcal{M}(\mathcal{G}_2).$$

If $\mathcal{M}(\cdot)$ is **injective** on $\mathfrak{G}$, then $\mathcal{L} := (\mathfrak{G}, \preceq)$ is the model-oriented poset (partially ordered set).

☞ satisfying reflexivity, transitivity and antisymmetry.

- We write $\mathcal{G}_1 \prec \mathcal{G}_2$ if $\mathcal{G}_1 \preceq \mathcal{G}_2$ and $\mathcal{G}_1 \neq \mathcal{G}_2$.
- $\mathcal{G}_1, \mathcal{G}_2$ are said to be **comparable** if $\mathcal{G}_1 \preceq \mathcal{G}_2$ or $\mathcal{G}_1 \succeq \mathcal{G}_2$.
- A **chain** is a set of graphs that are pairwise comparable: $\mathcal{G}_1 \preceq \mathcal{G}_2 \preceq \cdots \preceq \mathcal{G}_n$.
- We say $\mathcal{G}_1$ covers $\mathcal{G}_2$, written as $\mathcal{G}_1 \gtrdot \mathcal{G}_2$, if $\mathcal{G}_1 \succ \mathcal{G}_2$ and there is no $\mathcal{G}_3$ such that $\mathcal{G}_1 \succ \mathcal{G}_3 \succ \mathcal{G}_2$. We define

$$\text{Neighbors}(\mathcal{G}) := \{\mathcal{G}' : \mathcal{G}' \lessdot \mathcal{G}\} \cup \{\mathcal{G}' : \mathcal{G}' \gtrdot \mathcal{G}\}.$$

- **Model-oriented distance** $d_{\mathcal{L}}$:= shortest path distance via neighbors.

Model-oriented poset and distance

Model-oriented distance = Shortest path distance on Hasse diagram

- Fix $\mathrm{V} = \{1, 2, 3\}$ and let $\mathfrak{G} := \{\text{all (probabilistic) undirected graphs over } \mathrm{V}\}$. We adopt

$$\mathcal{M}^{\mathrm{o}}_{\mathrm{UG}}(\mathcal{G}) := \left\{P(X_1, X_2, X_3) : \mathrm{A} \perp\!\!\!\perp \mathrm{B} \mid \mathrm{C}\ [\mathcal{G}] \implies X_{\mathrm{A}} \perp\!\!\!\perp X_{\mathrm{B}} \mid X_{\mathrm{C}}\ [P]\right\}.$$

☞

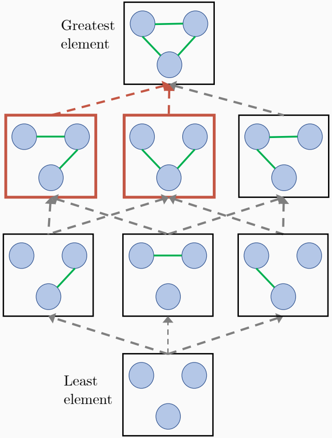
$$\mathcal{G}_1 \preceq \mathcal{G}_2 \iff \mathcal{G}_1 \text{ is a subgraph of } \mathcal{G}_2$$

$$\mathcal{G}_1 \lessdot \mathcal{G}_2 \iff \mathcal{G}_1 \text{ is a subgraph of } \mathcal{G}_2 \text{ with exactly one fewer edge.}$$

Model-oriented distance = Shortest path distance on Hasse diagram

- Fix $\mathrm{V} = \{1, 2, 3\}$ and let $\mathfrak{G} := \{$all (probabilistic) undirected graphs over $\mathrm{V}\}$. We adopt

$$\mathcal{M}^{\mathrm{o}}_{\mathrm{UG}}(\mathcal{G}) := \{P(X_1, X_2, X_3) : \mathrm{A} \perp\!\!\!\perp \mathrm{B} \mid \mathrm{C}\ [\mathcal{G}] \implies X_{\mathrm{A}} \perp\!\!\!\perp X_{\mathrm{B}} \mid X_{\mathrm{C}}\ [P]\} .$$



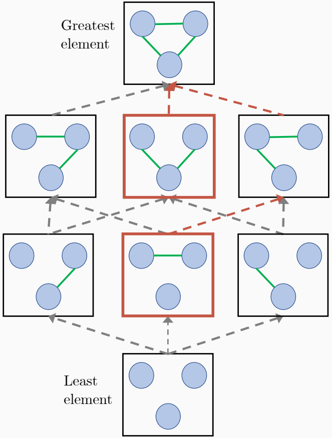
☞ **Hasse diagram**: $\mathcal{G}_1 \lessdot \mathcal{G}_2 \implies$ drawing $\mathcal{G}_1 \dashrightarrow \mathcal{G}_2$ upward.

$$d_{\mathcal{L}} = \min \{\mathrm{len}(p) : p \in \mathrm{P}_{\mathcal{L}}(\mathcal{G}_s, \mathcal{G}_t)\} , \quad \mathrm{P}_{\mathcal{L}} := \{\text{Paths on Hasse}\}.$$

Model-oriented distance = Shortest path distance on Hasse diagram

- Fix $\mathrm{V} = \{1, 2, 3\}$ and let $\mathfrak{G} := \{$all (probabilistic) undirected graphs over $\mathrm{V}\}$. We adopt

$$\mathcal{M}^{\mathrm{o}}_{\mathrm{UG}}(\mathcal{G}) := \{P(X_1, X_2, X_3) : \mathrm{A} \perp\!\!\!\perp \mathrm{B} \mid \mathrm{C}\ [\mathcal{G}] \implies X_{\mathrm{A}} \perp\!\!\!\perp X_{\mathrm{B}} \mid X_{\mathrm{C}}\ [P]\} \,.$$



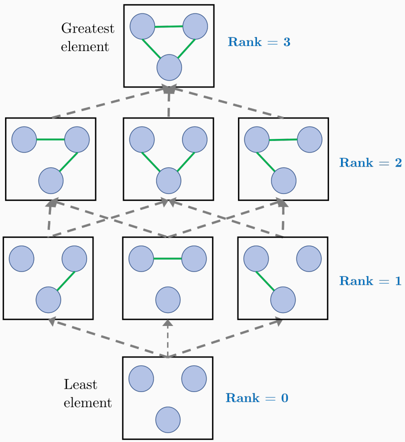
☞ **Hasse diagram**: $\mathcal{G}_1 \lessdot \mathcal{G}_2 \implies$ drawing $\mathcal{G}_1 \dashrightarrow \mathcal{G}_2$ upward.

$$d_{\mathcal{L}} = \min \{\mathrm{len}(p) : p \in \mathrm{P}_{\mathcal{L}}(\mathcal{G}_s, \mathcal{G}_t)\}\,, \quad \mathrm{P}_{\mathcal{L}} := \{\text{Paths on Hasse}\}.$$

Model-oriented distance = Shortest path distance on Hasse diagram

- Fix $\mathrm{V} = \{1,2,3\}$ and let $\mathfrak{G} := \{$all (probabilistic) undirected graphs over $\mathrm{V}\}$. We adopt

$$\mathcal{M}^{\mathrm{o}}_{\mathrm{UG}}(\mathcal{G}) := \{P(X_1, X_2, X_3) : \mathrm{A} \perp\!\!\!\perp \mathrm{B} \mid \mathrm{C}\ [\mathcal{G}] \implies X_{\mathrm{A}} \perp\!\!\!\perp X_{\mathrm{B}} \mid X_{\mathrm{C}}\ [P]\}\,.$$



☜ **Hasse diagram**: $\mathcal{G}_1 \lessdot \mathcal{G}_2 \implies$ drawing $\mathcal{G}_1 \dashrightarrow \mathcal{G}_2$ upward.

$$d_{\mathcal{L}} = \min\{\mathrm{len}(p) : p \in \mathrm{P}_{\mathcal{L}}(\mathcal{G}_s, \mathcal{G}_t)\}\,, \quad \mathrm{P}_{\mathcal{L}} := \{\text{Paths on Hasse}\}.$$

- $d_{\mathcal{L}}$ is well-defined if the Hasse diagram is connected, which holds if there exists a greatest element $\mathcal{G}_{\widehat{1}}$ or a least element $\mathcal{G}_{\widehat{0}}$. ☞ Typically, $\mathcal{G}_{\widehat{0}} =$ empty graph
- $\mathcal{L}$ is graded if it admits a rank function : $\mathfrak{G} \to \mathbb{N}$ s.t.

$$\mathcal{G}_1 \lessdot \mathcal{G}_2 \implies \mathrm{rank}(\mathcal{G}_1) + 1 = \mathrm{rank}(\mathcal{G}_2)$$

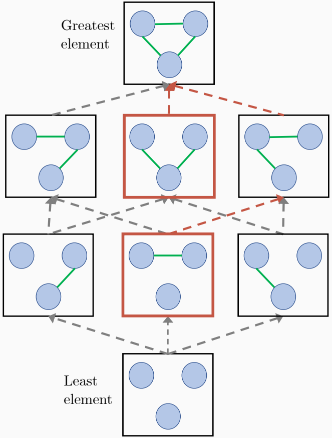
and $\mathrm{rank} = 0$ for a minimal element.

☞ If exists, $\mathrm{rank}(\cdot)$ is unique.

Model-oriented distance = Shortest path distance on Hasse diagram

- Fix $\mathrm{V} = \{1,2,3\}$ and let $\mathfrak{G} := \{$all (probabilistic) undirected graphs over $\mathrm{V}\}$. We adopt

$$\mathcal{M}^{\mathrm{o}}_{\mathrm{UG}}(\mathcal{G}) := \{P(X_1, X_2, X_3) : \mathrm{A} \perp\!\!\!\perp \mathrm{B} \mid \mathrm{C}\ [\mathcal{G}] \implies X_\mathrm{A} \perp\!\!\!\perp X_\mathrm{B} \mid X_\mathrm{C}\ [P]\}\,.$$



☜ For UGs, $\mathcal{L} \cong$ Boolean lattice over $\{(i,j) \in \mathrm{V} \times \mathrm{V} : i < j\}$, which implies

$$d_\mathcal{L} \equiv \mathrm{SHD}\,.$$

- $d_\mathcal{L}$ is well-defined if the Hasse diagram is connected, which holds if there exists a greatest element $\mathcal{G}_{\hat{1}}$ or a least element $\mathcal{G}_{\hat{0}}$. ☞ Typically, $\mathcal{G}_{\hat{0}}$ = empty graph
- $\mathcal{L}$ is graded if it admits a rank function : $\mathfrak{G} \to \mathbb{N}$ s.t.

$$\mathcal{G}_1 \lessdot \mathcal{G}_2 \implies \mathrm{rank}(\mathcal{G}_1) + 1 = \mathrm{rank}(\mathcal{G}_2)$$

and $\mathrm{rank} = 0$ for a minimal element.

☞ If exists, $\mathrm{rank}(\cdot)$ is unique.

Model-oriented distance

The model oriented distance can be viewed as

$$\boxed{d_{Lp}(\mathcal{G}_s, \mathcal{G}_t) = \min \#\{\text{basic operations to transform } \mathcal{G}_s \text{ to } \mathcal{G}_t\},}$$

where a basic operation is a covering relation defined by the model-oriented poset.

Model-oriented distance

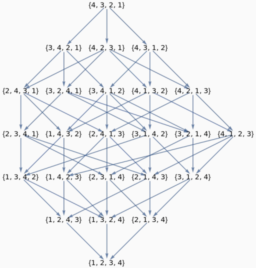
The model oriented distance can be viewed as

$$\boxed{d_{Lp}(\mathcal{G}_s, \mathcal{G}_t) = \min \#\{\text{basic operations to transform } \mathcal{G}_s \text{ to } \mathcal{G}_t\},}$$

where a basic operation is a covering relation defined by the model-oriented poset.

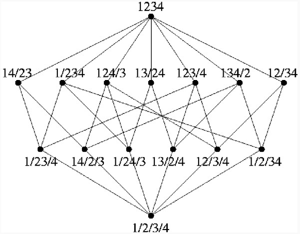
☞ This has been applied to other combinatorial objects.

Permutations / Rankings



Kendall's tau distance

Partitions / Clustering



Variation of information (Meila, 2005)

Distance between causal DAGs

* Now, let us fix a finite V and consider $\mathfrak{G} := \{\text{all DAGs over } \mathrm{V}\}$.

▶ For any $\mathcal{D} \in \mathfrak{G}$, we adopt the SWIG/FFRCISTG (Robins, 1986; Richardson & Robins, 2013; Zhao, 2025) **causal model**

$$\mathcal{M}^{\mathrm{c}}_{\mathrm{DAG}}(\mathcal{D}) = \{P(X(\cdot)) : P \text{ satisfies ❶ and ❷ with respect to } \mathcal{D}\},$$

❶ $\forall v \in \mathrm{V}$, $\mathrm{I} \subseteq \mathrm{V}$, $x \in \mathcal{X}$, $X_v(x_{\mathrm{I}}) = X_v(x_{\mathrm{I} \cap \mathrm{Pa}_{\mathcal{D}}(v)})$ w.p. 1, ☞ Counterfactual indexed by parents

❷ $\forall x \in \mathcal{X}$, the counterfactuals $\left\{X_v(x_{\mathrm{Pa}_{\mathcal{D}}(v)}) : v \in V\right\}$ are single-world mutually independent.

Distance between causal DAGs

✻ Now, let us fix a finite V and consider $\mathfrak{G} := \{\text{all DAGs over } \mathrm{V}\}$.

▶ For any $\mathcal{D} \in \mathfrak{G}$, we adopt the SWIG/FFRCISTG (Robins, 1986; Richardson & Robins, 2013; Zhao, 2025) **causal model**

$$\mathcal{M}^{\mathrm{c}}_{\mathrm{DAG}}(\mathcal{D}) = \{P(X(\cdot)) : P \text{ satisfies ❶ and ❷ with respect to } \mathcal{D}\},$$

❶ $\forall v \in \mathrm{V}$, $\mathrm{I} \subseteq \mathrm{V}$, $x \in \mathcal{X}$, $X_v(x_{\mathrm{I}}) = X_v(x_{\mathrm{I} \cap \mathrm{Pa}_{\mathcal{D}}(v)})$ w.p. 1, ☞ Counterfactual indexed by parents
❷ $\forall x \in \mathcal{X}$, the counterfactuals $\left\{X_v(x_{\mathrm{Pa}_{\mathcal{D}}(v)}) : v \in V\right\}$ are single-world mutually independent.

☞ It can be shown that $\mathcal{M}^{\mathrm{c}}_{\mathrm{DAG}}(\cdot)$ is injective. Further,

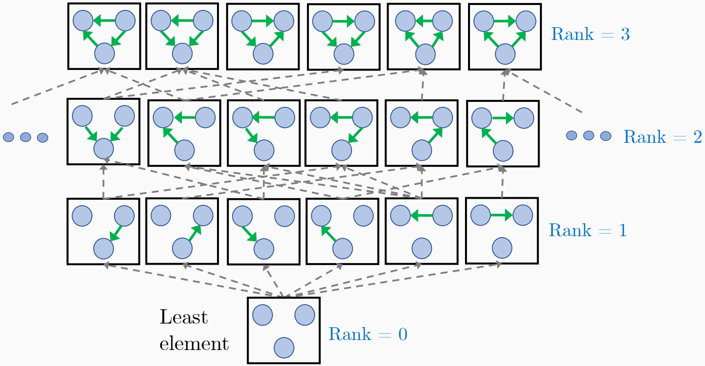
$$\mathcal{D}_1 \preceq \mathcal{D}_2 \iff \mathcal{D}_1 \text{ is a subgraph of } \mathcal{D}_2.$$

and

$$\mathcal{D}_1 \lessdot \mathcal{D}_2 \iff \mathcal{D}_1 \text{ is a subgraph of } \mathcal{D}_2 \text{ with exactly one fewer edge.}$$

Distance between causal DAGs

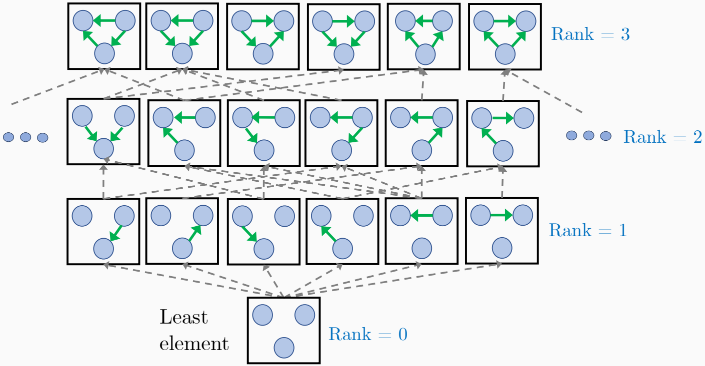
☞ DAG's poset is graded with rank = number of edges.



This poset is more complex than UG's, due to the acyclicity constraint.

Distance between causal DAGs

☞ DAG's poset is graded with rank = number of edges.



This poset is more complex than UG's, due to the acyclicity constraint.

☞ Nevertheless, we can still show $d_{\mathcal{L}}(\mathcal{D}_1, \mathcal{D}_2) \equiv \mathrm{SHD}(\mathcal{D}_1, \mathcal{D}_2)$.

Down-up/up-down distance

We consider down-up paths and up-down paths

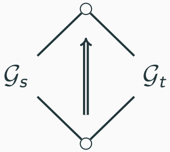
$$\mathrm{P}_{\mathcal{L},\downarrow\uparrow}(\mathcal{G}_s, \mathcal{G}_t) := \{\mathcal{G}_s \gtrdot \ldots \gtrdot \mathcal{G}^* \lessdot \ldots \lessdot \mathcal{G}_t\}, \quad \mathrm{P}_{\mathcal{L},\uparrow\downarrow}(\mathcal{G}_s, \mathcal{G}_t) := \{\mathcal{G}_s \lessdot \ldots \lessdot \mathcal{G}^* \gtrdot \ldots \gtrdot \mathcal{G}_t\}$$

and accordingly the distances

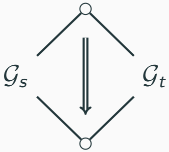
$$d_{\mathcal{L},\downarrow\uparrow}(\mathcal{G}_s, \mathcal{G}_t) := \min\{\mathrm{len}(p) : p \in \mathrm{P}_{\mathcal{L},\downarrow\uparrow}(\mathcal{G}_s, \mathcal{G}_t)\}, \; d_{\mathcal{L},\uparrow\downarrow}(\mathcal{G}_s, \mathcal{G}_t) := \min\{\mathrm{len}(p) : p \in \mathrm{P}_{\mathcal{L},\uparrow\downarrow}\}.$$

☞ They are well-defined, if and only if, respectively, $\mathcal{G}_{\hat{0}}$ and $\mathcal{G}_{\hat{1}}$ exist.

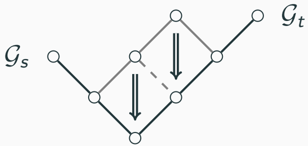
Semimodularity



Upper semimodularity



Lower semimodularity



Theorem When $\mathcal{G}_{\hat{0}}$ exists, we have

$$d_{\mathcal{L}} \equiv d_{\mathcal{L},\downarrow\uparrow} \iff \mathcal{L} \text{ is lower-semimodular;}$$

when $\mathcal{G}_{\hat{1}}$ exists, we have

$$d_{\mathcal{L}} \equiv d_{\mathcal{L},\uparrow\downarrow} \iff \mathcal{L} \text{ is upper-semimodular.}$$

☞ This generalizes earlier results in Monjardet (1981) and Foldes and Radeleczki (2021).

Distance between causal DAGs

Proposition DAG's poset is **lower semimodular** and a **meet semi-lattice**.

- $\mathcal{L}$ is a meet (join) semi-lattice if every $\mathcal{G}_1, \mathcal{G}_2 \in \mathfrak{G}$ admits $\mathcal{G}_1 \wedge \mathcal{G}_2$ $(\mathcal{G}_1 \vee \mathcal{G}_2)$, i.e., the greatest lower bound (least upper bound).

Distance between causal DAGs

Proposition DAG's poset is **lower semimodular** and a **meet semi-lattice**.

▶ $\mathcal{L}$ is a meet (join) semi-lattice if every $\mathcal{G}_1, \mathcal{G}_2 \in \mathfrak{G}$ admits $\mathcal{G}_1 \wedge \mathcal{G}_2$ ($\mathcal{G}_1 \vee \mathcal{G}_2$), i.e., the greatest lower bound (least upper bound).

Suppose $\mathcal{D}_3 = (\mathrm{V}, \mathrm{E}_3)$ covers two different DAGs $\mathcal{D}_1 = (\mathrm{V}, \mathrm{E}_1)$ and $\mathcal{D}_2 = (\mathrm{V}, \mathrm{E}_2)$.

We have $\mathrm{E}_1 \not\subseteq \mathrm{E}_2$, $\mathrm{E}_2 \not\subseteq \mathrm{E}_1$ and $|\mathrm{E}_1 \triangle \mathrm{E}_2| = 2$.

It follows that $\mathcal{D}_1, \mathcal{D}_2 \gtrdot \mathcal{D}_4 := (\mathrm{V}, \mathrm{E}_1 \cap \mathrm{E}_2) = \mathcal{D}_1 \wedge \mathcal{D}_2$.

Distance between causal DAGs

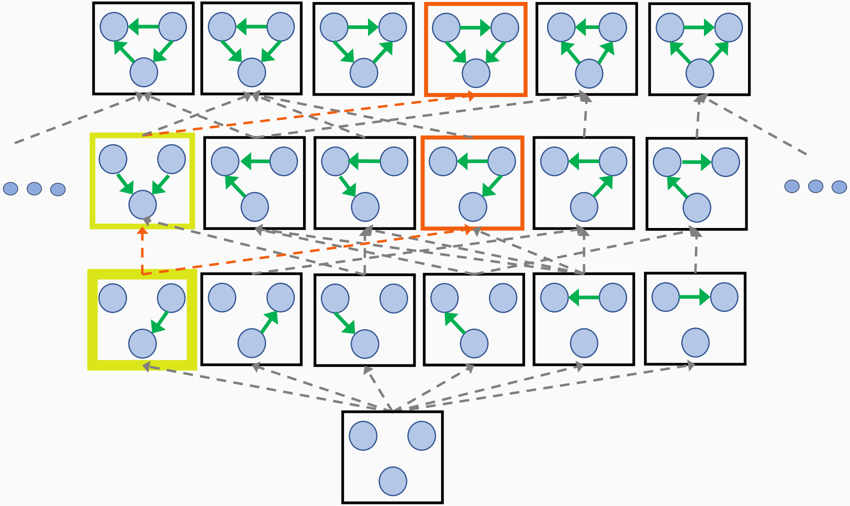
Proposition DAG's poset is **lower semimodular** and a **meet semi-lattice**.

▶ $\mathcal{L}$ is a meet (join) semi-lattice if every $\mathcal{G}_1, \mathcal{G}_2 \in \mathfrak{G}$ admits $\mathcal{G}_1 \wedge \mathcal{G}_2$ ($\mathcal{G}_1 \vee \mathcal{G}_2$), i.e., the greatest lower bound (least upper bound).

☞ For any pair of DAGs $\mathcal{D}_1, \mathcal{D}_2$,

$$
\begin{aligned}
& d_{\mathcal{L}}(\mathcal{D}_1, \mathcal{D}_2) \\
\text{(lower semimodularity)} \quad & = d_{\mathcal{L},\downarrow\uparrow}(\mathcal{D}_1, \mathcal{D}_2) \\
\text{(meet semi-lattice)} \quad & = d_{\mathcal{L},\downarrow}(\mathcal{D}_1, \mathcal{D}_1 \wedge \mathcal{D}_2) + d_{\mathcal{L},\uparrow}(\mathcal{D}_1 \wedge \mathcal{D}_2, \mathcal{D}_2) \\
\text{(gradedness)} \quad & = [\mathrm{rank}(\mathcal{D}_1) - \mathrm{rank}(\mathcal{D}_1 \wedge \mathcal{D}_2)] + [\mathrm{rank}(\mathcal{D}_2) - \mathrm{rank}(\mathcal{D}_1 \wedge \mathcal{D}_2)] \\
\text{(rank)} \quad & = |\mathrm{E}_1 \setminus (\mathrm{E}_1 \cap \mathrm{E}_2)| + |\mathrm{E}_2 \setminus (\mathrm{E}_1 \cap \mathrm{E}_2)| \\
& = |\mathrm{E}_1 \triangle \mathrm{E}_2| = \mathrm{SHD}(\mathcal{D}_1, \mathcal{D}_2).
\end{aligned}
$$

Distance between causal DAGs



Structural properties of graph posets

☞ Characterizing and computing the model-oriented distance critically depends on the structural properties of poset.

Graph class	Model	Least element	Greatest element	Graded	Semi-lattice	Semimodular
UG	prob.	empty graph	full graph	✓	join + meet	lower + upper
DAG	causal	empty graph	✗	✓	meet	lower
CPDAG	prob.	empty graph	full graph	✓	✗	✗
Polytree CPDAG	prob.	empty graph	✗	✓	✗	lower
MPDAG	causal	empty graph	full graph	✗	✗	✗
Polytree MPDAG	causal	empty graph	✗	✓	✗	lower

☞ For CPDAGs and MPDAGs, $d_{\mathcal{L}}$ no longer admits a closed form and we must develop algorithms for computing the distance.

Algorithms

Probabilistic CPDAGs

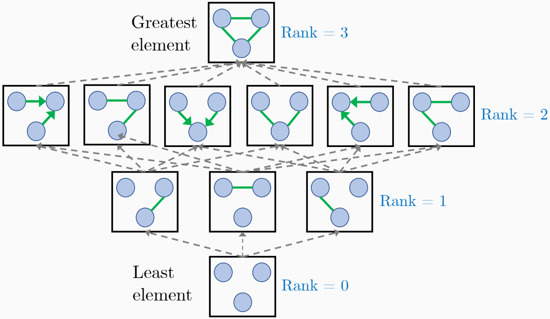
A CPDAG (Completed Partially Directed Acyclic Graph) / essential graph represents a class of Markov equivalent DAGs. For a finite V, let $\mathfrak{G} := \{$all CPDAGs over $\mathrm{V}\}$ and

$$\mathcal{M}^{\mathrm{o}}_{\mathrm{CPDAG}}(\mathcal{G}) := \{P : A \perp\!\!\!\perp_d B \mid C \text{ in any (and hence every) DAG } D \in [\mathcal{G}] \implies X_A \perp\!\!\!\perp X_B \mid X_C \ [P]\}.$$

Probabilistic CPDAGs

A CPDAG (Completed Partially Directed Acyclic Graph) / essential graph represents a class of Markov equivalent DAGs. For a finite V, let $\mathfrak{G} := \{$all CPDAGs over $\mathrm{V}\}$ and

$$\mathcal{M}^{\mathrm{o}}_{\mathrm{CPDAG}}(\mathcal{G}) := \{P : A \perp\!\!\!\perp_d B \mid C \text{ in any (and hence every) DAG } D \in [\mathcal{G}] \implies X_A \perp\!\!\!\perp X_B \mid X_C \; [P]\}.$$



CPDAG Poset

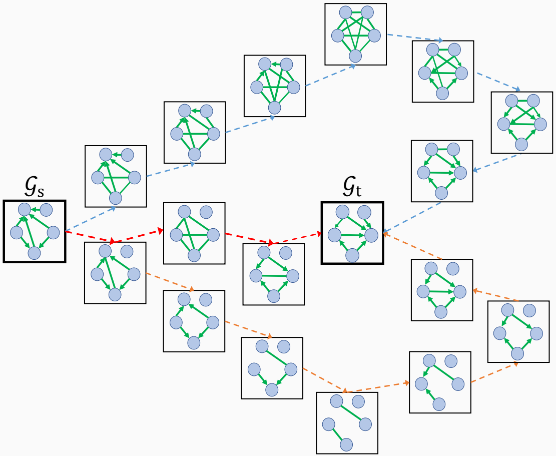
- $\mathcal{L}$ is graded with $\mathrm{rank}(\cdot) = $ total $\#$ edges.
- $\mathcal{G}_s \preceq \mathcal{G}_t \iff$ there is a sequence of DAGs

$$[\mathcal{G}_s] \ni \mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_k \in [\mathcal{G}_t]$$

  such that every $\mathcal{D}_i$ is transformed from $\mathcal{D}_{i-1}$ by either (i) reversing a covered edge or (ii) adding an edge (Chickering, 2002).
- $\mathcal{G}_s \lessdot \mathcal{G}_t \iff$ there exist $\mathcal{D}_s \in [\mathcal{G}_s], \mathcal{D}_t \in [\mathcal{G}_t]$ s.t. $\mathcal{D}_s$ is a subgraph of $\mathcal{D}_t$ with exactly one fewer edge.

Distance between CPDAGs

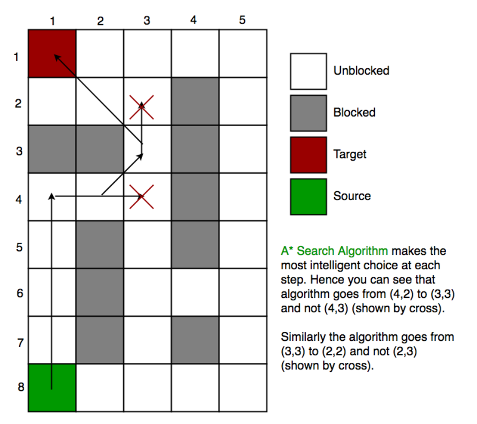
☞ $d_{\mathcal{L}} \neq \mathrm{SHD}$ in general. Without semimodularity, $d_{\mathcal{L},\uparrow\downarrow}, d_{\mathcal{L},\downarrow\uparrow}$ are only upper bounds on $d_{\mathcal{L}}$.

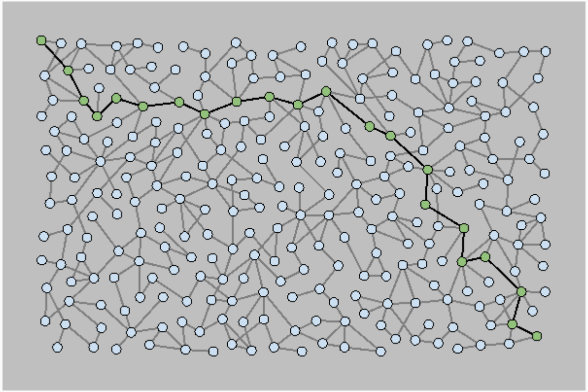


- $d_{\mathcal{L}} = 4$ (zigzag) but $d_{\mathcal{L},\uparrow\downarrow} = d_{\mathcal{L},\downarrow\uparrow} = 8$.

Computing the distance

A* (Hart et al., 1968), a great achievement in **early-day AI**, is a heuristic-guided search algorithm to find a shortest path.





A* algorithm with branch and bound

- We designed a version of A* that depends on 3 subroutines:

1. $\textsc{EnumNeighbors}(\mathcal{G})$ ☞ Need characterization of '$\lessdot$'
2. A lower bound h of the distance: ☞ Admissible heuristic that guides the search

$$0 \leq h(\mathcal{G}) \leq d_{\mathcal{L}}(\mathcal{G}, \mathcal{G}_t)$$

3. An upper bound u on the distance: ☞ Pruning and early termination

$$d_{\mathcal{L}}(\mathcal{G}, \mathcal{G}_t) \leq u(\mathcal{G}) \leq +\infty$$

☞ When $h \equiv 0$ and $u \equiv +\infty$, the algorithm reduces to breadth-first or depth-first search depending the implementation of its priority queue.

A* algorithm with branch and bound

- We designed a version of A* that depends on 3 subroutines:

1. ENUMNEIGHBORS($\mathcal{G}$) ☞ Need characterization of '$\lessdot$'
2. A lower bound h of the distance: ☞ Admissible heuristic that guides the search

$$0 \leq h(\mathcal{G}) \leq d_{\mathcal{L}}(\mathcal{G}, \mathcal{G}_t)$$

3. An upper bound u on the distance: ☞ Pruning and early termination

$$d_{\mathcal{L}}(\mathcal{G}, \mathcal{G}_t) \leq u(\mathcal{G}) \leq +\infty$$

☞ When $h \equiv 0$ and $u \equiv +\infty$, the algorithm reduces to breadth-first or depth-first search depending the implementation of its priority queue.

Theorem Suppose $\mathfrak{G}$ is finite and the associated model-oriented poset is connected. Supplied with the subroutines ENUMNEIGHBORS$(\cdot)$, $u(\cdot)$ and $h(\cdot)$ that satisfy the conditions, the A* algorithm will terminate and return $d_{\mathcal{L}}(\mathcal{G}_s, \mathcal{G}_t)$.

A* for CPDAGs

❶ **EnumNeighbors($\mathcal{G}$)** consists of zero or more covered edge reversal and one edge addition or removal. ☞ These correspond to INSERT and DELETE of Chickering's GES

A* for CPDAGs

❶ **EnumNeighbors($\mathcal{G}$)** consists of zero or more covered edge reversal and one edge addition or removal. ☞ These correspond to INSERT and DELETE of Chickering's GES

❷ **Upper bound** $u(\mathcal{G}, \mathcal{G}_t) = \min\left(d_{\mathcal{L},\uparrow\downarrow}, d_{\mathcal{L},\downarrow\uparrow}\right)$, which itself can be computed from another A*.

A* for CPDAGs

❶ **EnumNeighbors($\mathcal{G}$)** consists of zero or more covered edge reversal and one edge addition or removal. ☞ These correspond to INSERT and DELETE of Chickering's GES

❷ **Upper bound** $u(\mathcal{G}, \mathcal{G}_t) = \min\left(d_{\mathcal{L},\uparrow\downarrow}, d_{\mathcal{L},\downarrow\uparrow}\right)$, which itself can be computed from another A*.

❸ **Lower bound**

$$d_{\mathcal{L}}(\mathcal{G}_s, \mathcal{G}_t) \geq |\,\mathsf{sk}(\mathcal{G}_s) \triangle\, \mathsf{sk}(\mathcal{G}_t)| + 2\, m_{\mathrm{op}}(\mathrm{S}(\mathcal{G}_s, \mathcal{G}_t)),$$

where m_{op} is the minimal number of operations (adding a covered edge or removing an edge) that align the set of colliders S that differ between $\mathcal{G}_s$ and $\mathcal{G}_t$.

A* for CPDAGs

❶ **EnumNeighbors($\mathcal{G}$)** consists of zero or more covered edge reversal and one edge addition or removal. ☞ These correspond to INSERT and DELETE of Chickering's GES

❷ **Upper bound** $u(\mathcal{G}, \mathcal{G}_t) = \min\left(d_{\mathcal{L},\uparrow\downarrow}, d_{\mathcal{L},\downarrow\uparrow}\right)$, which itself can be computed from another A*.

❸ **Lower bound**

$$d_{\mathcal{L}}(\mathcal{G}_s, \mathcal{G}_t) \geq |\,\mathrm{sk}(\mathcal{G}_s) \triangle \mathrm{sk}(\mathcal{G}_t)| + 2\, m_{\mathrm{op}}(\mathrm{S}(\mathcal{G}_s, \mathcal{G}_t)),$$

where m_{op} is the minimal number of operations (adding a covered edge or removing an edge) that align the set of colliders S that differ between $\mathcal{G}_s$ and $\mathcal{G}_t$.

These are **critical** because the $\mathfrak{G}$ can be astronomically large!

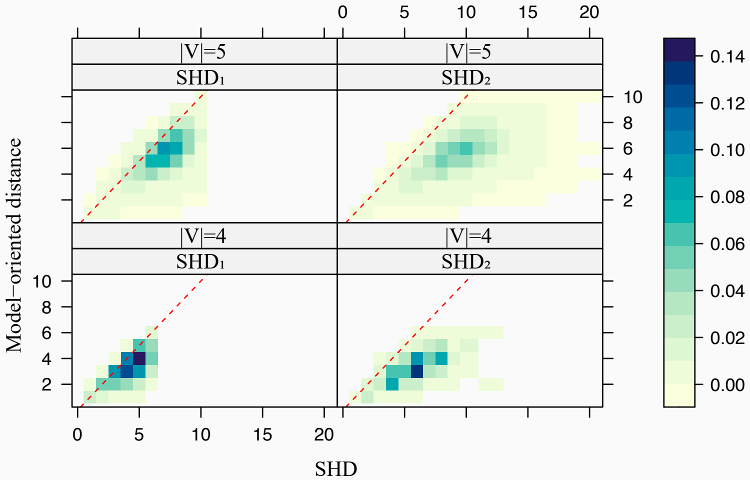
$$\text{For } \mathfrak{G} = \{\text{CPDAGs over 13 vertices}\}, \quad |\mathfrak{G}| > 5 \times 10^{30}.$$

☞ For a pair randomly drawn from an Erdös–Rényi model over $\mathfrak{G}$ with edge prob. 0.2, our A* algorithm implemented in C++ takes about **0.1 second** on average to compute the distance!

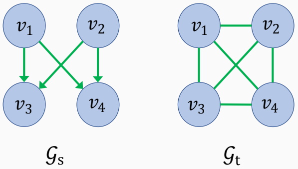
✻ We thank Jian Kang, Carter Lembo and Arnav Mazumder at the University of Washington.

Numerical results

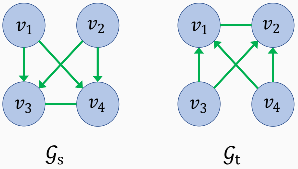
☞ Joint distribution of $(d_{\mathcal{L}}, \mathrm{SHD})$ computed over all pairs of CPDAGs over 4 or 5 vertices, where the identity line is drawn as dashed.



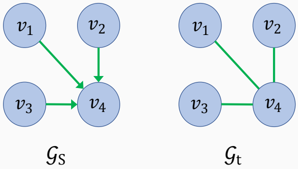
Pairs of graphs with large discrepancies



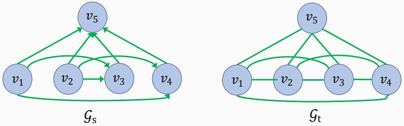
(a) $d_{\mathcal{L}} = 2$, $\mathrm{SHD}_1 = 6$, $\mathrm{SHD}_2 = 8$



(b) $d_{\mathcal{L}} = 2$, $\mathrm{SHD}_1 = 6$, $\mathrm{SHD}_2 = 12$



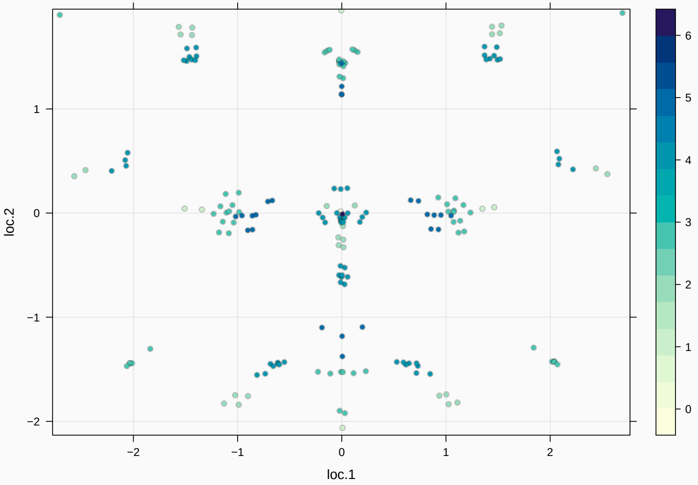
(c) $d_{\mathcal{L}} = 4$, $\mathrm{SHD}_1 = 3$, $\mathrm{SHD}_2 = 3$

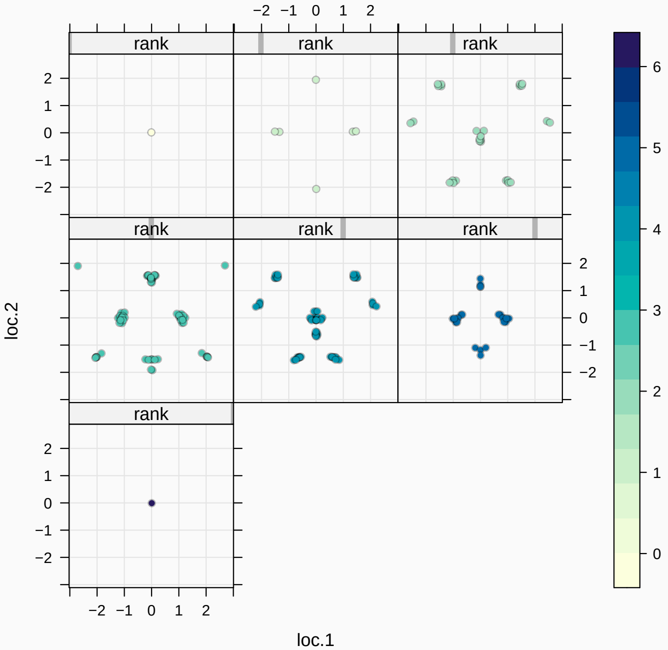


(d) $d_{\mathcal{L}} = 2$, $\mathrm{SHD}_1 = 10$, $\mathrm{SHD}_2 = 12$

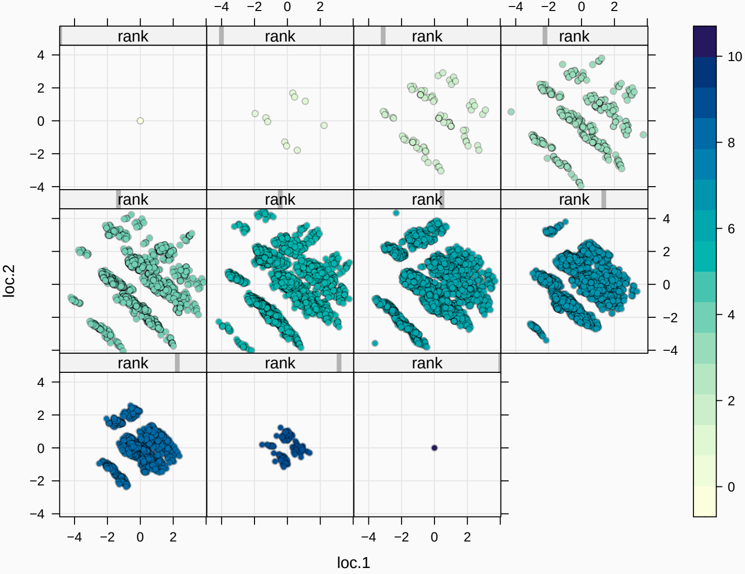
Embedded in $\mathbb{R}^2$

- Euclidean embedding of 185 CPDAGs over 4 vertices via multidimensional scaling (MDS).




rank
rank
rank
rank
rank
rank
rank
loc.2
loc.1
−2
−1
0
1
2
3
4
5
6

- 8782 CPDAGs over 5 vertices.



Restricting to polytree CPDAGs

The model-oriented distance is adapted to $\mathfrak{G}$ and $\mathcal{M}$.
If for some reason, only a subset $\mathfrak{G}' \subset \mathfrak{G}$ is worth considering, $d_{\mathcal{L}}$ can be redefined with respect to the sub-poset induced by $(\mathfrak{G}', \mathcal{M})$.

Restricting to polytree CPDAGs

The model-oriented distance is adapted to $\mathfrak{G}$ and $\mathcal{M}$.
If for some reason, only a subset $\mathfrak{G}' \subset \mathfrak{G}$ is worth considering, $d_{\mathcal{L}}$ can be redefined with respect to the sub-poset induced by $(\mathfrak{G}', \mathcal{M})$.

☞ For example, we also considered

$$\mathfrak{G}_{\mathrm{poly}} := \{\text{CPDAGs over } \mathrm{V} \text{ whose skeletons do not contain cycles}\},$$

which has been studied for structure learning (Rebane & Pearl, 1987; Sepehr & Materassi, 2019; Grüttemeier et al., 2021; Tramontano et al., 2022).

Polytree CPDAGs have a poset $\mathcal{L}$ that is lower-semimodular, so

$$d_{\mathcal{L}} \equiv d_{\mathcal{L},\downarrow\uparrow}.$$

Causal MPDAGs

A MPDAG (Maximally Oriented Partially Directed Acyclic Graph) represents a class of Markov equivalent DAGs subject to certain background knowledge (e.g., v_1 temporally precedes v_2). We interpret it as a causal model

$$\mathcal{M}^{c}_{\mathrm{MPDAG}}(\mathcal{G}) = \bigcup \{\mathcal{M}^{c}_{\mathrm{DAG}}(\mathcal{D}) : \mathcal{D} \text{ is a DAG represented by } \mathcal{G}\}.$$

MPDAG Poset

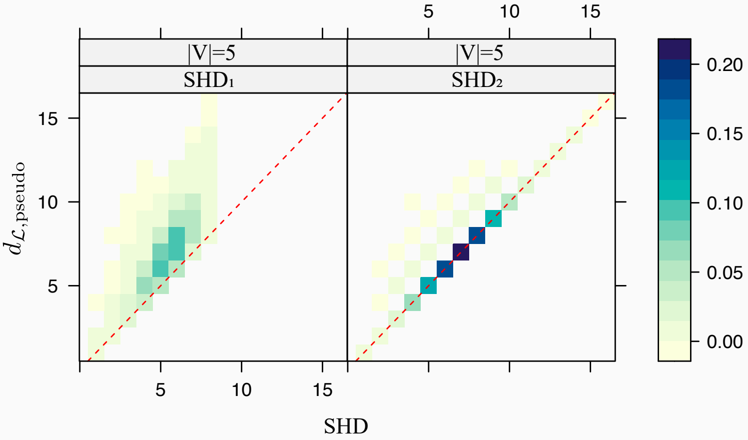
1. $\mathcal{G}_s \preceq \mathcal{G}_t \iff \forall \mathcal{D}_s \in [\mathcal{G}_s],\ \exists \mathcal{D}_t \in [\mathcal{G}_t], \text{s.t. } \mathcal{D}_s$ is a subgraph of $\mathcal{D}_t$.
2. The poset is **not graded** when $|\mathrm{V}| \geq 4$.
3. Between any MPDAG $\mathcal{G}$ and the empty graph $\mathcal{G}_{\hat{0}}$, there exists a maximal chain $\mathcal{G}_{\hat{0}} =: \mathcal{G}_0 \lessdot \cdots \lessdot \mathcal{G}_l := \mathcal{G}$ of length

$$\boxed{\mathrm{pseudo\text{-}rank}(\mathcal{G}) := \#\text{directed edges} + 2 \times \#\text{undirected edges},}$$

which can be used to define an approximation to $d_{\mathcal{L}}$.

Distance between MPDAGs

☞ Joint distribution of $(d_{\mathcal{L},\text{pseudo-rank}}, \text{SHD})$ computed over all pairs of all pairs of polytree MPDAGs over 5 vertices.



Summary

Summary and future work

✻ Don't judge a book by its cover. Similarly, we should not judge a graph by merely what it looks like!

☞ Rather, we should look at a graph ➊ through the model it represents and ➋ relative to the space of all models in consideration.

Summary and future work

✻ Don't judge a book by its cover. Similarly, we should not judge a graph by merely what it looks like!

☞ Rather, we should look at a graph ❶ through the model it represents and ❷ relative to the space of all models in consideration.

Future directions

1. Applying to other graphs, e.g., acyclic directed mixed graphs, local independence graphs (Didelez, 2008; Mogensen & Hansen, 2020)
2. Improve algorithms to scale up to even larger graphs.
3. New perspective on Chickering's GES: its greedy search is essentially traveling on a up-down path. What can be said about these algorithms?
4. Uncertainty quantification for structure learning

$$P\left(\widehat{\mathcal{G}}_l \preceq \mathcal{G} \preceq \widehat{\mathcal{G}}_u\right) \approx 95\%.$$

THANKS

References

Chickering, D. M. (2002).**Optimal structure identification with greedy search.** *Journal of Machine Learning Research*, *3*(Nov), 507–554.

Didelez, V. (2008).**Graphical models for marked point processes based on local independence.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(1), 245–264.

Ferguson, D., & Meyer, F. G. (2023).**Theoretical analysis and computation of the sample fréchet mean of sets of large graphs for various metrics.** *Information and Inference: A Journal of the IMA*, *12*(3), 1347–1404. https://doi.org/10.1093/imaiai/iaad002

Foldes, S., & Radeleczki, S. (2021).**On distances and metrics in discrete ordered sets.** *Mathematica Bohemica*, *146*(3), 251–262.

Grüttemeier, N., Komusiewicz, C., & Morawietz, N. (2021).**On the parameterized complexity of polytree learning.** *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, 4221–4227.

- Hart, P. E., Nilsson, N. J., & Raphael, B. (1968).**A formal basis for the heuristic determination of minimum cost paths.** *IEEE Transactions on Systems Science and Cybernetics*, *4*(2), 100–107.
- Meila, M. (2005).**Comparing clusterings: An axiomatic view.** *Proceedings of the Twenty-Second International Conference on Machine Learning (ICML)*, 577–584.
- Mogensen, S. W., & Hansen, N. R. (2020).**Markov equivalence of marginalized local independence graphs.** *The Annals of Statistics*, *48*(1), 539–559.
- Monjardet, B. (1981).**Metrics on partially ordered sets - a survey.** *Discrete Mathematics*, *35*, 173–184. https://api.semanticscholar.org/CorpusID:11792278
- Rebane, G., & Pearl, J. (1987).**The recovery of causal poly-trees from statistical data.** *Proceedings of the Third Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 175–182.
- Richardson, T. S., & Robins, J. M. (2013). **Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality (tech. rep. No. 128).** Center for the Statistics and the Social Sciences, University of Washington Series.

Robins, J. M. (1986).**A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect.** *Mathematical Modelling*, *7*, 1393–1512.

Sepehr, F., & Materassi, D. (2019).**An algorithm to learn polytree networks with hidden nodes.** *Proceedings of the Thirty-Third Annual Conference on Neural Information Processing Systems (NeuRIPS)*, 15084–15093.

Tramontano, D., Monod, A., & Drton, M. (2022).**Learning linear non-gaussian polytree models.** *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, (UAI)*, *180*, 1960–1969.

Wang, Y. S., Kolar, M., & Drton, M. (2025).**Confidence sets for causal orderings.** *Journal of the American Statistical Association*, (just-accepted), 1–25.

Zhao, Q. (2025).**On statistical and causal models associated with acyclic directed mixed graphs.** *arXiv preprint arXiv:2501.03048*.