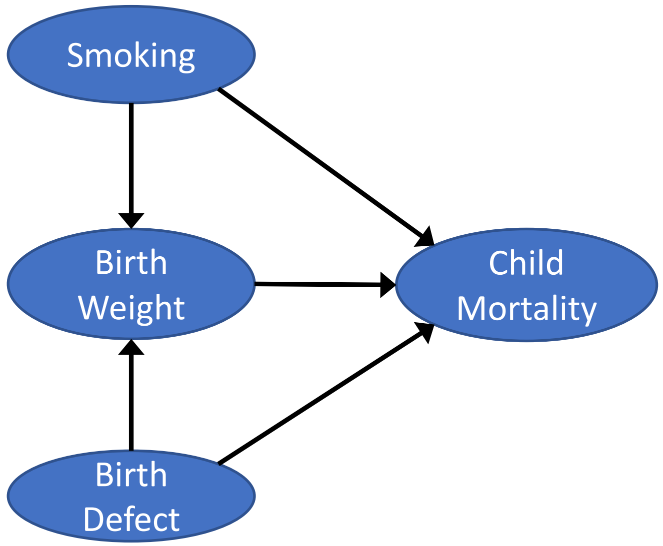
Rensselaer

Differentially Private Causal Discovery

Burak Varıcı

April 8, 2022



- Directed Acyclic Graph (DAG): cause-effect relationships
- **Causal discovery (structure learning):**
  - Learning the structure of this graph.
  - Answer causal questions reading the graph.
- Trustworthiness? Better reasoning/explainability than association models.
- **Constraint-based methods** and score-based methods
- Conditional Independence (CI) tests to rule out edges. Data privacy concerns.

How to perform differentially private causal discovery?

Rensselaer

- PC algorithm [1]: Well-known constraint-based algorithm for **causally sufficient models**.
- EM-PC algorithm [2]: Exponential mechanism to ensure privacy of CI tests. Inefficient though.
- **Priv-PC [3]:** More recent and successful algorithm. Provides simple theoretical results.
  - Do not directly use the p-value of CI tests. Privatize the process.
  - Sub-sampled sparse vector technique (SVT): filter out unlikely edges with little privacy cost.
  - After pruning process, use Laplace mechanism to check the remaining edges with larger privacy budget.
- Latent confounders = > **causally insufficient models?** Not explored yet.
- **FCI algorithm [4]:** Counterpart of PC for causally insufficient models.

Apply privatization technique of Priv-PC to FCI to derive **Priv-FCI**.

FCI algorithm to DP Priv-FCI

- Learn skeleton with CI tests.
  - Start from complete undirected graph.
  - Test edge for (i, j) by conditioning on $S \subset adj(i)\ or\ adj(j)$.
  - If $i \perp j \mid S$, delete $i - j$ edge, record S to $sepset(i, j)$ and $sepset(j, i)$
- Orient V-structures (no CI tests)
- Update skeleton with CI tests.
  - Test edge for (i, j) by conditioning on $S \subset posdsep(i)\ or\ posdsep(j)$.
  - If $i \perp j \mid S$, delete $i - j$ edge, record S to $sepset(i, j)$ and $sepset(j, i)$
- Orient V-structures (no CI tests)
- Apply orientation rules (no CI tests)

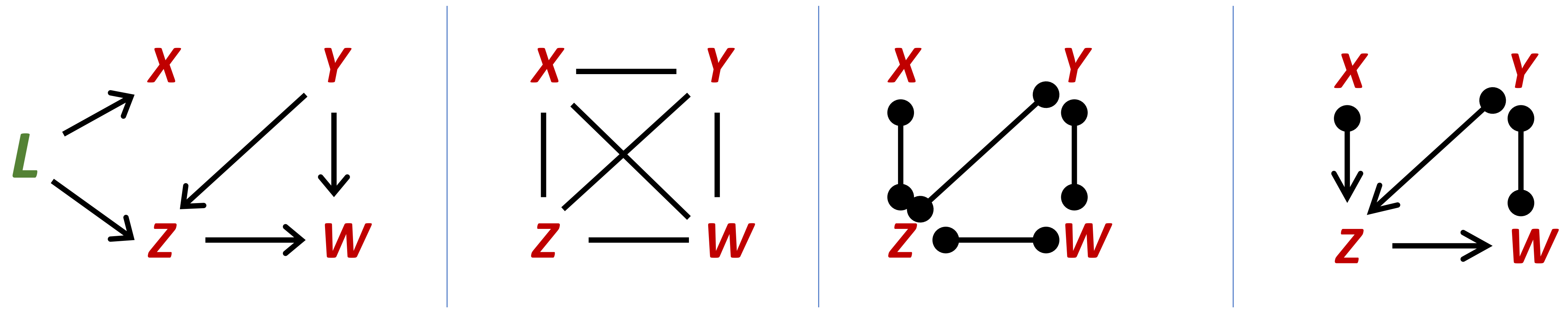
Pseudo-code for FCI algorithm

- If we make $i \perp j \mid S$ decision mechanism differentially private, we obtain differentially private **Priv-FCI** algorithm .

- Pick a CI test: conditional Spearman's ρ, χ^2-test, Kendall's τ.
- Standard (non-private) decision mechanism for variables i, j and set S.

$$\text{If } \tau(i, j \mid S) \geq \alpha \quad \rightarrow \quad i \perp j \mid S$$

e.g. $X \perp_d W \mid Y, Z$



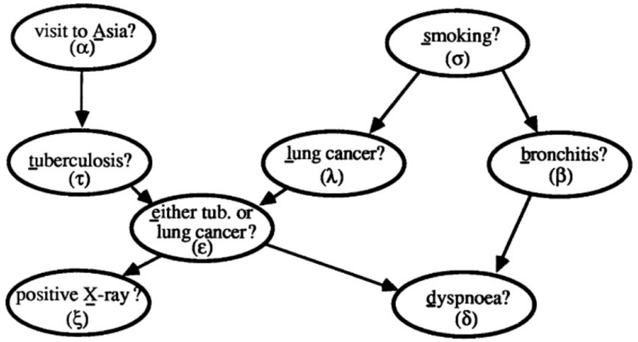
Edge types 

sieve-and-examine procedure of Priv-PC

- Sparse vector technique (SVT): filter out unlikely edges with little privacy cost.
- q:subsampling ratio, ϵ:privacy parameter, t:threshold tweak, Δ:sensitivity on full dataset
- SVT privacy cost: $\epsilon' = \ln(\frac{e^{\frac{\epsilon}{2}}-1}{q} + 1)$, private threshold: $\alpha' = \alpha - t + Lap(\frac{2\Delta}{\sqrt{q\epsilon'}})$.
- Private decision mechanism
- If $\tau(i,j \mid S) + Lap\left(\frac{4\Delta}{\sqrt{q\epsilon'}}\right) \geq \alpha'$ (pruning part, only few edges will pass)
  - If $\tau(i,j \mid S) + Lap\left(\frac{2\Delta}{\epsilon}\right) \geq \alpha \quad \rightarrow \quad i \perp j \mid S$.
  - Resample data and reset private threshold $\alpha' = \alpha - t + Lap(\frac{2\Delta}{\sqrt{q\epsilon'}})$.
- Recall definition: $\mathbb{P}[\mathcal{A}(\mathcal{D}) \in \mathcal{S}] \leq e^{\epsilon}\mathbb{P}[\mathcal{A}(\mathcal{D}') \in \mathcal{S}] + \delta$
- One shot of this process is ϵ-private. Advanced composition theorem for total privacy cost.

Experiments

- Benchmark synthetic datasets from the literature.
  - Earthquake [5], Cancer [5], Asia [6], Survey [7].
- Output of the algorithm: Partial Ancestral Graphs.
- Metric: F1 scores for edge recovery:
  - Skeleton: compare to true edges without regarding edge orientation.
  - Orientation: also check the orientation of the edge ($\circ\!-\!\circ, \circ\!\rightarrow, \rightarrow, \leftrightarrow$).
- No competitive methods exist for causally insufficient models.
- Compare with non-private version, classical FCI.

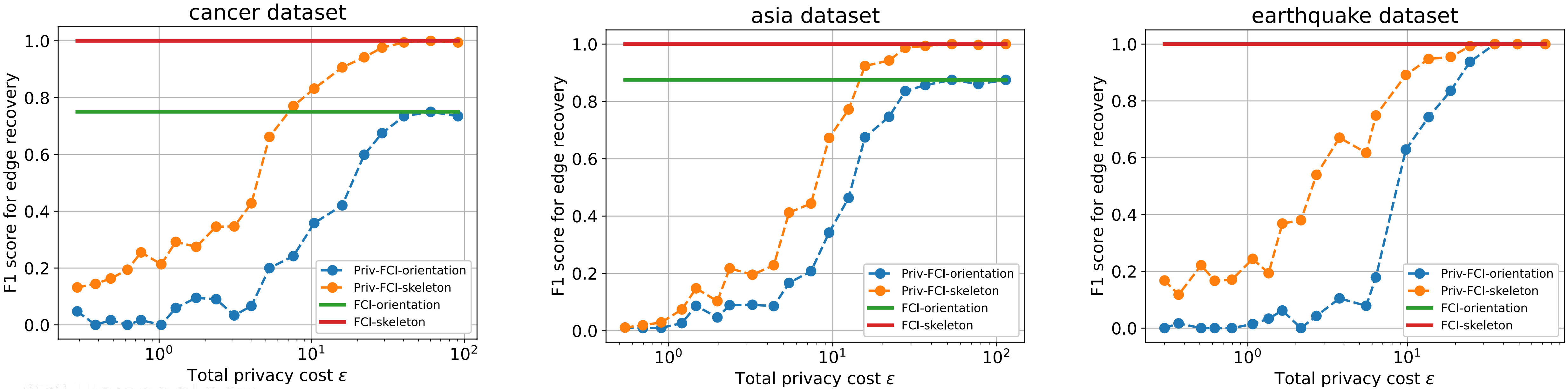


Ground truth causal graph for Asia dataset.

Rensselaer

Experiments

- Priv-FCI algorithm is run for different values of privacy constraints.
- Experiments are repeated 20 times.
- As expected, performance gets closer to non-private FCI as privacy budget grows.



Experiments

- Priv-FCI algorithm is run for different values of privacy constraints.
- Feasible runtime (disclaimer: very small models).

Dataset	# nodes	# edges	Type	Runtime (s)
Earthquake	5	4	Binary	1.46
Cancer	5	4	Binary	1.43
Asia	8	10	Binary	4.59
Survey	6	6	Discrete	1.02

Conclusion and Future Directions

- Extension of Priv-PC to Priv-FCI is indeed possible.
- Runtime of the algorithm, which is one benefit of Priv-PC over previous work, is still reasonable for Priv-FCI.
- Observed privacy costs of accurate DP-causal discovery is still high.
- How to integrate CI tests with infinite sensitivity like χ^2-test?
- Some tricks like tweaking a bias (not included here): no theoretical analysis.

References

[1] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. Causation, prediction, and search. MIT press, 2000.

[2] Depeng Xu, Shuhan Yuan, and Xintao Wu. Differential privacy preserving causal graph discovery, In 2017 IEEE Symposium on Privacy-Aware Computing (PAC), pages 60–71. IEEE, 2017.

[3] Wang, L., Pang, Q., & Song, D. (2020). Towards practical differentially private causal graph discovery. *Advances in Neural Information Processing Systems*, *33*, 5516-5526.

[4] Spirtes, P. (2001, January). An anytime algorithm for causal inference. In *International Workshop on Artificial Intelligence and Statistics* (pp. 278-285). PMLR.

[5] Kevin B Korb and Ann E Nicholson. Bayesian artificial intelligence. CRC press, 2010.

[6] Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society. 50(2):157–194, 1988.

[7] Marco Scutari and Jean-Baptiste Denis. Bayesian networks: with examples in R. CRC press, 2014.