Testing Causal Models with Hidden Variables in Polynomial Delay via Conditional Independencies




Hyunchai Jeong*[1] , Adiba Ejaz*[2], Jin Tian[3] and Elias Bareinboim[2]

[1]Purdue University, [2]Columbia University, [3]Mohamed bin Zayed University of Artificial Intelligence

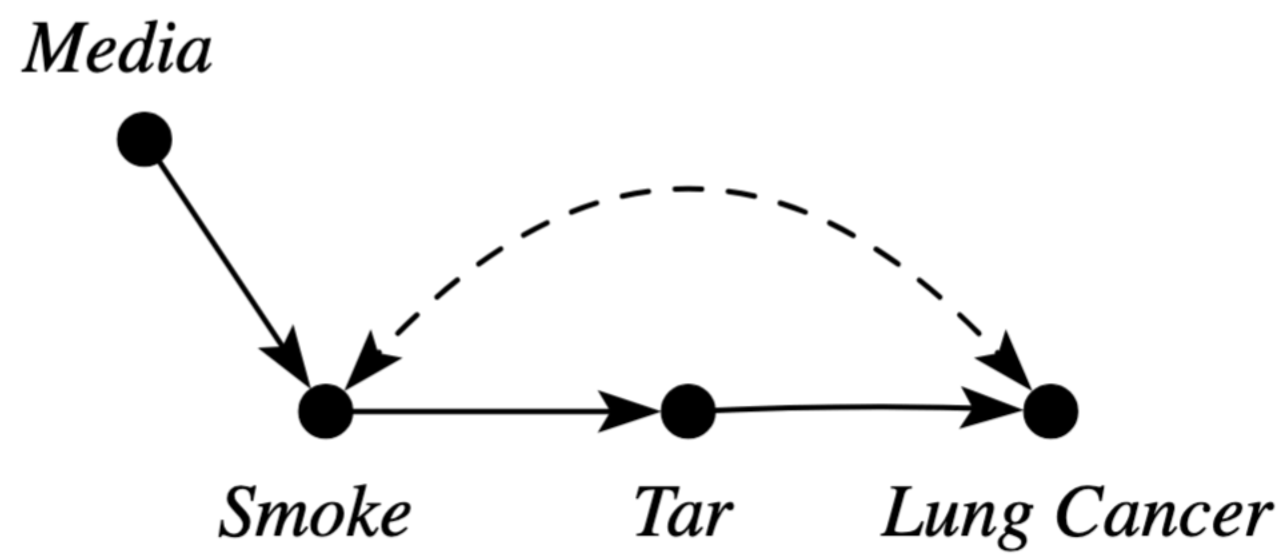
The Thirty-Ninth AAAI Conference on Artificial Intelligence, 2025

Causal Effect Identification

① Query

$P(lc \mid do(s))$

② Causal Graph



③ Data

$P(m, s, lc, t)$

Causal Inference Engine

Identifiable?

Yes

No

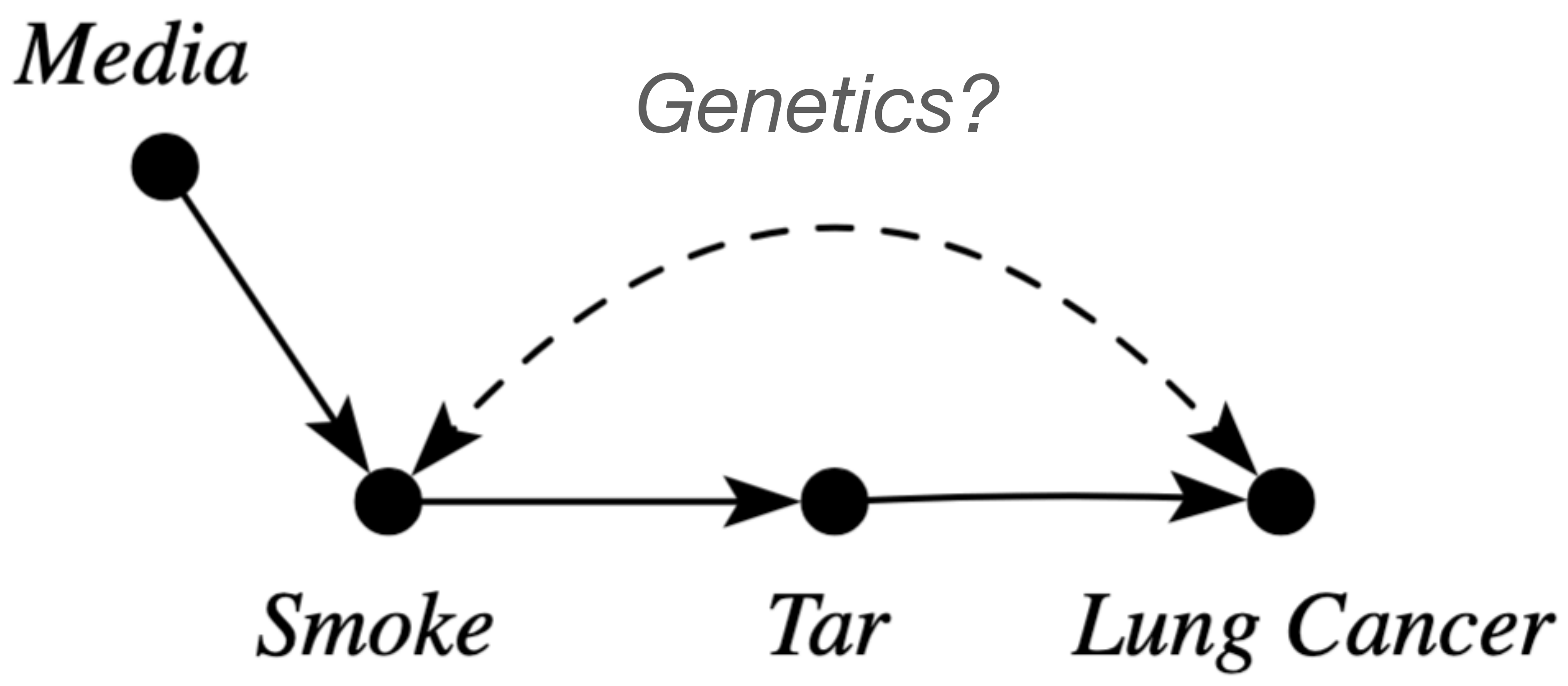
How do we test this?

$$P(lc \mid do(s)) = \sum_t P(t \mid s) \sum_{s'} P(lc \mid s', t) P(s')$$

Causal Graphs

- Causal graphs encode causal assumptions needed for inference (Bareinboim et al. 2020).
- $Smoke \rightarrow Tar$: $Smoke$ causes Tar.
- $Smoke \leftrightarrow Lung\ Cancer$: $Smoke$ and $Lung\ Cancer$ have unobserved confounding.

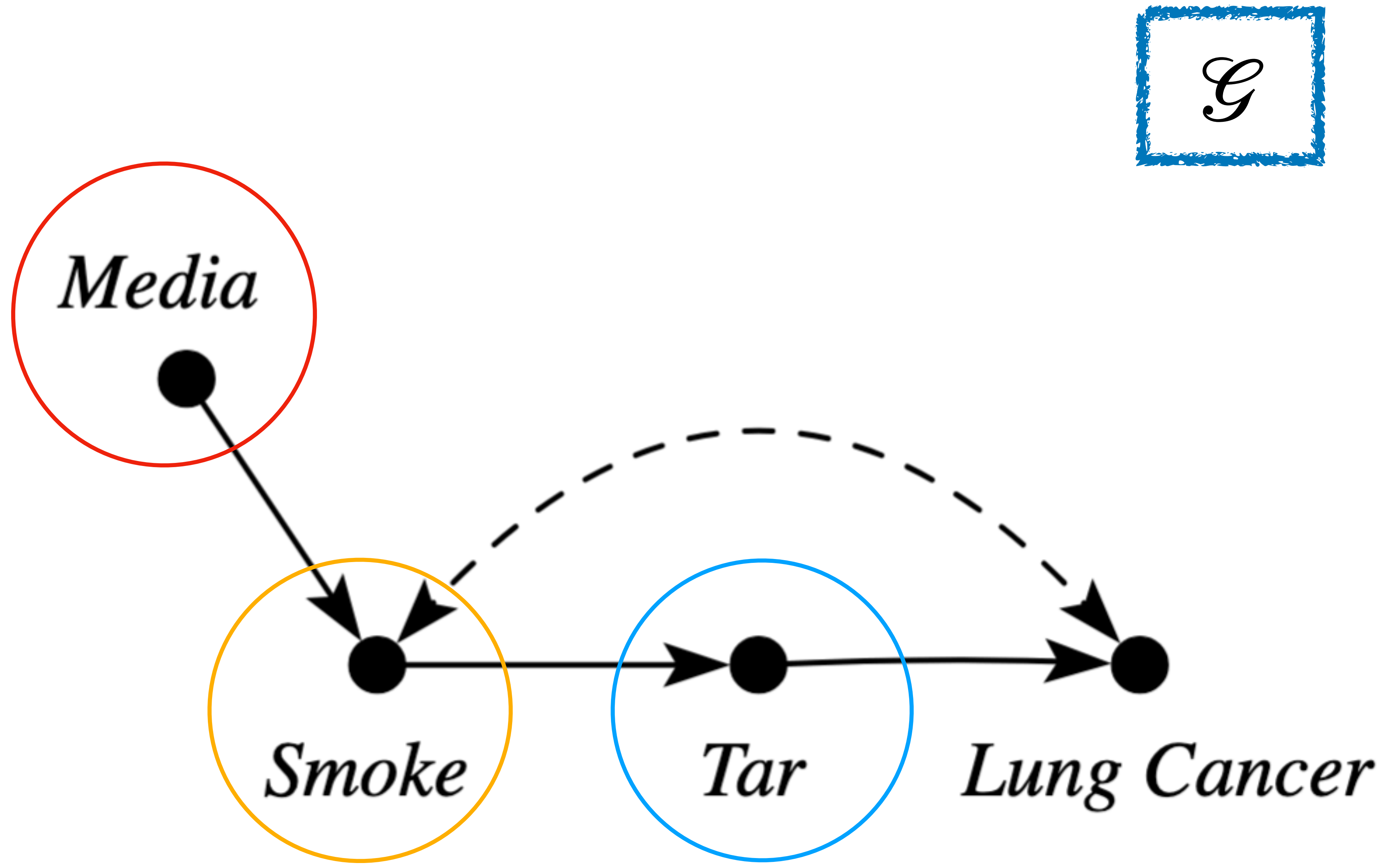
$\mathcal{G}$



Motivation

How do we evaluate the causal graph using data?

Test if conditional independencies (CIs) implied by the model hold in the observational data!

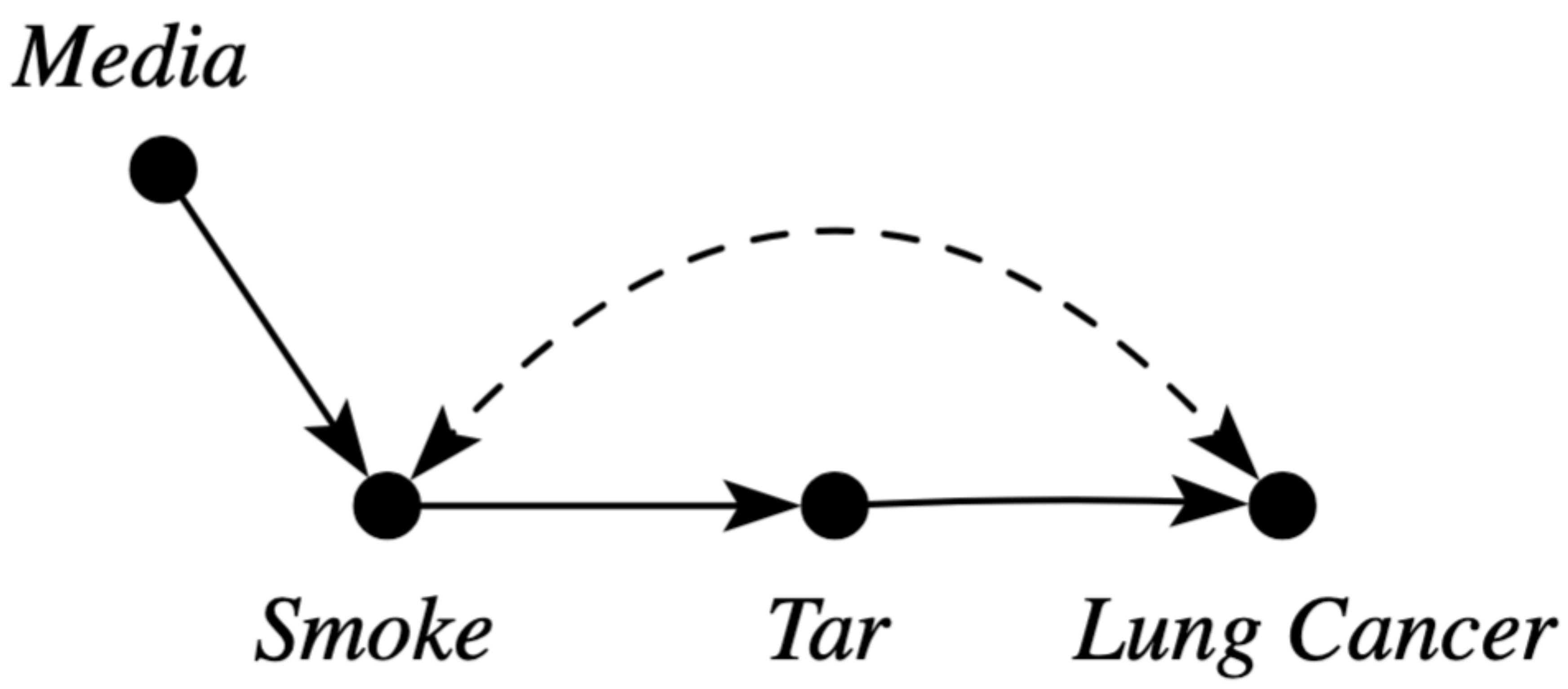


$Media \perp Tar \mid Smoke.$

Motivation

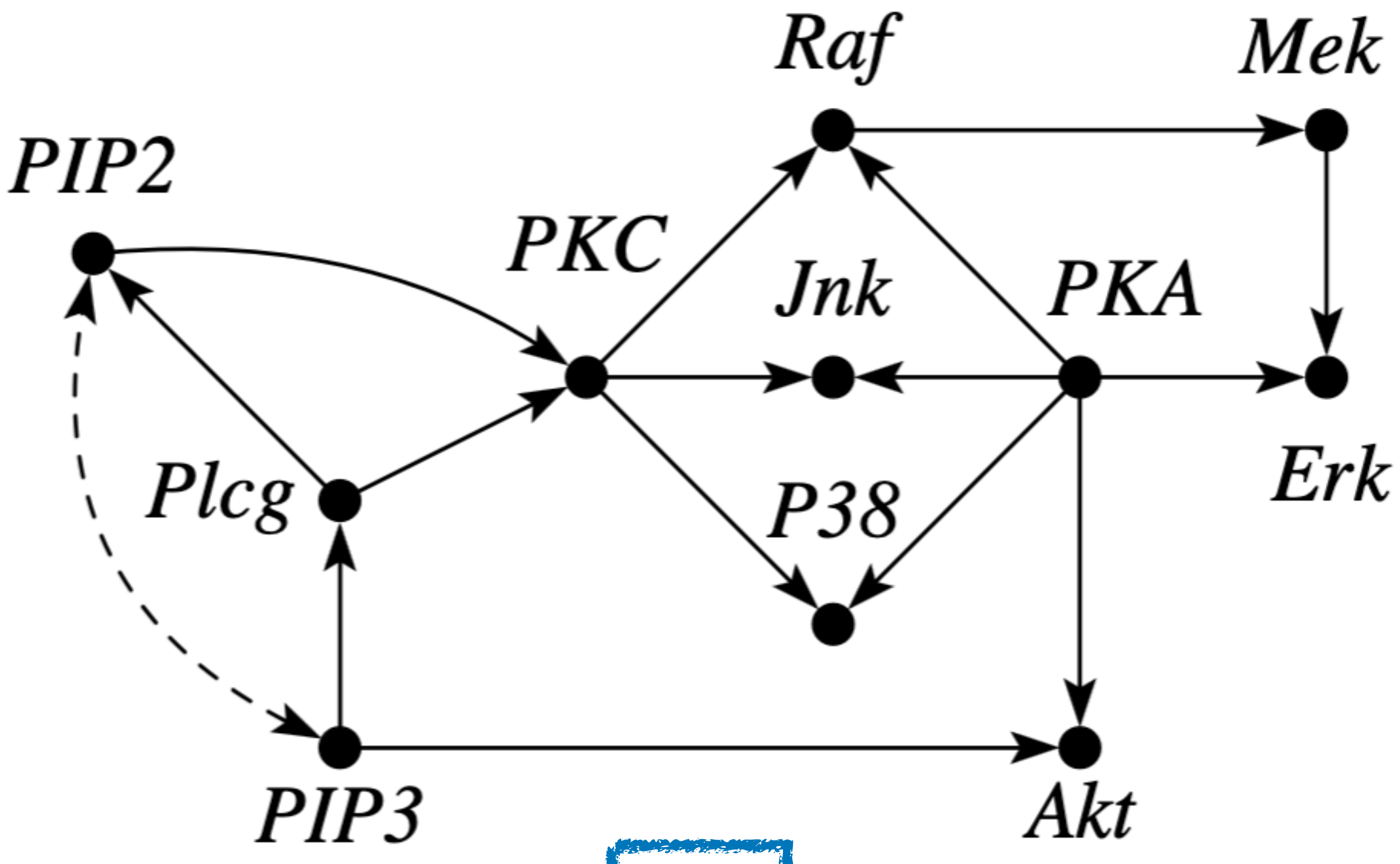
How many CIs do we need to test?

$n = 4, m = 4$. 1 CI encoded in $\mathscr{G}$



$\mathscr{G}$

$n = 11, m = 16$. 76580 CIs encoded in $\mathscr{G}^S$



$\mathscr{G}^S$

Protein signalling network (Sachs et al. 2005)

Sachs, K.; Perez, O.; Pe'er, D.; Lauffenburger, D. A.; and Nolan, G. P. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. Science, 308(5721): 523–529.

Motivation

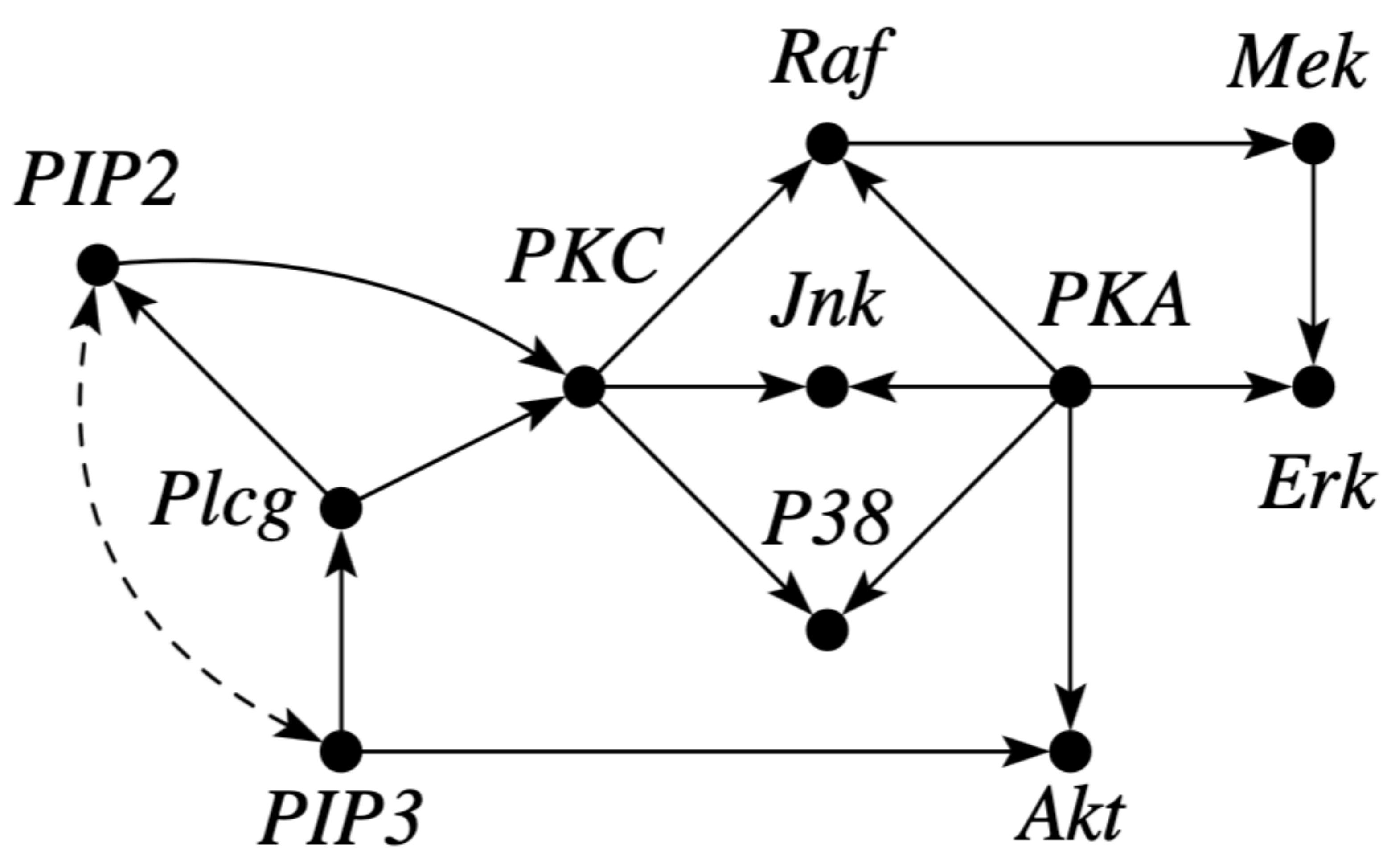
- Global Markov property (GMP): $\Theta(4^n)$ CIs.

Are we stuck?

- Local Markov property (LMP) : a 'basis' of CIs which imply all others.

$\mathscr{G}$ consistent with GMP $\iff$ $\mathscr{G}$ consistent with LMP

$\mathscr{G}$



$n = 11$, $m = 16$. 76580 CIs encoded in $\mathscr{G}$!

10 CIs imply all other 76570 CIs.

Local Markov Property

Definition 1 (LMP (Pearl 88, Lauritzen et al. 90))

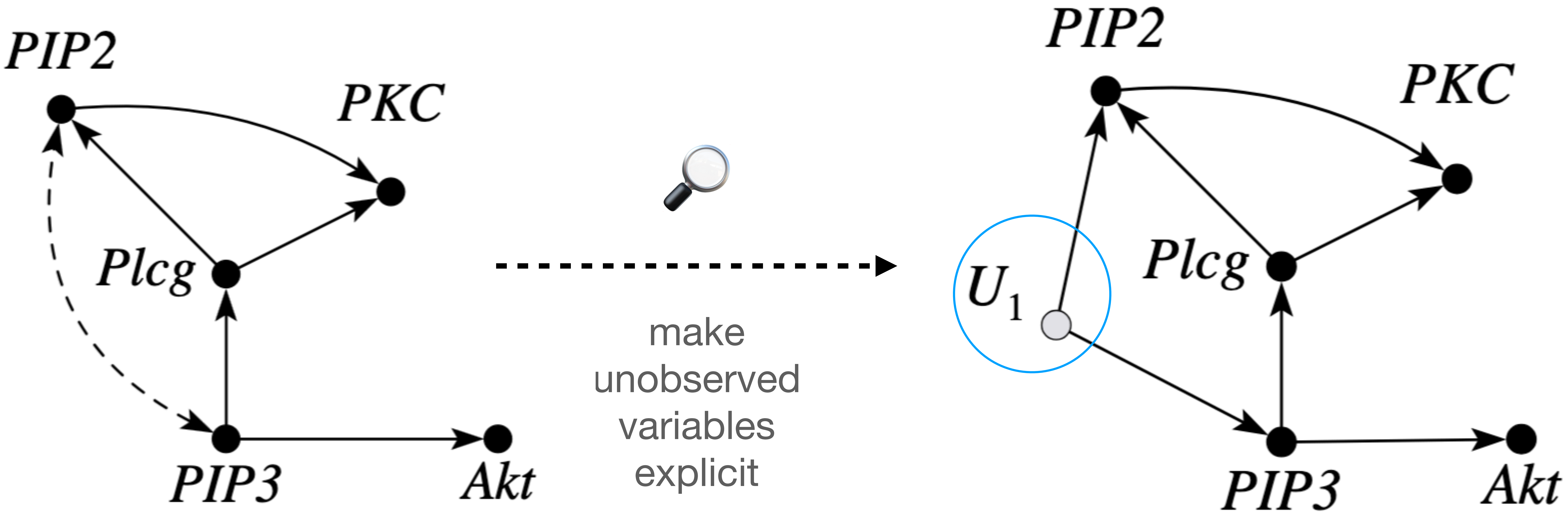
Given a causal graph $\mathcal{G}$ over variables $\mathbf{V}$, a probability distribution $P(\mathbf{v})$ satisfies *the local Markov property* (LMP) for $\mathcal{G}$ if, for any $X \in \mathbf{V}$,

$$X \perp \textit{non-desc}(X) \backslash pa(X)) \mid pa(X) \text{ in } P(\mathbf{v})$$

Lauritzen, S. L.; Dawid, A. P.; Larsen, B. N.; and Leimer, H. G. G. 1990. Independence Properties of Directed Markov Fields. Networks, 20(5): 491–505.
Pearl, J. 1988. Probabilistic Reasoning in Intelligent Systems. San Mateo, CA: Morgan Kaufmann

The Hidden Variable Problem

- $PIP2 \perp PIP3, Akt \mid Plcg$? No
- $PIP2 \perp PIP3, Akt \mid Plcg, U_1$? Yes



Local Markov Property

Definition 1 (LMP (Pearl 88, Lauritzen et al. 90))

Given a causal graph $\mathcal{G}$ over variables $\mathbf{V}$, a probability distribution $P(\mathbf{v})$ satisfies *the local Markov property* (LMP) for $\mathcal{G}$ if, for any $X \in \mathbf{V}$,

$$X \perp \textit{non-desc}(X) \setminus pa(X)) \mid pa(X) \text{ in } P(\mathbf{v})$$

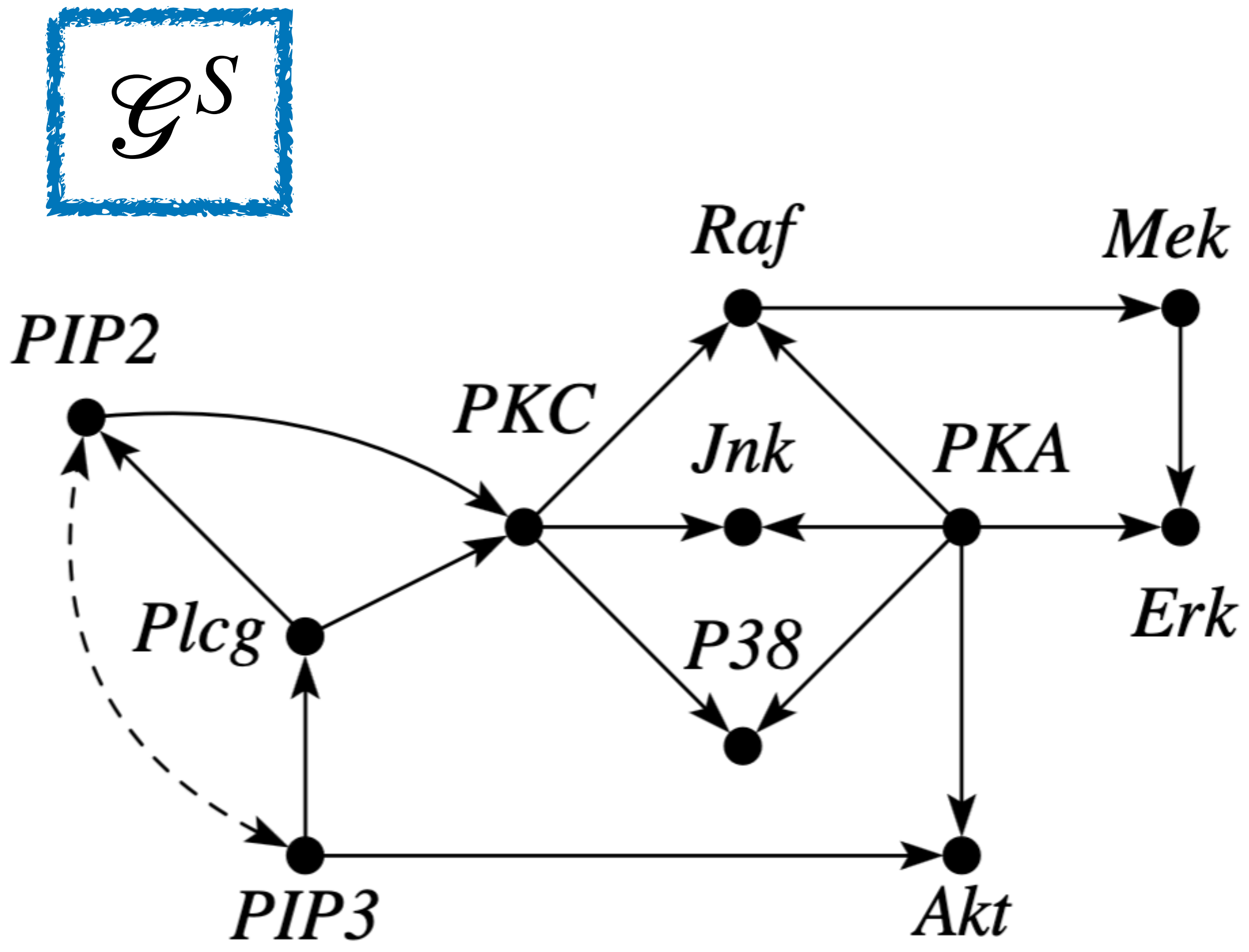
Assumes causal sufficiency!

Lauritzen, S. L.; Dawid, A. P.; Larsen, B. N.; and Leimer, H. G. G. 1990. Independence Properties of Directed Markov Fields. Networks, 20(5): 491–505.
Pearl, J. 1988. Probabilistic Reasoning in Intelligent Systems. San Mateo, CA: Morgan Kaufmann

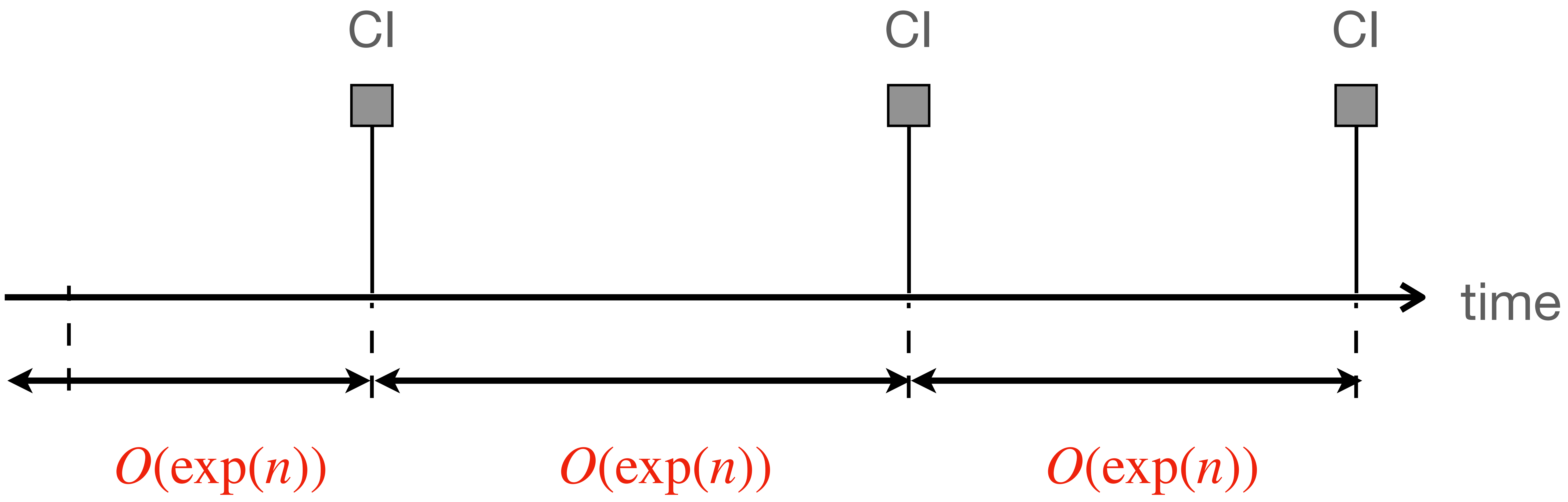
Motivation

When hidden variables and non-parametric data distributions are involved:

- Local Markov Properties exist, but take exponential time to enumerate!

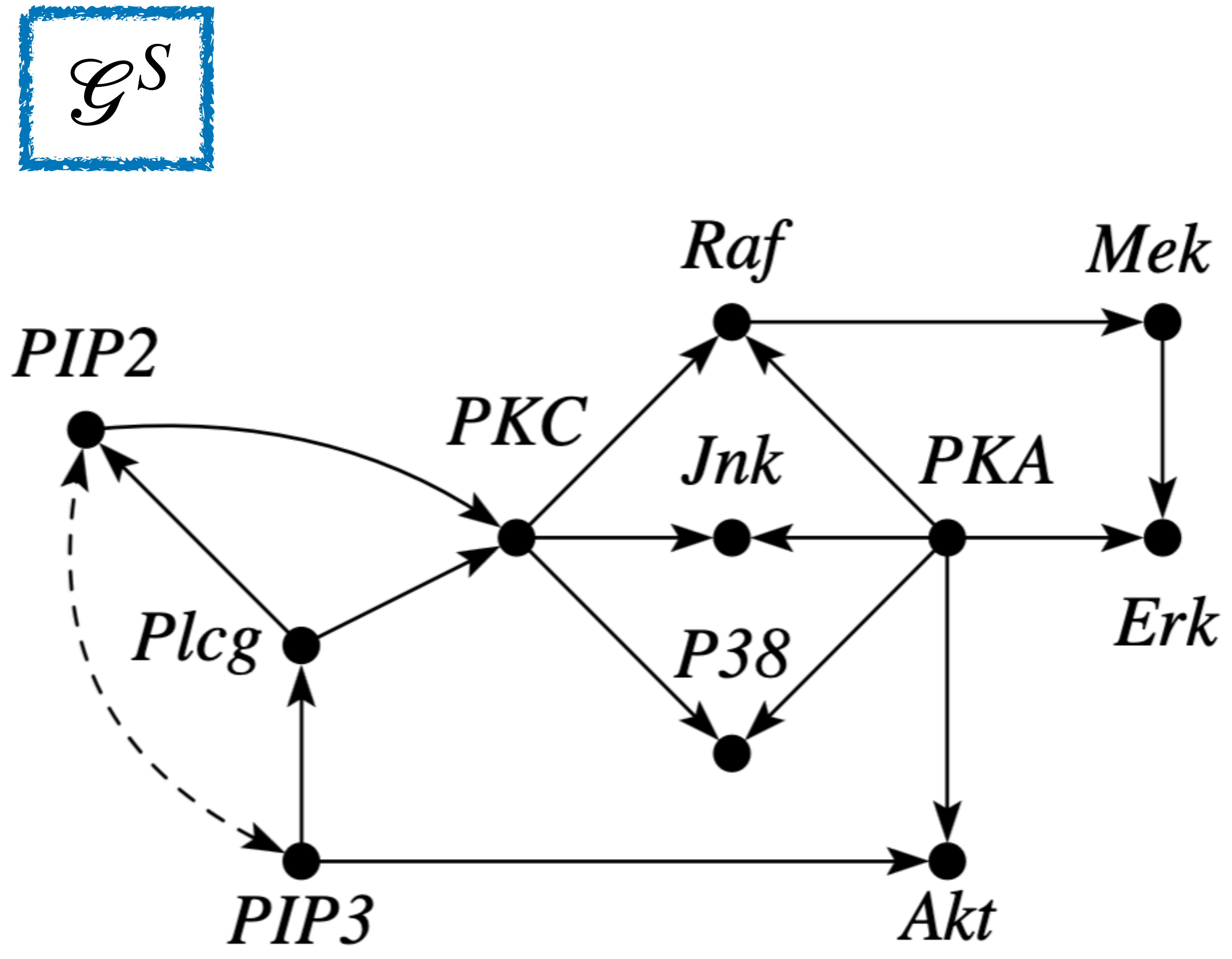


Need to test only 10 CIs, but don't know how to enumerate them!



Result I - C-LMP

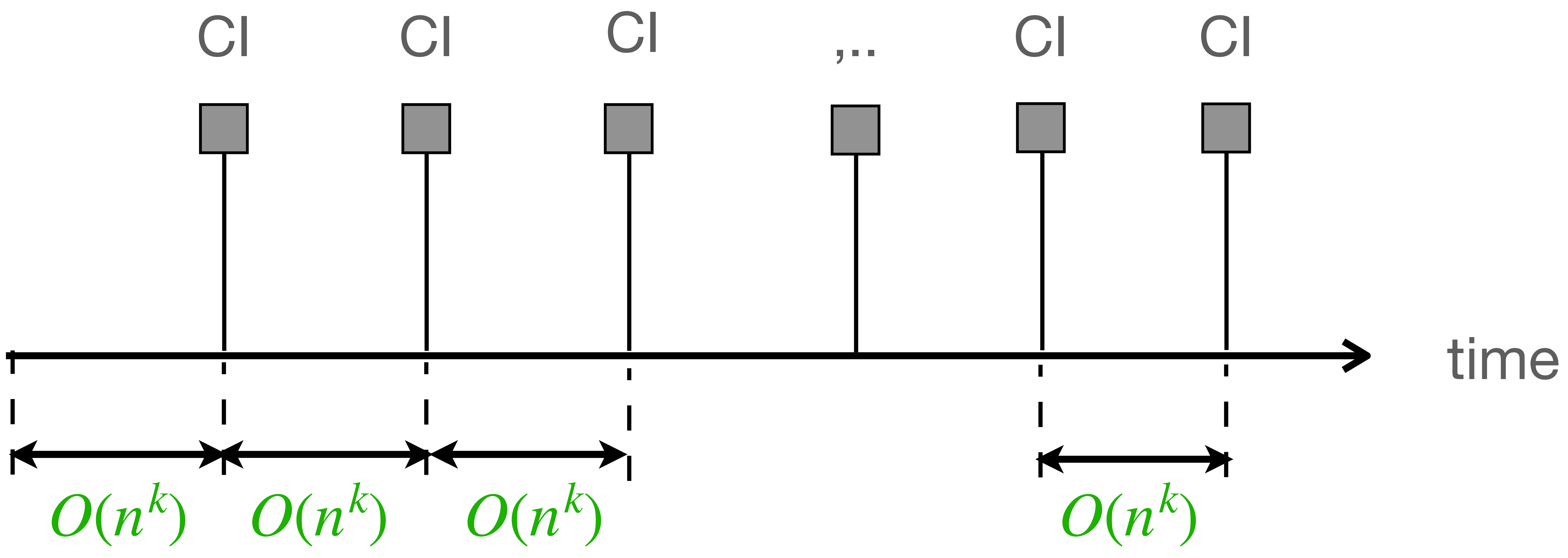
(i) Introduce the C-component local Markov property (C-LMP) for causal graphs with hidden variables.
- Exponentially fewer CI tests than the global Markov property.

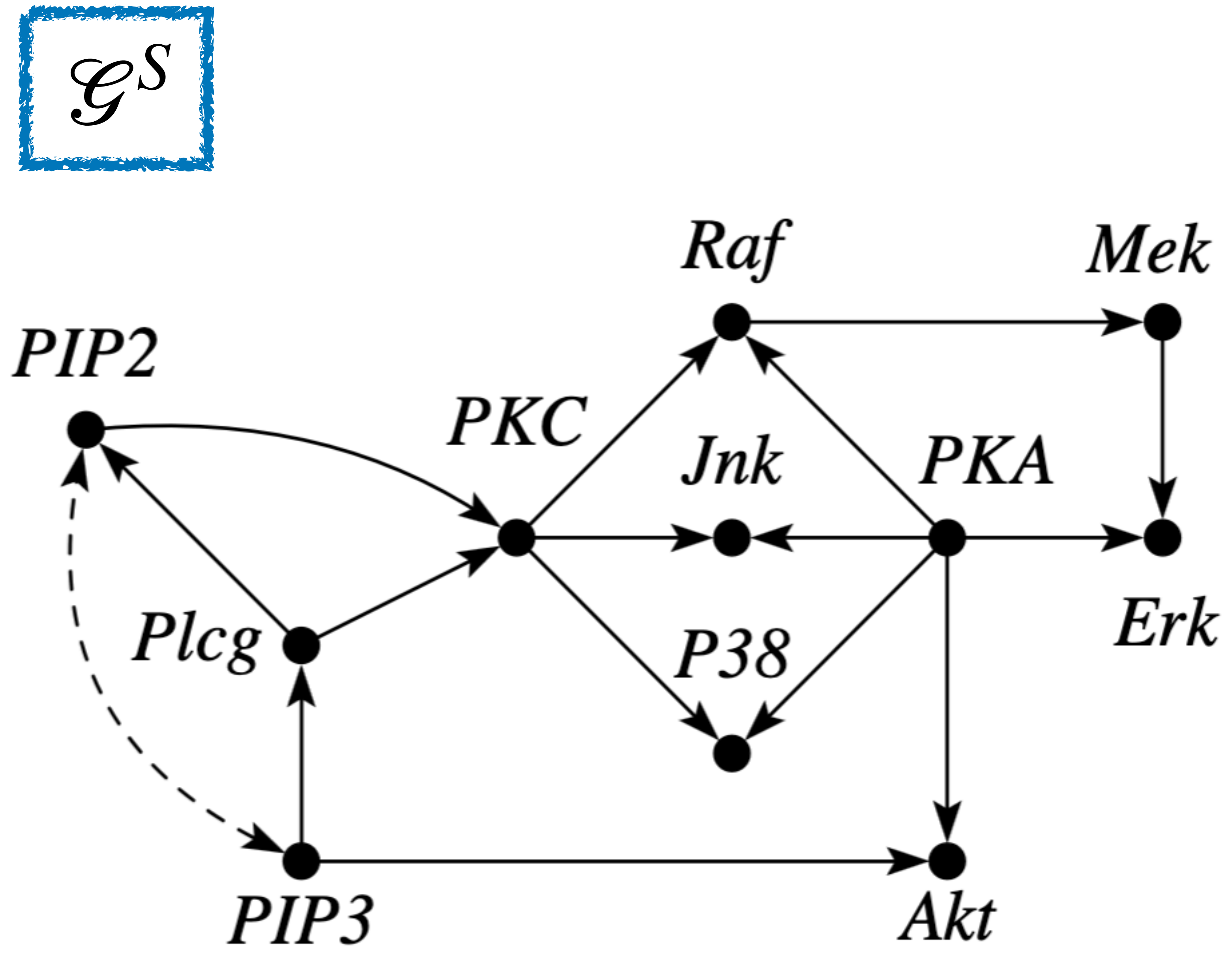


C-LMP invokes 10 CIs which imply all other 76570 CIs encoded in $\mathcal{G}$.

Result II - `ListCI`

(i) Present `ListCI`, an algorithm for listing CIs invoked by C-LMP in polynomial delay





`ListCI` lists 10 CIs in poly-time intervals which imply all other 76570 CIs encoded in $\mathcal{G}$

C-component Local Markov Property

Definition 2 (C-LMP)

Given a casual graph $\mathcal{G}$ and a consistent ordering $\mathbf{V}^{\prec}$, a probability distribution $P(\mathbf{v})$ satisfies *the c-component local Markov property* (C-LMP) for $\mathcal{G}$ w.r.t $\mathbf{V}^{\prec}$ if, for any $X \in \mathbf{V}^{\prec}$ and ancestral c-component $\mathbf{C}$ relative to X,

$$X \perp \mathbf{W} \mid \mathbf{Z} \text{ in } P(\mathbf{v}) \text{ where}$$

$$\mathbf{W} = \mathbf{V}^{\leq X} \backslash (De(Sp(\mathbf{C}) \backslash Pa(\mathbf{C})) \cup Pa(\mathbf{C}))$$

$$\mathbf{Z} = Pa(\mathbf{C}) \backslash \{X\}.$$

Compare: $X \perp \textit{non-desc}(X) \backslash pa(X)) \mid pa(X)$ in

Example

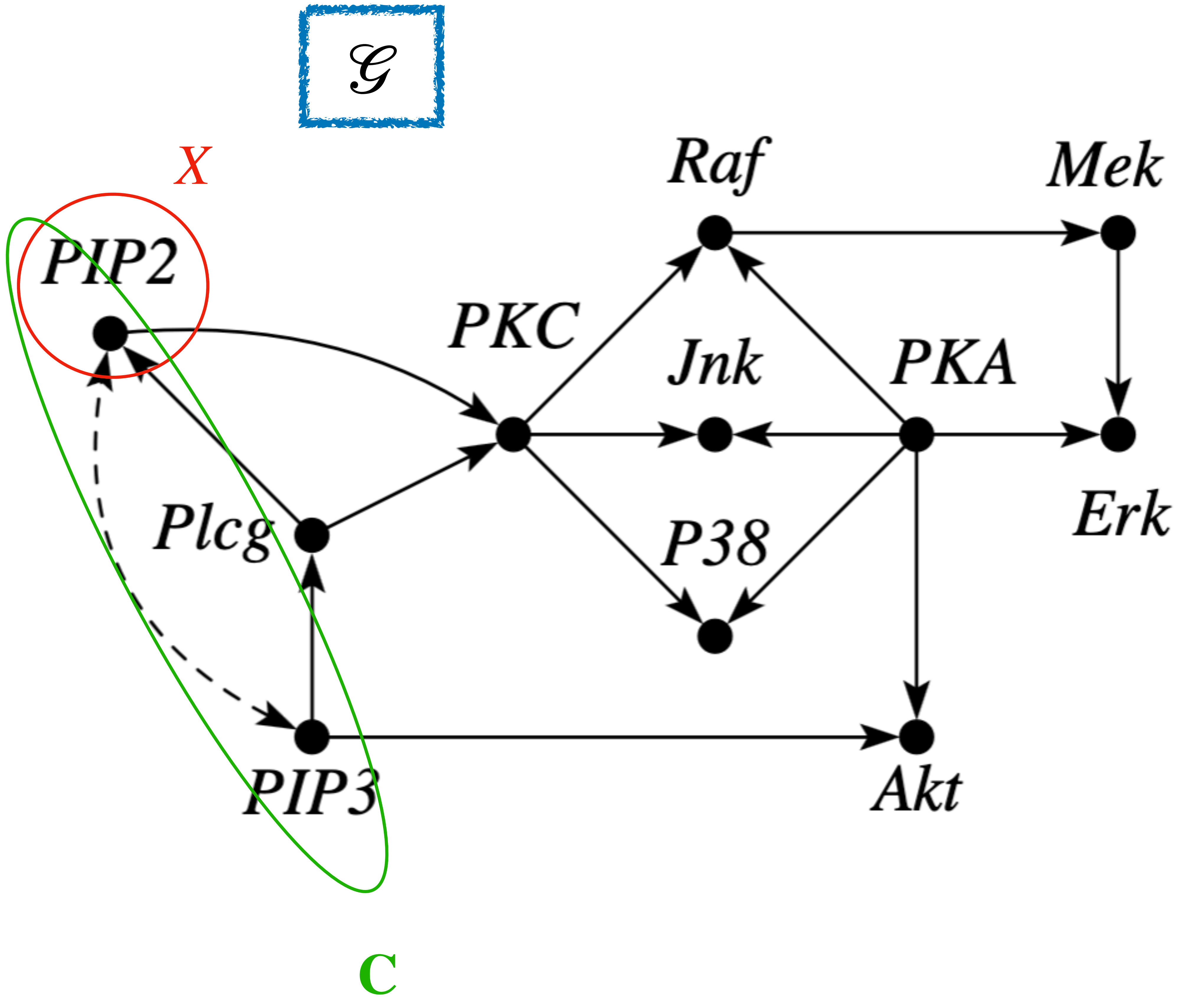
1. Let $X = PIP2$ and an ancestral c-component $\mathbf{C} = \{PIP2,\ PIP3\}$.

2. Then, we have the CI: $X \perp \mathbf{W} \mid \mathbf{Z}$

- $\mathbf{W} = \mathbf{V}^{\leq X} \setminus (De(Sp(\mathbf{C}) \setminus Pa(\mathbf{C})) \cup Pa(\mathbf{C})) = Akt, PKA$

- $\mathbf{Z} = Pa(\mathbf{C}) \setminus \{X\} = PIP3, Plcg$

$\therefore PIP2 \perp Akt, PKA \mid PIP3, Plcg.$



$PIP2 \perp PIP3,\ Akt \mid Plcg$? No

$PIP2 \perp Akt, PKA \mid PIP3, Plcg$? Yes

Properties of C-LMP (1/2)

Theorem 1 (Equivalence of C-LMP and GMP)

Given $\mathscr{G}$ and a consistent ordering $\mathbf{V}^{\prec}$, a probability distribution $P(\mathbf{v})$ satisfies *the c-component local Markov property* for $\mathscr{G}$ w.r.t $\mathbf{V}^{\prec}$ if and only if it satisfies the global Markov property for $\mathscr{G}$.

Test CIs invoked by C-LMP instead of GMP!

Properties of C-LMP (2/2)

Theorem 2 (1-1 correspondence between ACs and CIs invoked by C-LMP, informal)

Given $\mathscr{G}$, $\mathbf{V}^{\prec}$, and $X \in \mathbf{V}^{\prec}$, every CI invoked by *the c-component local Markov property* for X can be generated by exactly one ancestral c-component $\mathbf{C}$.

To generate all CIs invoked by C-LMP, generate all $\mathbf{C}$ for each X!

Result II - Listing CIs Invoked by C-LMP

C-LMP may invoke exponentially many CIs in the worst case.

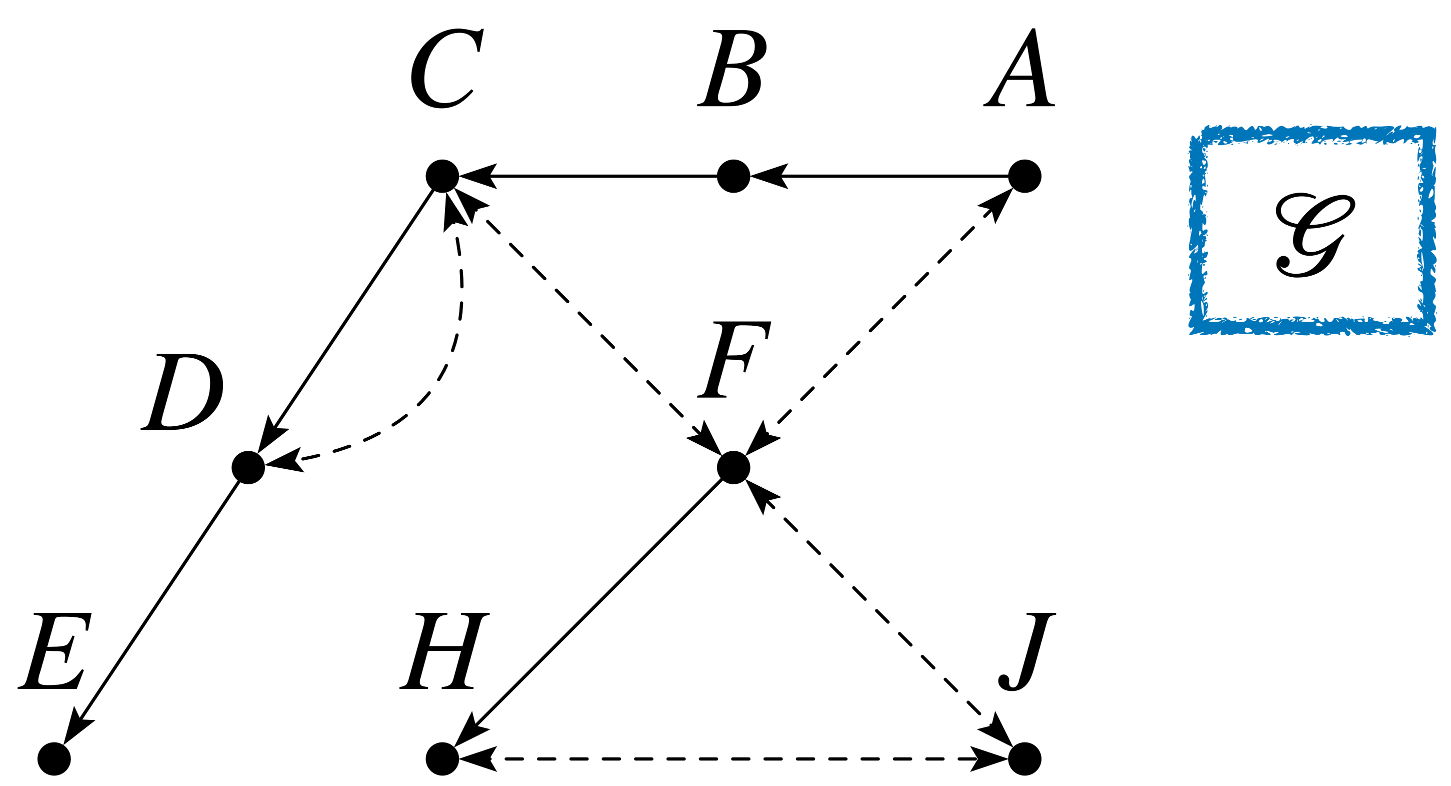
- $O(n2^s)$ with $s \leq n$ (size of the largest c-component).

We present **`ListCI`**, an algorithm that enumerates all and only all non-vacuous CIs in poly-delay.

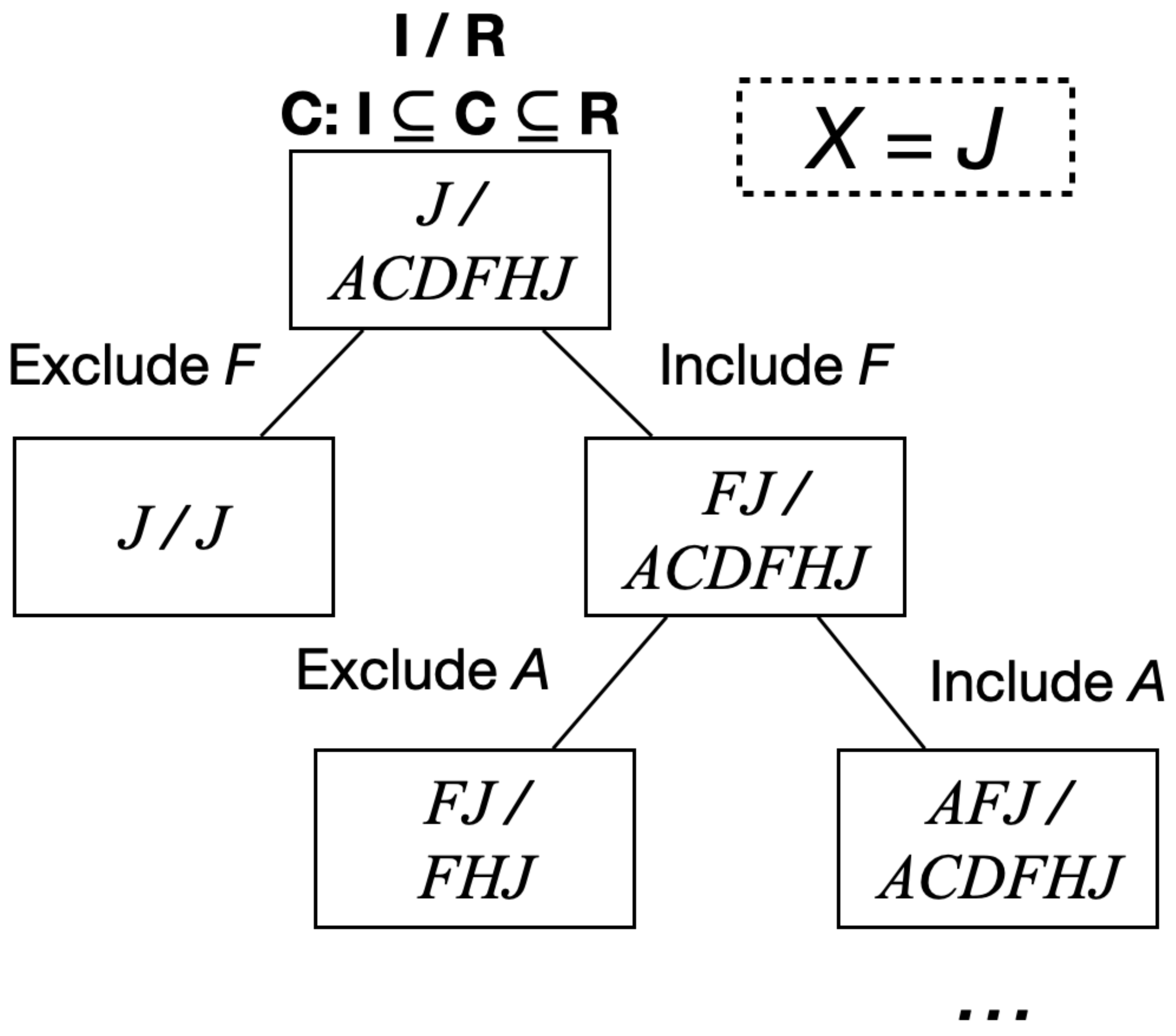
Listing Algorithm - `ListCI`

`ListCI`: For every variable X,

1. **Root**: search space of all possible ACs containing X.
2. Choose some variable Y (in a clever way).
    1. **Left child:** ACs containing X but not Y.
    2. **Right child:** ACs containing X and Y.

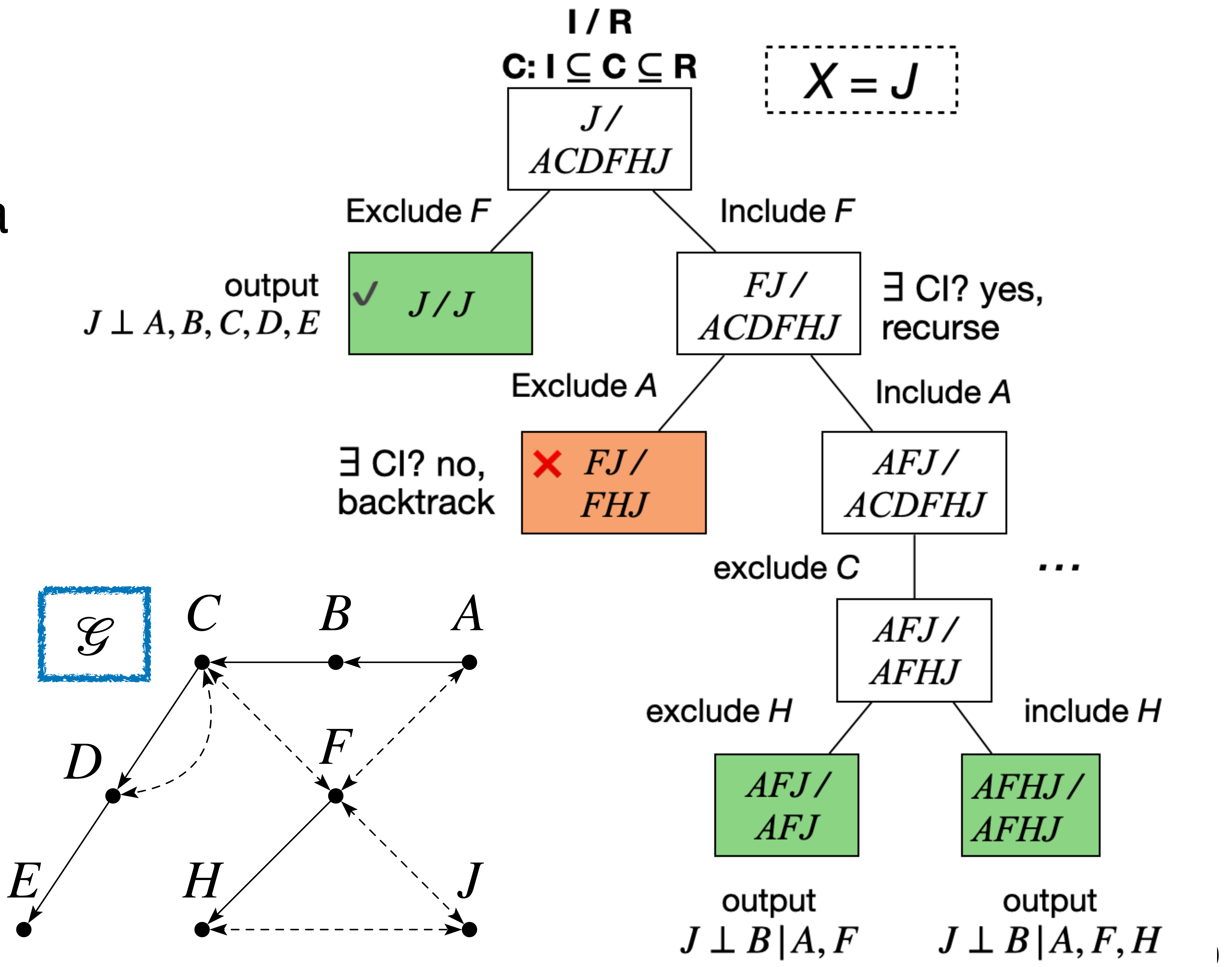


Problem! Exponentially many vacuous CIs, i.e., $X \perp \emptyset \mid \mathbf{Z}$



Listing Algorithm - `ListCI`

3. On each child, check (in polynomial time!) if the search space contains a non-vacuous CI.
    1. If not, backtrack.
    2. Otherwise, recurse.
4. Leaf node: search space contains exactly one AC. Output the corresponding CI.



Listing Algorithm - `ListCI`

Theorem 3 (Correctness of `ListCI`)

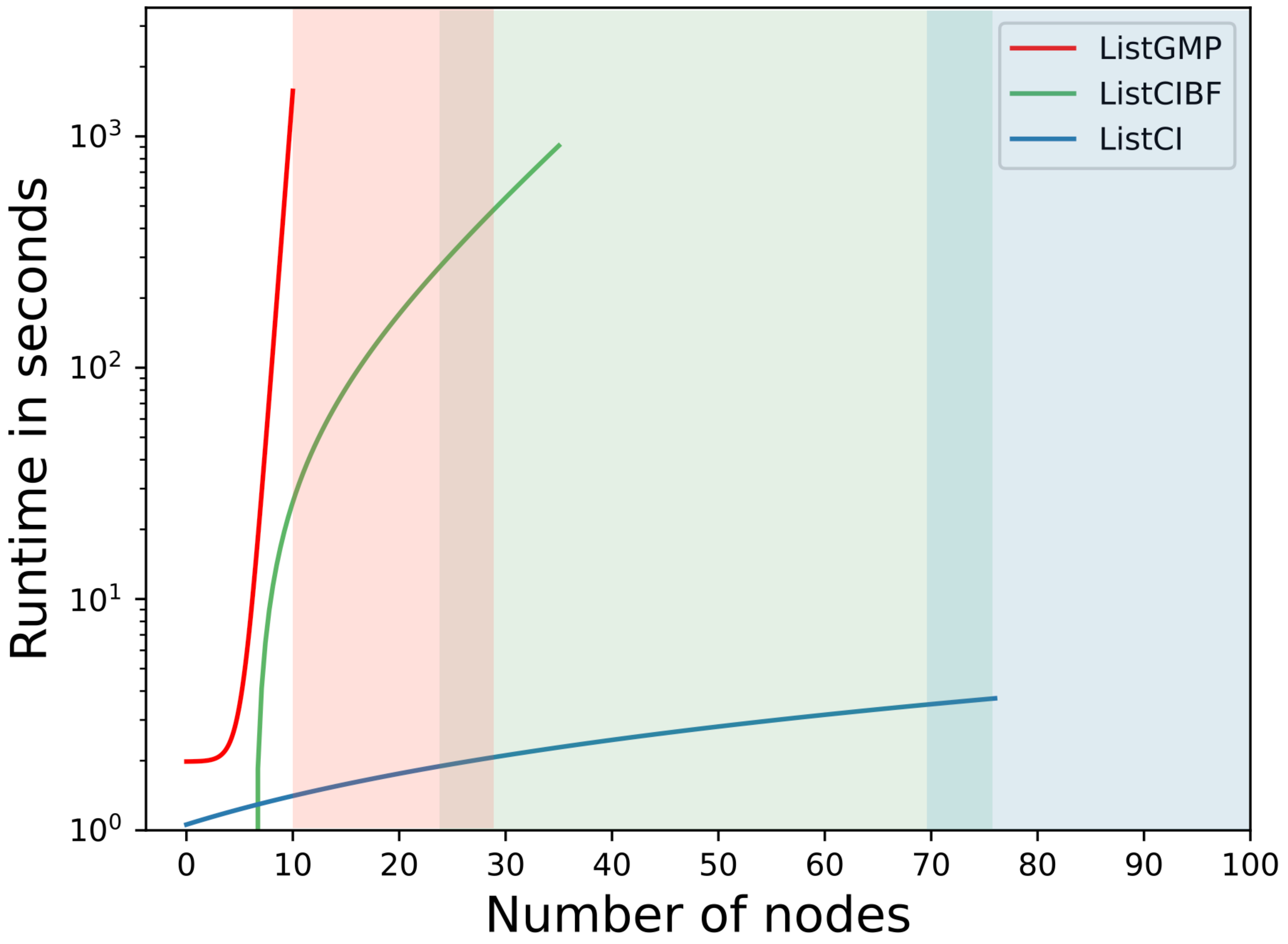
Given $\mathcal{G}$ and a consistent ordering $\mathbf{V}^{\prec}$, $\texttt{ListCI}(\mathcal{G}, \mathbf{V}^{\prec})$ enumerates all and only all non-vacuous CIs invoked by the c-component local Markov property for $\mathcal{G}$ w.r.t $\mathbf{V}^{\prec}$.

1. Complexity
   - Time: $O(n^2(n+m))$ delay. Space: $O(n(n+m))$.

Empirical Performance

Comparison of `ListCI` with two baseline algorithms.

- Algorithms: `ListGMP`, `ListCIBF` (brute force C-LMP).
- Graphs: real-world instances on up to 100 nodes from `bnlearn` package (Scutari 2010).
- Conclusion: `ListCI` is **orders of magnitude faster** than baselines.



y-axis: log scale.

Colored region: algorithm timed out (i.e., spent > 1 hour).

Scutari, M. 2010. Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software, 35(3): 1–22

Testing the Sachs graph

1. Dataset: Protein signaling dataset
   - Continuous, 853 observations.

2. Null hypothesis of dependence.

3. Result: 7 out of 10 CIs invoked by C-LMP resulted in $p \geq 0.05$.

4. Conclusion: $\mathscr{G}$ may need revision.
   - CIs that are violated may guide experts in revision process.

CIs implied by C-LMP for GS	p-value
PIP3 $\perp$ {PKA}	0.175
Plcg $\perp$ {PKA} | {PIP3}	0.081
Akt $\perp$ {Plcg} | {PIP3, PKA}	0.370
PIP2 $\perp$ {Akt, PKA} | {PIP3, Plcg}	0.648
PKC $\perp$ {Akt, PIP3, PKA} | {PIP2, Plcg}	0.318
Raf $\perp$ {Akt, PIP2, PIP3, Plcg} | {PKA, PKC}	0.036
P38 $\perp$ {Akt, PIP2, PIP3, Plcg, Raf} | {PKA, PKC}	0.680
Jnk $\perp$ {Akt, P38, PIP2, PIP3, Plcg, Raf} | {PKA, PKC}	0.002
Mek $\perp$ {Akt, Jnk, P38, PIP2, PIP3, PKA, PKC, Plcg} | {Raf}	0.544
Erk $\perp$ {Akt, Jnk, P38, PIP2, PIP3, PKC, Plcg, Raf} | {Mek, PKA}	0.000

Conclusions

1. Researchers can evaluate a causal graph prior to inference by enumerating CIs encoded in the graph if they are consistent with the observational data.

2. Testing all CIs encoded in the graph is both impractical and unnecessary. We introduce C-LMP, which implies all CIs invoked by GMP.

3. We introduced C-LMP, and we developed `ListCI` that enumerates all CIs invoked by C-LMP in polynomial delay.