Large-Scale Causal Structure Learning: Challenges and New Methods

Journée NormaSTIC 2025
Université de Caen Normandie

June 26, 2025

Shuyu Dong[1] (L2S, CentraleSupélec)

Joint work with Michèle Sebag[1], Kento Uemura[2], Akito Fujii[2], Shuang Chang[2], Yoseke Koyanagi[2], and Koji Maruhashi[2]

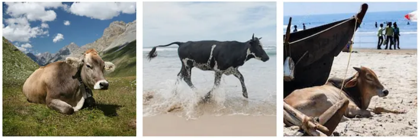
[1] *INRIA TAU team, LISN, Université Paris-Saclay*
[2] *Fujitsu Laboratories*

Outline

- 1. Causality, Causal Discovery, and Related Work
- 2. DCILP: A Distributed Approach
- 3. Conclusion and Perspectives

Causality in A Few Examples

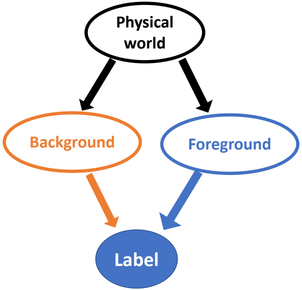
Image classification:



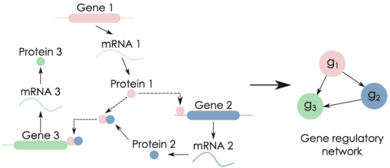
(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94



Biomedical sciences:



(Huynh-Thu and Sanguinetti, 2018)

Applications:

- Cell state engineering
- Drug discovery

Objective of this talk: Causal Structural Model ← Causal Structure Learning

Causal Structure Learning

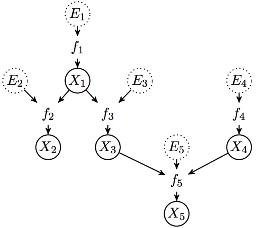
Definition (linear causal model): $X_i = \sum_{j=1}^{d} B_{ji} X_j + \epsilon_i$ for all $i = 1, \dots, d$.

B: weighted adjacency matrix of a **directed acyclic graph (DAG)** $\mathcal{G}$

ϵ_i: noise variable, $\epsilon_i \perp\!\!\!\perp X_j$ for all $j \in \mathbf{Pa}_B(i) := \{k \in [d] : B_{ki} \neq 0\}$

(e.g., Peters et al. (2017))

$$
\begin{aligned}
X_1 &= \epsilon_1 \\
X_2 &= B_{12} X_1 + \epsilon_2 \\
X_3 &= B_{13} X_1 + \epsilon_3 \\
X_4 &= \epsilon_4 \\
X_5 &= B_{35} X_3 + B_{45} X_4 + \epsilon_5
\end{aligned}
$$



$$
\begin{cases}
X_1 = f_1(E_1) \\
X_2 = f_2(X_1, E_2) \\
X_3 = f_3(X_1, E_3) \\
X_4 = f_4(E_4) \\
X_5 = f_5(X_3, X_4, E_5)
\end{cases}
$$

(Kalainathan et al., 2022)

Remarks:

- **Markov property**: $P(X_1, \dots, X_d) = \Pi_{i=1}^{d} P(X_i | \mathbf{Pa}_B(i))$.
- Acyclic and **sparse**: B is also a sparse matrix in most applications

Problem statement: Given samples $\mathcal{X}$ of $(X_1, \dots, X_d)$, learn a DAG matrix B that best fits $\mathcal{X}$.

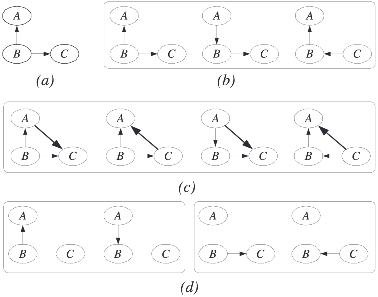
Causal Structure Learning: Related Work

Discrete methods: maximum likelihood within the set of DAGs:

- **Acyclicity** is a complex combinatorial constraint (NP-hardness (Chickering, 1996)).
- Minimize $f(B) = -\log p(B; \mathcal{X})$ by *enumerating* different DAGs $\rightsquigarrow$ **combinatorial problem**
- *Learning* the cofficients for the nonzeros of B $\rightsquigarrow$ **continuous optimization**

GES algorithm (Chickering, 2002): greedy search to maximize the Bayesian information criterion (BIC)

$$S(\mathcal{G}; \mathcal{X}) = \log p(\mathcal{X} | \mathcal{G}, \widehat{\theta}) - \frac{d}{2} \log(n).$$



Causal Structure Learning: Related Work

Discrete methods: **maximum likelihood** within the set of DAGs:

- **Acyclicity** is a complex combinatorial constraint (NP-hardness (Chickering, 1996)).
- Minimize $f(B) = -\log p(B; \mathcal{X})$ by *enumerating* different DAGs $\rightsquigarrow$ **combinatorial problem**
- *Learning* the cofficients for the nonzeros of B $\rightsquigarrow$ **continuous optimization**

Continuous optimization:

Theorem (Zheng et al., 2018): *The graph of $B \in \mathbb{R}^{d \times d}$ is a DAG if and only if*

$$h(B) := \mathrm{tr}(\exp(B \odot B)) - d = 0.$$

Non-combinatorial Optimization **NOTEARS** (Zheng et al., 2018)

$$\min_{B \in \mathbb{R}^{d \times d}} \quad f(B) + \lambda \|B\|_{\ell_1} \qquad \Leftrightarrow \qquad \min_{B \in \mathbb{R}^{d \times d}} \quad f(B) + \lambda \|B\|_{\ell_1}$$

$$\text{s.t.} \quad \mathrm{tr}(\exp(B \odot B)) - d = 0 \qquad\qquad \text{s.t.} \quad B \in \mathbf{DAG}(d)$$

- $h(B)$ continuous and differentiable
- Cost: function evaluation of $B \to \mathrm{tr}(\exp(B \odot B))$ and its gradients $\rightsquigarrow$ $O(d^3)$
- Augmented Lagrangian method ... Nonconvex nonsmooth problem

Continuous Optimization Methods

Theorem (Zheng et al., 2018): *The graph of $B \in \mathbb{R}^{d \times d}$ is a DAG if and only if*

$$h(B) := \operatorname{tr}(\exp(B \odot B)) - d = 0.$$

Proof (sketch): For $\mathbb{B} \in \{0,1\}^{d \times d}$ and any $k \geq 1$,

$$\operatorname{tr}(\mathbb{B}^k) = \text{amount of } k\text{-cycles.}$$

Total amount of all cycles:

$$\operatorname{tr}(\exp(\mathbb{B})) = \operatorname{tr}\left(I + \sum_{k \geq 1} \frac{1}{k!} \mathbb{B}^k\right) = d + \sum_{k \geq 1} \frac{1}{k!} \operatorname{tr}(\mathbb{B}^k).$$

Trick to generalize $\mathbb{B}$ to weighted adjacency B: the Hadamard product where $(B \odot B)_{ij} = B_{ij}^2$. □

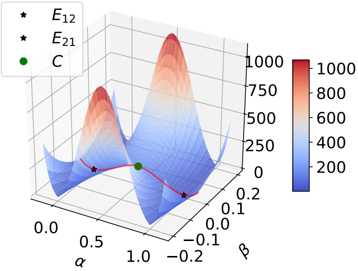
Zheng et al. (2018): DAGs with NOTEARS: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, volume 31.

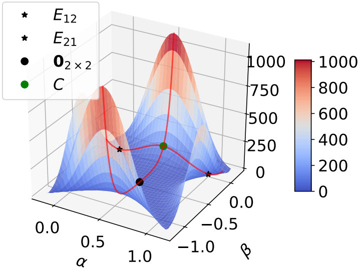
Continuous Optimization Methods

Theorem (Zheng et al., 2018): *The graph of $B \in \mathbb{R}^{d\times d}$ is a DAG if and only if*

$$h(B) := \mathrm{tr}(\exp(B \odot B)) - d = 0.$$

Landscape of $h(B)$ near $C = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}$, in two different subspaces of $\mathbb{R}^{2\times 2}$:





$$E_{12} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad E_{21} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

Continuous Optimization Methods

Non-combinatorial Optimization **NOTEARS** (Zheng et al., 2018)

$$\min_{B\in\mathbb{R}^{d\times d}} \quad f(B)+\lambda\|B\|_{\ell_1} \qquad \Leftrightarrow \qquad \min_{B\in\mathbb{R}^{d\times d}} \quad f(B)+\lambda\|B\|_{\ell_1}$$

$$\text{s.t.} \quad \operatorname{tr}(\exp(B\odot B))-d=0 \qquad\qquad \text{s.t.} \quad B\in \mathbf{DAG}(d)$$

The continuous opt. approach may induce **heavy bias** in the estimated causal order!

(*Var-sortability bias* (Reisach et al., 2021))

$$\begin{aligned} X_1 &= \epsilon_1\\ X_4 &= \epsilon_4\\ X_2 &= B_{12}X_1+\epsilon_2\\ X_3 &= B_{13}X_1+\epsilon_3\\ X_5 &= B_{35}X_3+B_{45}X_4+\epsilon_5 \end{aligned}$$

- Homogeneous data ($\mathrm{var}(\epsilon_i)$ **equal**):
  Order of $\{\mathrm{var}(X_i)\}_{i=1,\ldots,d}$ consistent with causal order
- Heteorgeneous data ($\mathrm{var}(\epsilon_i)$ **non-equal**):
  Order of $\{\mathrm{var}(X_i)\}_{i=1,\ldots,d}$ no longer consistent $\rightsquigarrow$ bias through the gradient $\nabla f(B)$

Challenges in Causal Structure Learning

For continuous optimization methods:

- Nonconvexity
- Heavy bias on **heterogeneous** data

For discrete & graphical methods:

- **Acyclicity** is a complex combinatorial constraint (NP-hardness (Chickering, 1996)).
- The set of DAGs is **huge**!

The size of $\mathbb{DAG}(d) := \{B \in \{0,1\}^{d \times d} : \mathcal{G}(B) \text{ is a DAG}\}$ grows as

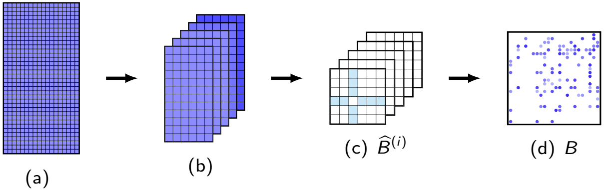
$$|\mathbb{DAG}(d)| \approx d! 2^{d^2/2}.$$

Outline

Divide-and-Conquer in Three Phases

DCILP (Dong et al., 2025): Distributed causal discovery using ILP

1. **Phase 1**: divide $\boldsymbol{X} = (X_1, \ldots, X_d)$ into different subsets $\mathbf{S}_1, \mathbf{S}_2, \ldots$
2. **Phase 2**: learn subgraph from data restricted to $\mathbf{S}_i$ separately
3. **Phase 3**: aggregate subgraphs



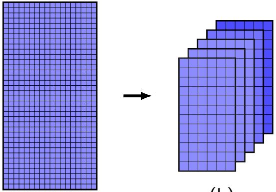
How it differs with the related work (Gao et al., 2017; Gu and Zhou, 2020; Mokhtarian et al., 2021):

- Phase 2: **parallel** instead of sequential
- Phase 3: **integer programming**-based instead of rule-based

---

Dong et al. (2025): SD, M. Sebag, K. Uemura, A. Fujii, S. Chang, Y. Koyanagi, K. Maruhashi. DCILP: a distributed approach for large-scale causal structure learning. In *the 39th Annual AAAI Conference on Artificial Intelligence* (AAAI-25). URL https://doi.org/10.1609/aaai.v39i15.33795.

DCILP Phase-1: Divide by Markov Blankets



(a) (b)

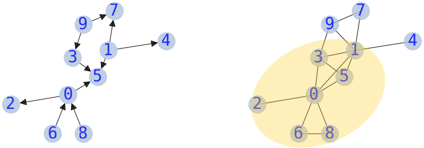
Definition (e.g., Peters et al. (2017)): The Markov blanket $\mathbf{MB}(X_i)$ of a variable X_i is the smallest set $M \subset \boldsymbol{X}$ such that

$$X \perp\!\!\!\perp \boldsymbol{X} \backslash (M \cup \{X_i\}) \text{ given } M.$$

Property (example of $\mathbf{MB}(X_0)$)

- **Parent** nodes: X_6, X_8
- **Children** nodes: X_2, X_5
- **Spouse** nodes: X_3, X_1

$\boldsymbol{B}$ $\qquad$ $\boldsymbol{\mathcal{M}(B)}$



Theorem (Loh and Bühlmann, 2014): *Under a faithfulness assumption, the Markov blankets can be identified via the support of the precision matrix:* $\mathcal{M}(B) = \mathrm{Supp}((\mathrm{cov}(\boldsymbol{X}))^{-1})$.

DCILP Phase-2: Parallel Computing

Algorithm 1 (DCILP) Distributed causal discovery using ILP

(Phase-1) Divide:

$$\text{Estimate Markov blanket } \mathbf{MB}(X_i) \text{ for } i = 1, \ldots, d$$

(Phase-2) for $i = 1, \ldots, d$ **do in parallel**

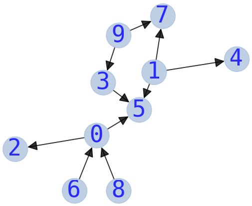
  $A^{(i)} \leftarrow$ Causal discovery on $\mathbf{S}_i := \mathbf{MB}(X_i) \cup \{X_i\}$ # using GES (Chickering, 2002) or # DAGMA (Bello et al., 2022)

  $\widehat{B}^{(i)}_{j,k} \leftarrow A^{(i)}_{j,k}$ if $j = i$ or $k = i$, and 0 otherwise

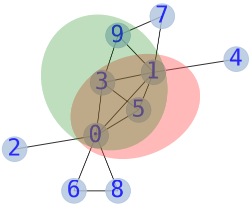
(Phase-3) Conquer:

$$B \leftarrow \text{Reconciliation from } \{\widehat{B}^{(i)}, i = 1 \ldots d\} \text{ through the ILP}$$

$\boldsymbol{B}^\star$



Phase-2 on estimated $\boldsymbol{\mathcal{M}}$ (all MBs)



DCILP Phase-3

Algorithm 1 (DCILP) Distributed causal discovery using ILP

(Phase-1) Divide:

Estimate Markov blanket $\mathbf{MB}(X_i)$ for $i = 1, \ldots, d$

(Phase-2) for $i = 1, \ldots, d$ **do in parallel**

$A^{(i)} \leftarrow$ Causal discovery on $\mathbf{S}_i := \mathbf{MB}(X_i) \cup \{X_i\}$ # using GES (Chickering, 2002) or # DAGMA (Bello et al., 2022)

$\widehat{B}^{(i)}_{j,k} \leftarrow A^{(i)}_{j,k}$ if $j = i$ or $k = i$, and 0 otherwise

(Phase-3) Conquer:

$B \leftarrow$ Reconciliation from $\{\widehat{B}^{(i)}, i = 1 \ldots d\}$ through the ILP

Question: how to **aggregate all** the subgraphs $\widehat{B}^{(i)}$?

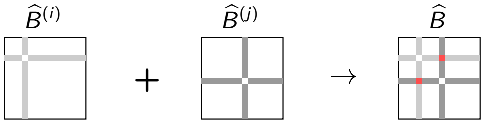
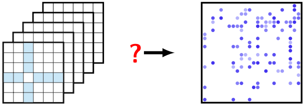


Figure: Merge conflict in concatenation of two local results.

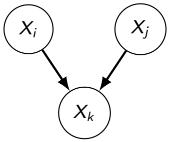


(a) $\widehat{B}^{(i)}$ (b) B

DCILP Phase-3: Causal Structure in Binary Variables

Idea: *Correct* the edges in $\widehat{B} = \sum_i \widehat{B}^{(i)}$ with respect to all the Markov blankets $\mathcal{M}$ (assume $\mathcal{M} = \mathcal{M}(B^\star)$)

- A **necessary condition** for a candidate B is: $\boldsymbol{\mathcal{M}(B) = \mathcal{M}(B^\star)}$.
  How to: auxiliary variables depending on B_{ij}
  - $B_{ij} = 1$ if $X_i \to X_j$.
  - $V_{ijk} = V_{jik} = 1$ if there is a v-structure $(X_i \to X_k \leftarrow X_j)$
  - $S_{ij} = S_{ji} = 1$ if X_i and X_j are spouses, i.e., $\exists k$, $V_{ijk} = 1$.



- **Consistency** among the variables B, S and V

Our discovery: $\mathcal{M}(B) = \mathcal{M}(B^\star)$ can be translated into binary linear constraints on (B, S, V).

DCILP Phase-3: the ILP Formulation

$$\max_{B,S,V} \langle B, \sum_{i=1}^{d} \widehat{B}^{(i)} \rangle \quad \text{subject to}$$

Ensure $\mathcal{M}(B) = \mathcal{M}(B^\star)$

$$B_{ij} = 0, \quad S_{ij} = S_{ji} = 0 \qquad \text{if } X_i \notin \mathbf{MB}(X_j) \qquad (1)$$

$$B_{ij} + B_{ji} + S_{ij} \geq 1 \qquad \text{if } X_i \in \mathbf{MB}(X_j) \qquad (2)$$

$$B_{ij} + B_{ji} \leq 1 \qquad \text{if } X_i \in \mathbf{MB}(X_j) \qquad (3)$$

Consistency of (B, S, V)

$$V_{ijk} \leq B_{ik}, \ V_{ijk} \leq B_{jk}, \qquad \text{if } \{i,j,k\} \subset (\mathbf{S}_i \cap \mathbf{S}_j \cap \mathbf{S}_k) \qquad (4)$$

$$B_{ik} + B_{jk} \leq 1 + V_{ijk}, \qquad \text{if } \{i,j,k\} \subset (\mathbf{S}_i \cap \mathbf{S}_j \cap \mathbf{S}_k) \qquad (5)$$

$$V_{ijk} \leq S_{ij}, \ S_{ij} \leq \sum_k V_{ijk} \qquad \text{if } \{i,j,k\} \subset (\mathbf{S}_i \cap \mathbf{S}_j \cap \mathbf{S}_k) \qquad (6)$$

for all i, j, k such that $i \neq j$, $j \neq k$, $k \neq i$: $\quad (\mathbf{S}_i := \mathbf{MB}(X_i) \cup \{X_i\})$

Proposition: *Under the Markov property assumption (distribution of* $\boldsymbol{X}$ *agreeing with* $B^\star$*): given the correct* **MB***s, the sought causal graph* $B^\star$ *and the underlying structures* $(S^\star, V^\star)$ *satisfy the ILP constraints (1)–(6).*

DCILP: Experiments

Algorithm 1 (DCILP) Distributed causal discovery using ILP

1: **(Phase-1) Divide**:

Estimate Markov blanket $\mathbf{MB}(X_i)$ for $i = 1, \ldots, d$

2: **(Phase-2) for** $i = 1, \ldots, d$ **do in parallel**

3: $A^{(i)} \leftarrow$ Causal discovery on $\mathbf{S}_i := \mathbf{MB}(X_i) \cup \{X_i\}$ using GES or DAGMA (Bello et al., 2022)

4: $\widehat{B}^{(i)}_{j,k} \leftarrow A^{(i)}_{j,k}$ if $j = i$ or $k = i$, and 0 otherwise

5: **(Phase-3) Conquer:**

$B \leftarrow$ Reconciliation from $\{\widehat{B}^{(i)}, i = 1 \ldots d\}$ through the ILP

- *Phase 1*: empirical precision matrix estimator
- *Phase 2*: Parallellized on $\min(2d, 400)$ CPU cores.
  Running on Ruche (Mesocentre Paris-Saclay)
- *Phase 3*: implementation with Gurobi tools

DCILP - Experiments: ILP versus the Naive Merge

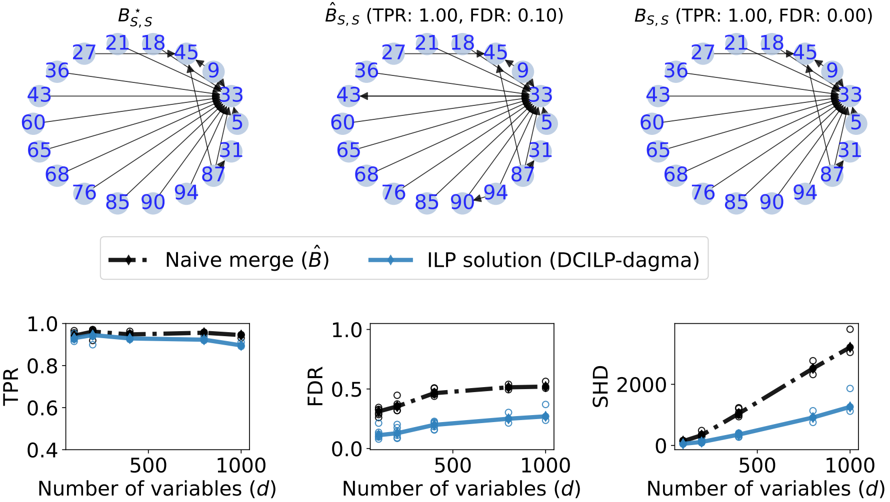


Figure: Comparing with the naive merge $\widehat{B}$: DCILP on SF3 data.

DCILP: Experiments

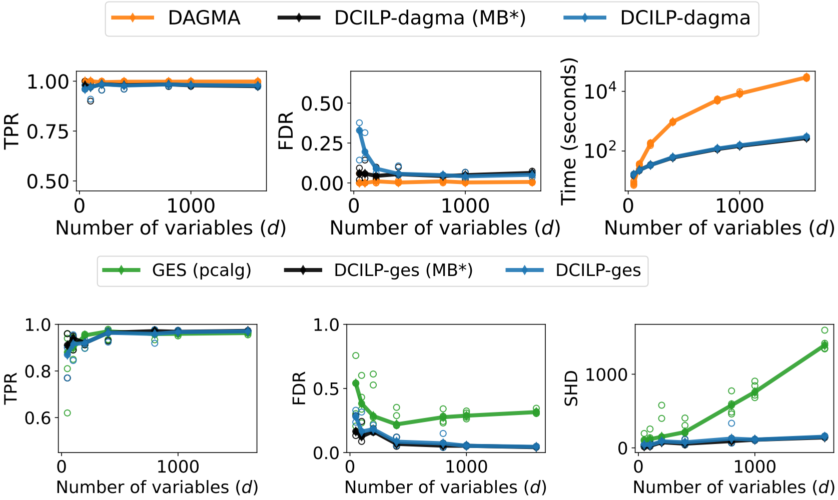


Figure: Comparison with DAGMA (Bello et al., 2022) and GES (Chickering, 2002) on ER2 data.

DCILP: Experiments

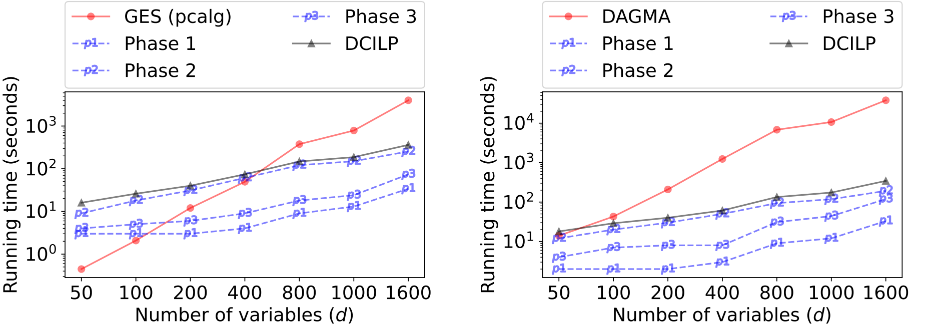
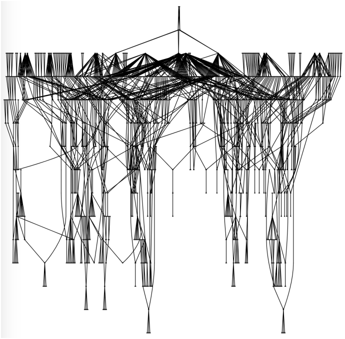


Figure: Running time comparisons with GES and DAGMA.

DCILP: Experiments on MUNIN network

- A DAG with $d = 1041$ nodes (`https://www.bnlearn.com/bnrepository/`)
- Medical expert-system model based on electromyographs (EMG)



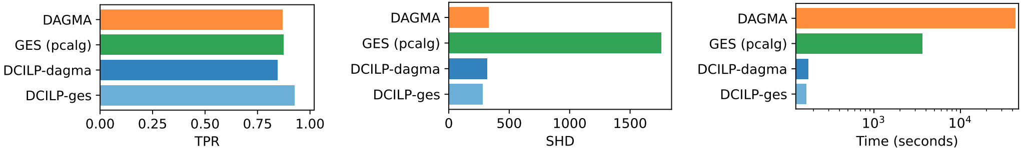


Figure: Results on the MUNIN network data.

Outline

Conclusion

- A **distributed approach**: DCILP leverages parallel computing while ensuring an optimized merge of local solutions via an ILP-based algorithm.
- **Modularity**: DCILP allows for new alternative subroutines for Phase 1 and Phase 2.
- Significant improvement in scalability for learning sparse and large causal graphs.

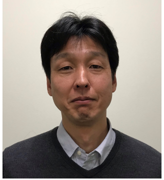
Perspectives:

- Extend applicability:
  - Nonlinear models
  - Robustness to change of scales in the measurements/observations
- Adapt to the learning of denser causal graphs
- Causal modeling with latent variables

Acknowledgement

TAU, INRIA Saclay

Fujitsu Laboratories



Thank You !

References

Bello, K., Aragam, B., and Ravikumar, P. (2022). DAGMA: Learning dags via m-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35:8226–8239.

Chickering, D. M. (1996). Learning bayesian networks is NP-complete. *Learning from data: Artificial intelligence and statistics V*, pages 121–130.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

Dong, S., Sebag, M., Uemura, K., Fujii, A., Chang, S., Koyanagi, Y., and Maruhashi, K. (2025). DCILP: A distributed approach for large-scale causal structure learning. *The 39th Annual AAAI Conference on Artificial Intelligence (AAAI-25)*.

Gao, T., Fadnis, K., and Campbell, M. (2017). Local-to-global bayesian network structure learning. In *International Conference on Machine Learning*, pages 1193–1202. PMLR.

Gu, J. and Zhou, Q. (2020). Learning big gaussian bayesian networks: Partition, estimation and fusion. *Journal of machine learning research*, 21(158):1–31.

Huynh-Thu, V. A. and Sanguinetti, G. (2018). Gene regulatory network inference: an introductory survey. In *Gene regulatory networks: Methods and protocols*, pages 1–23. Springer.

References (cont.)

Kalainathan, D., Goudet, O., Guyon, I., Lopez-Paz, D., and Sebag, M. (2022). Structural agnostic modeling: Adversarial learning of causal graphs. *Journal of Machine Learning Research*, 23(219):1–62.

Loh, P.-L. and Bühlmann, P. (2014). High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1):3065–3105.

Mokhtarian, E., Akbari, S., Ghassami, A., and Kiyavash, N. (2021). A recursive Markov boundary-based approach to causal structure learning. In *The KDD'21 Workshop on Causal Discovery*, pages 26–54. PMLR.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Reisach, A. G., Seiler, C., and Weichwald, S. (2021). Beware of the simulated dag! causal discovery benchmarks may be easy to game. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27772–27784.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31.