Inferring causality from a mixture of observations and interventions

IA et Santé: Approches Interdisciplinaires

Nantes, 29 juin - 1er juillet, 2022

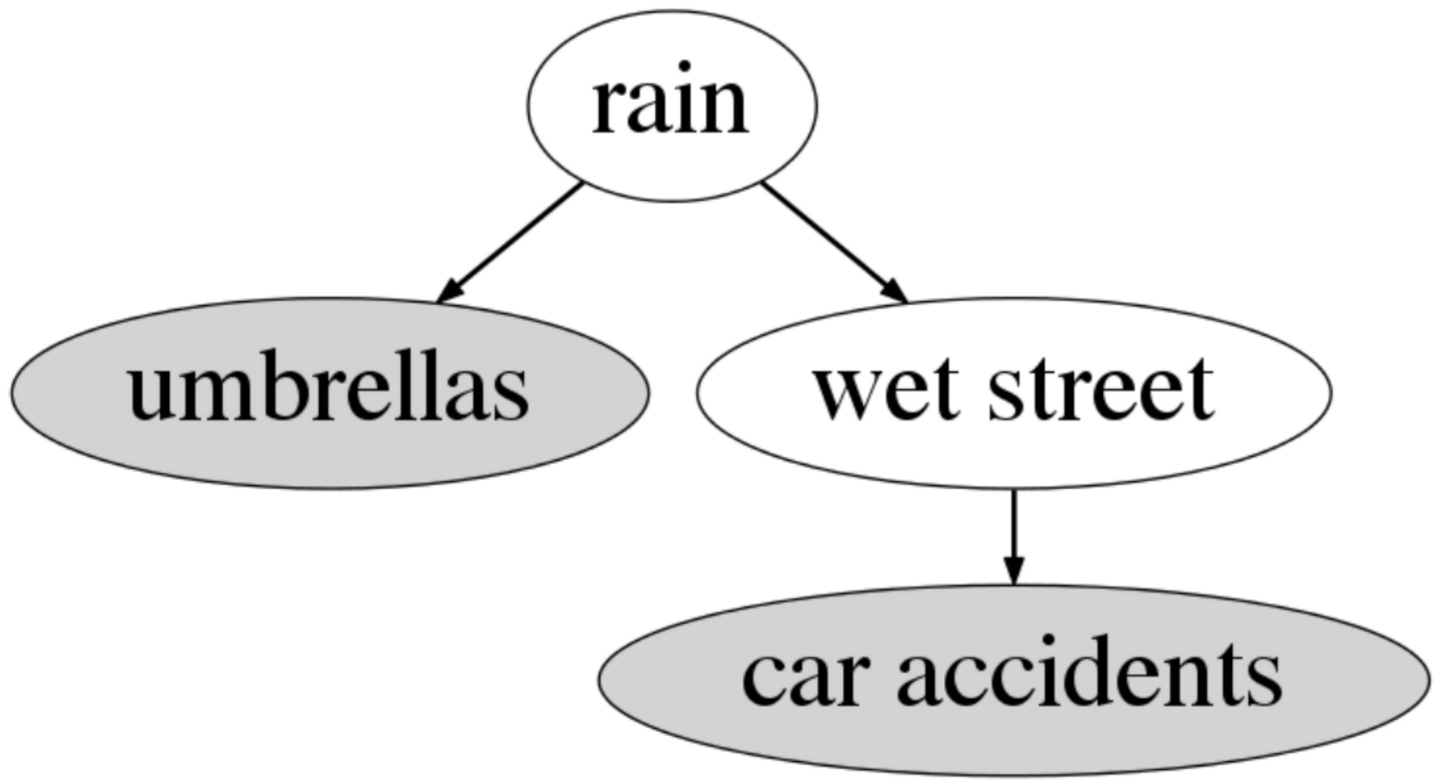
G. Nuel

LPSM
LABORATOIRE DE PROBABILITÉS
STATISTIQUE & MODÉLISATION

SORBONNE
UNIVERSITÉ

cnrs

Correlation is not Causation

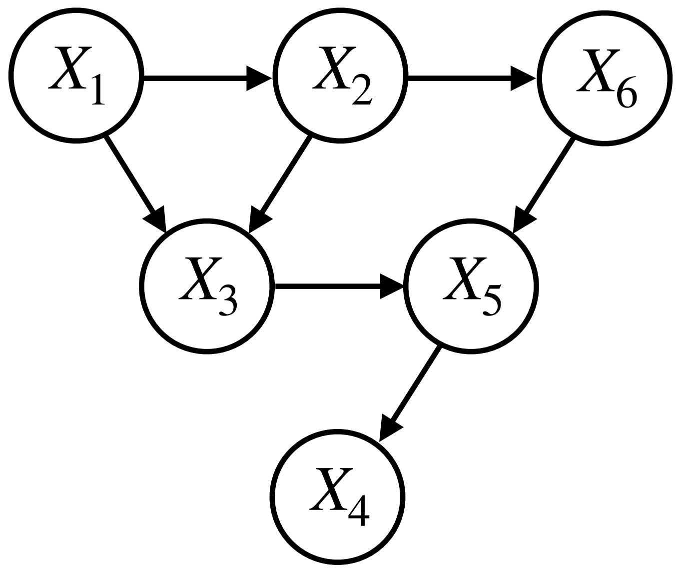


umbrellas and car accidents are correlated

But:

- provoking car accidents does not make appear umbrellas
- distributing umbrellas in the street does not provoke car accidents

Directed Acyclic Graphs

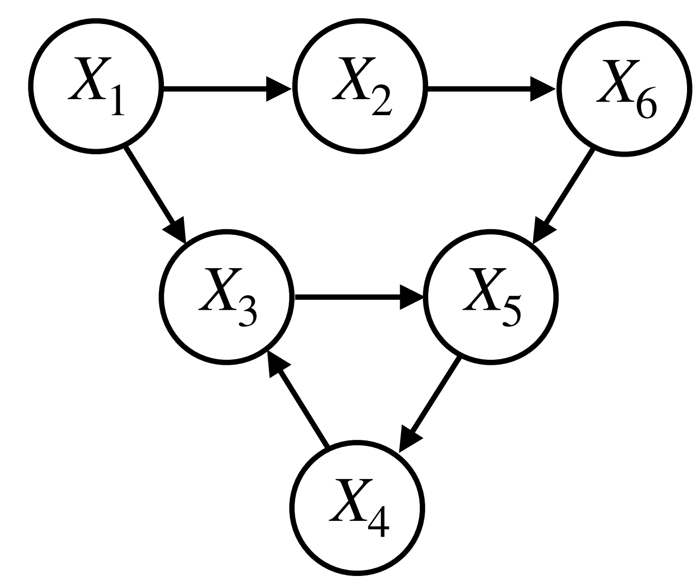


Definition (DAG):
$G = (V = \{1,\ldots,n\}, E \subset V \times V)$ is a *Directed Acyclic Graph* if and only if it has no cycle

Remark: loops are ok (*e.g.* $X_1 - X_2 - X_3$)

Theorem (topological ordering):
$G = (V = \{1,\ldots,n\}, E \subset V \times V)$ is a Directed Acyclic Graph if and only if it exists a t*opological ordering* $\sigma_1, \ldots, \sigma_n$ such that $\forall i, j \in V$ such that $(i, j) \in E$ we have $\sigma_i < \sigma_j$

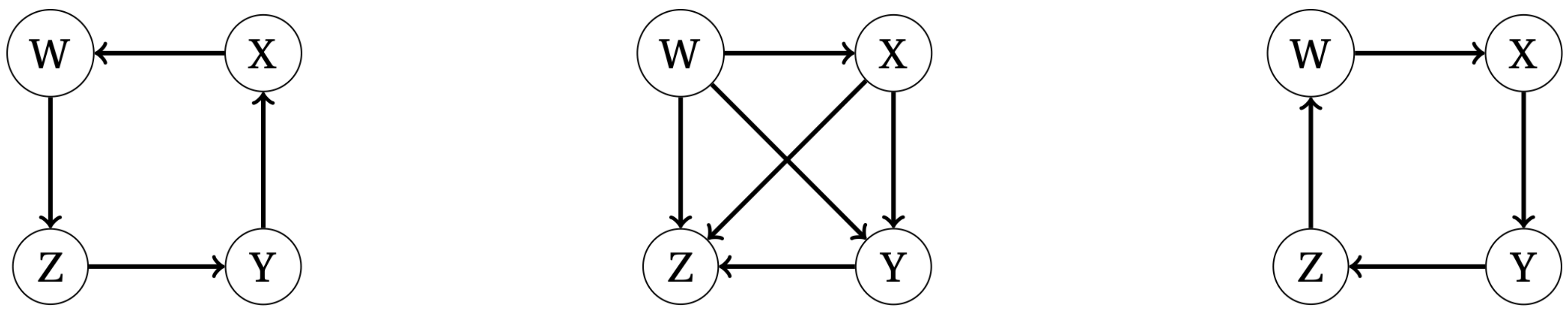
Example: $\sigma = (1,2,3,6,5,4)$ is a topological ordering



not a DAG

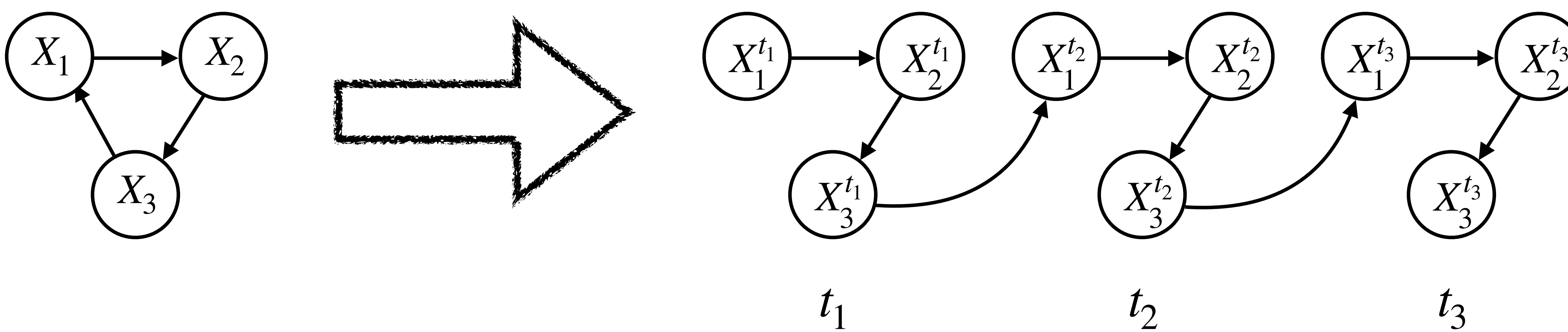
Directed Acyclic Graphs

Why not cyclic directed graph ?

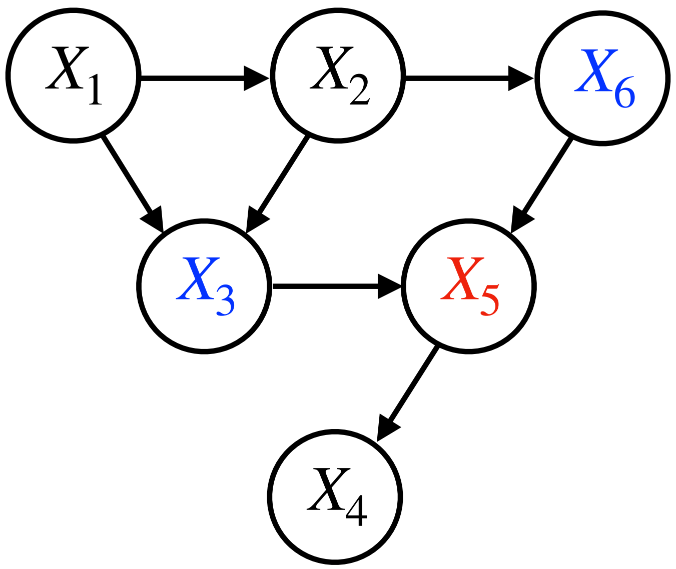


Three models with the same distribution (Monneret 2019)

What about feedback loops ?



Bayesian Networks



Definition (Bayesian Network):
$(G, \mathbb{P})$ is a Bayesian Network if and only if $G = (V, E)$ is a DAG and

$$\mathbb{P}(X) = \prod_{j=1}^{n} \mathbb{P}(X_j \mid X_{\mathrm{pa}_j})$$

with $\mathrm{pa}_j = \{i \in V, (i, j) \in E\}$

Example:

$$\mathbb{P}(X_1, \ldots, X_6) = \mathbb{P}(X_1)\mathbb{P}(X_2 \mid X_1)\mathbb{P}(X_3 \mid X_1, X_2)\mathbb{P}(X_6 \mid X_2)\mathbb{P}(X_5 \mid X_3, X_6)\mathbb{P}(X_4 \mid X_5)$$

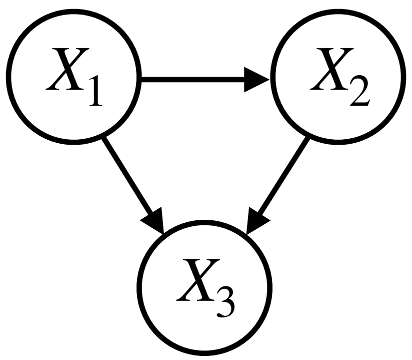
NB: the topological ordering provide a generative procedure

Example of conditional distribution:

With *Gaussian Bayesian Networks* $\mathbb{P}(X_i \mid X_{\mathrm{pa}_j} = Z) \sim \mathcal{N}(Z\beta, \sigma^2)$ but we can use any GLM: binomial $\mathscr{B}(n, \mathrm{softmax}(Z\beta))$, Poisson $\mathscr{P}(e^{Z\beta})$, etc.

Markov Equivalence Class

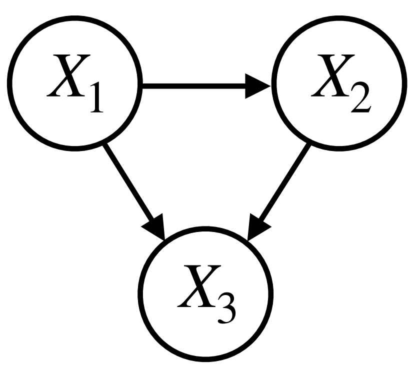
Simulation



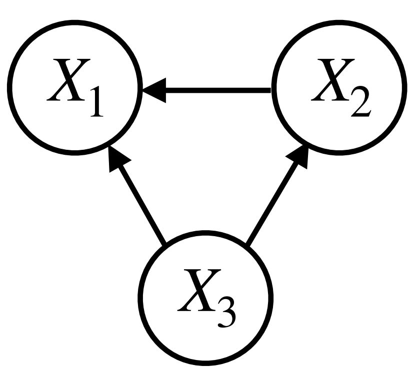

```
set.seed(42)
x1=rnorm(1000)
x2=-0.3*x1+rnorm(1000)
x3=1.2*x1+0.5*x2+rnorm(1000)
```

	x1	x2	x3
[1,]	1.3709584	1.9137710	2.85261368
[2,]	-0.5646982	0.6935316	-0.60879604
[3,]	0.3631284	0.8617949	-0.85808419
[4,]	0.6328626	0.1871146	-1.15371251
[5,]	0.4042683	-1.1172139	-1.36529329
[6,]	-0.1061245	-0.5656456	-0.04433397

Estimations



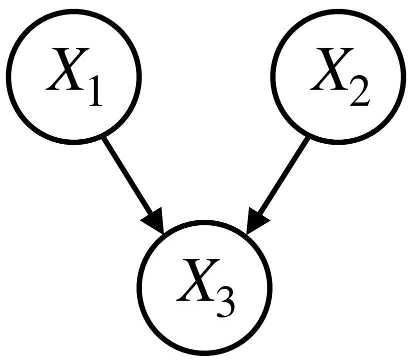

```
> reg=list(lm(x1~1),lm(x2~x1),lm(x3~x1+x2))
> sum(sapply(reg,logLik))
[1] -4272.506
```




```
> reg=list(lm(x1~x2+x3),lm(x2~x3),lm(x3~1))
> sum(sapply(reg,logLik))
[1] -4272.506
```

Markov Equivalence Class

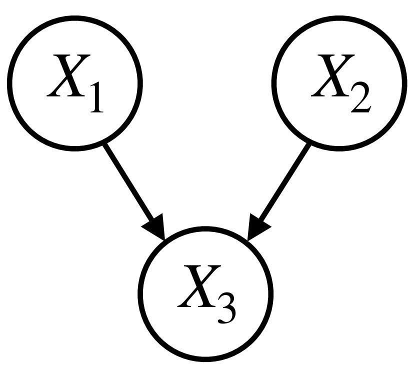
Simulation



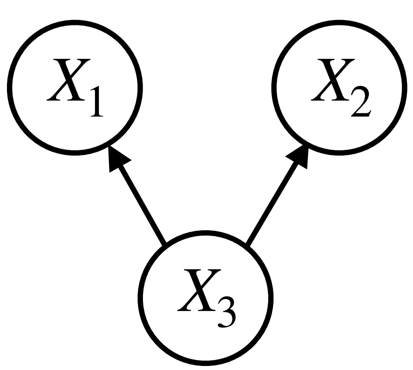

```r
set.seed(42)
x1=rnorm(1000)
x2=rnorm(1000)
x3=1.2*x1+0.5*x2+rnorm(1000)
```

```
             x1         x2          x3
[1,]  1.3709584  1.9137710  2.85261368
[2,] -0.5646982  0.6935316 -0.60879604
[3,]  0.3631284  0.8617949 -0.85808419
[4,]  0.6328626  0.1871146 -1.15371251
[5,]  0.4042683 -1.1172139 -1.36529329
[6,] -0.1061245 -0.5656456 -0.04433397
```

Estimations




```r
> reg=list(lm(x1~1),lm(x2~1),lm(x3~x1+x2))
> sum(sapply(reg,logLik))
[1] -4314.239
```



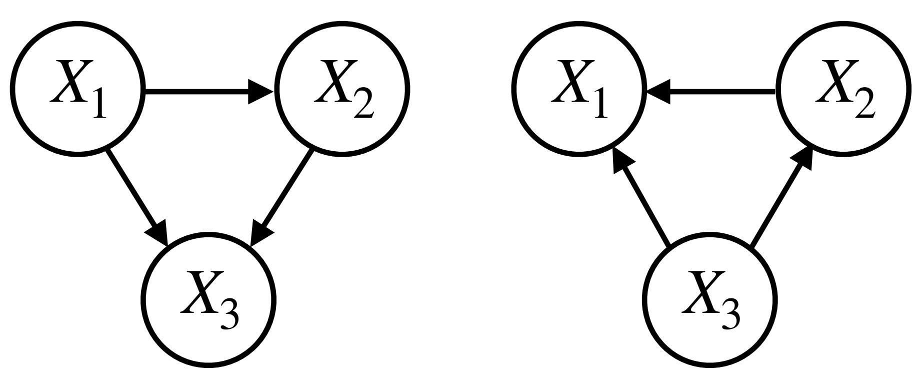

```r
> reg=list(lm(x1~x3),lm(x2~x3),lm(x3~1))
> sum(sapply(reg,logLik))
[1] -4405.149
```

Markov Equivalence Class

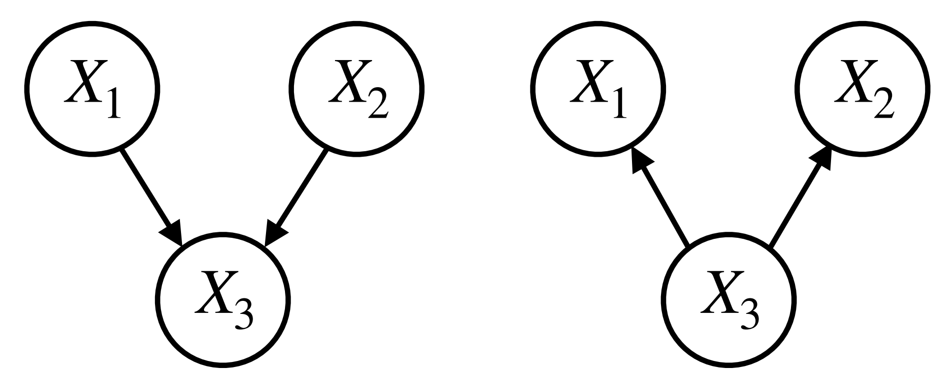
Definition (skeleton) :
The *skeleton* of a DAG is the *undirected* graph induced by its (directed) edges

Definition (v-structure):
(A, B, C) is a *v-structure* of a DAG iff:
$A \to B \leftarrow C$ without $A \to C$ nor $A \leftarrow C$

Theorem (2.1 in Andersson *et al* 1997)**:**
Two DAGs are Markov equivalent if and only if they have the same *skeleton* and the same *v-structures* (also called *immoralities*).



Same skeleton no v-structures

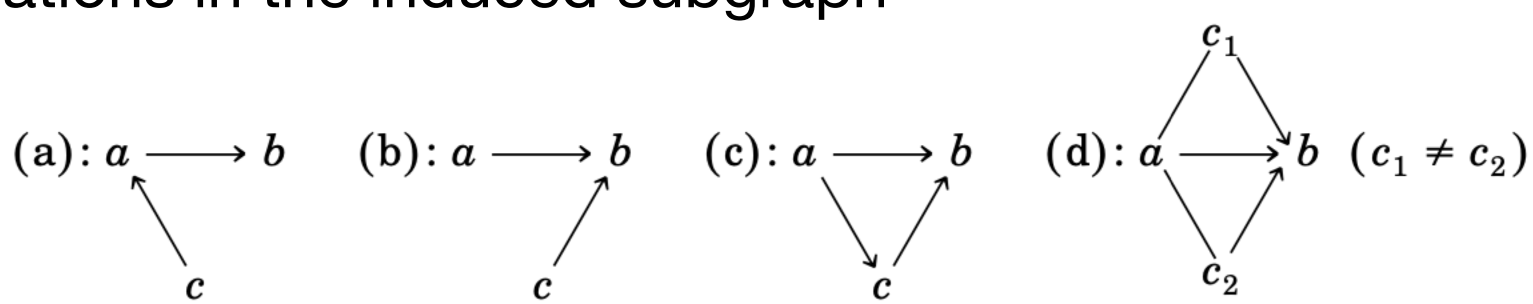


Same skeleton different v-structures

CPDAG: Completed Partially Directed Acyclic Graph

Definition (CPDAG):
The CPDAG (also called essential graph) is a PDAG representing the MEC of a DAG. Directed edge iff shared by all DAGs, undirected otherwise.

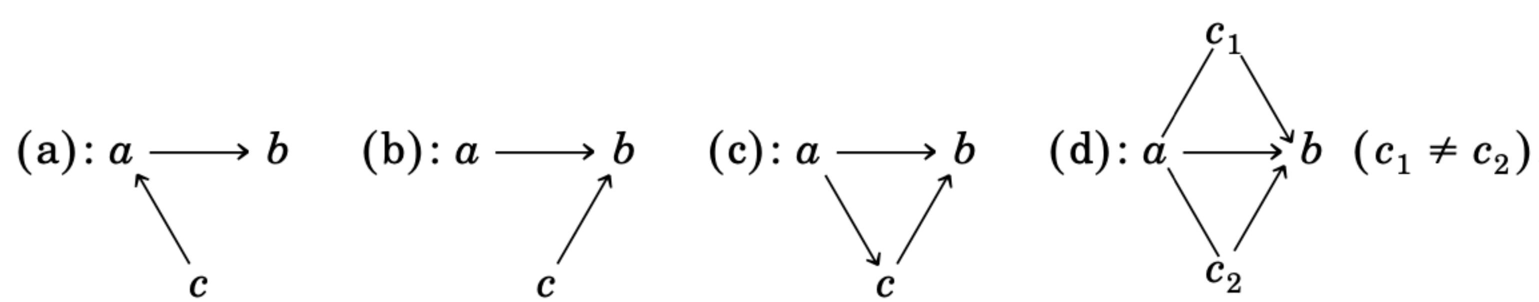
Definition (strongly protected arrows):
$a \to b$ is *strongly protected* in G if $a \to b$ occurs in at least one of the following configurations in the induced subgraph



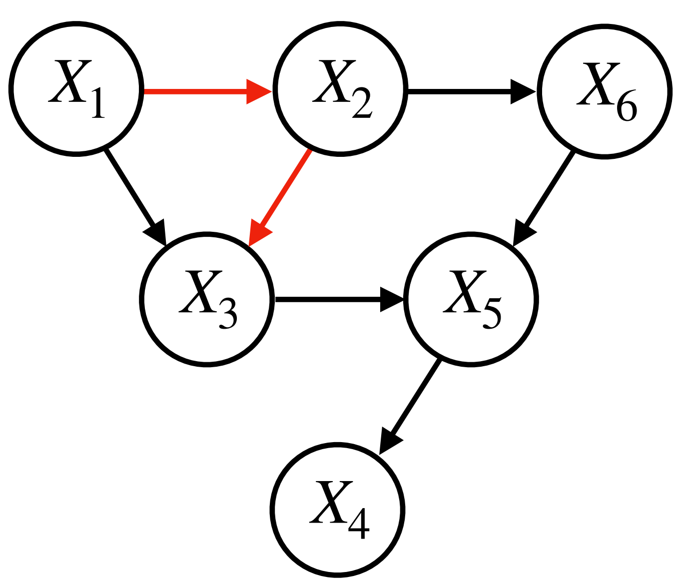
Definition 3.3 from Andersson *et al* (1997)

Algorithm (Algo 1, Hauser & Bühlmann, 2012): we can build a CPDAG from a DAG by dropping all arrows not strongly protected, updating the edges, and repeat until convergence

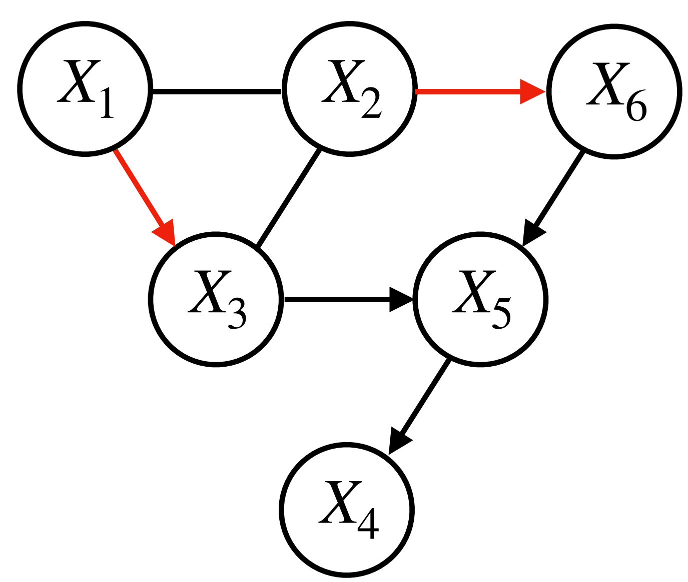
CPDAG: Completed Partially Directed Acyclic Graph



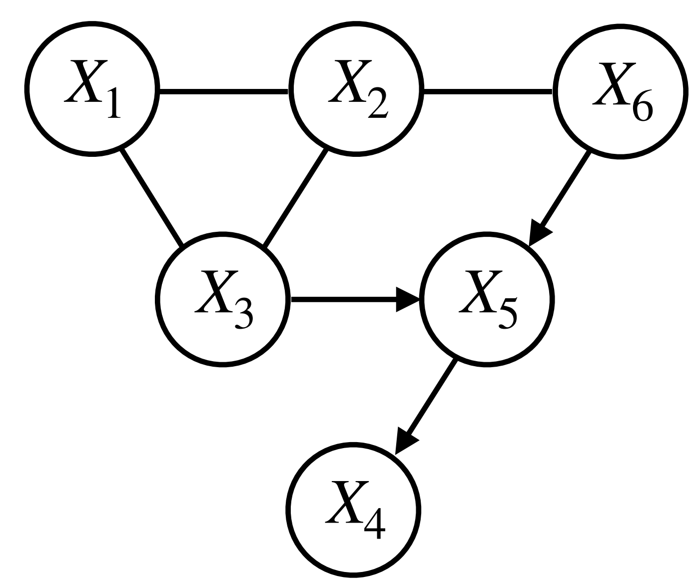
$a \rightarrow b$ strongly protected (Andersson *et al*,1997)



Initial DAG



Intermediary PDAG



CPDAG

Posterior DAG Distribution

$$\mathbb{P}(G|\text{data}) \propto \underbrace{\mathbb{P}(G)}_{\text{DAG prior}} \times \overbrace{\int_\theta \underbrace{\mathbb{P}(\text{data}|G,\theta)}_{\text{likelihood}} \times \underbrace{\mathbb{P}(\theta|G)}_{\text{param. prior}} \times d\theta}^{\text{integrated likelihood}}$$

Rather than integrating the likelihood, we use the following approximation:

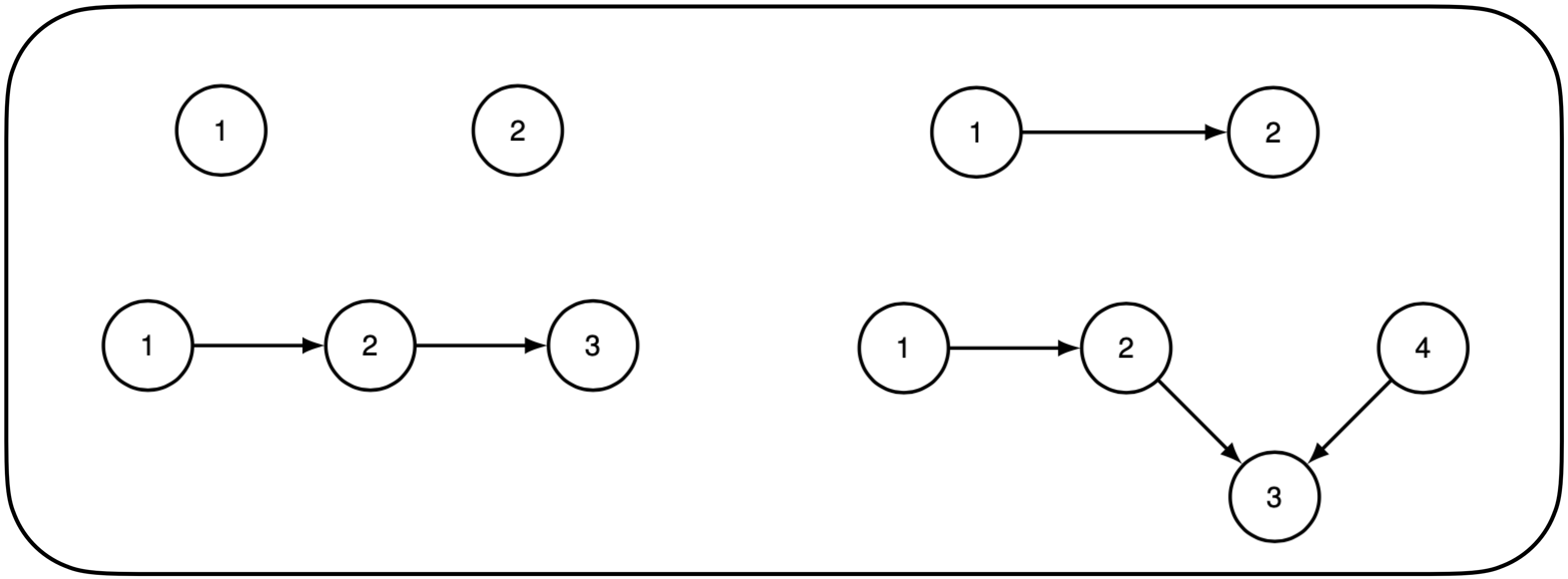
$$\log \mathbb{P}(G|\text{data}) \simeq \text{Cst.} + \log \mathbb{P}(G) + \operatorname{loglik}(\hat{\theta}|G) - \operatorname{pen}(G)$$

where the penalty function can be either:

- $\operatorname{pen}(G) = \frac{1}{2}\sum_j (|\mathrm{pa}_j| + 2)\log n_j$ (BIC)
- $\operatorname{pen}(G) = \frac{1}{2}\sum_j \left\{(|\mathrm{pa}_j| + 2)\log n_j + \log \binom{|\mathrm{pa}_j|}{p-1}\right\}$ (eBIC)

Toy-examples

Four reference DAGs

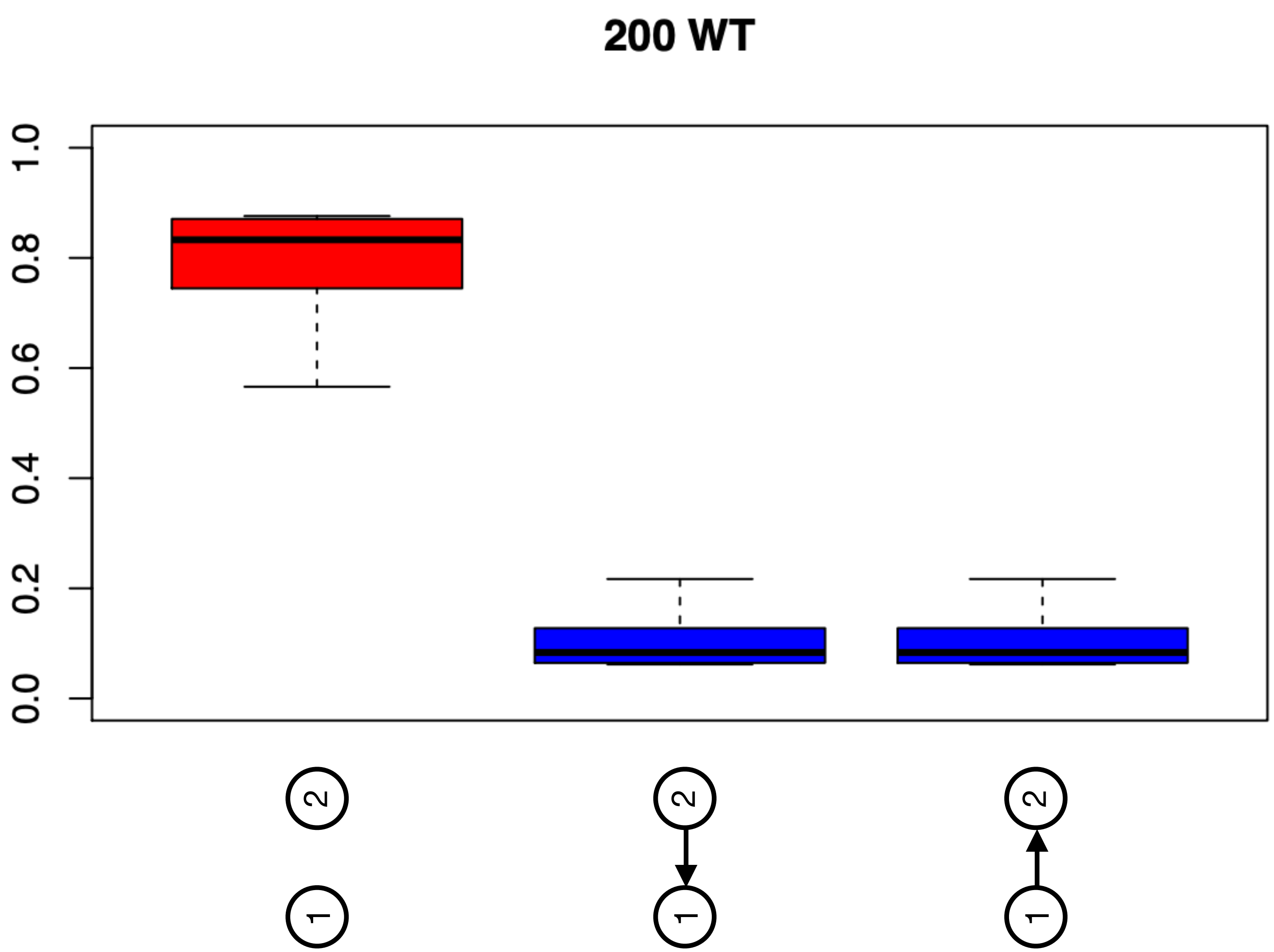


Experiments:

- Simulate 200 observations using a GBN
- Exhaustive search over the DAG space
- Posterior $\mathbb{P}(G \,|\, \text{data})$ over 100 replicates

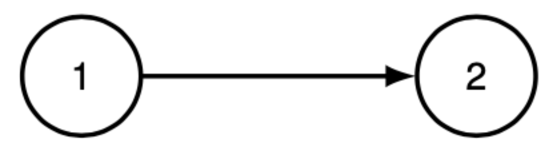
$p = 2$ search over 3 DAGs

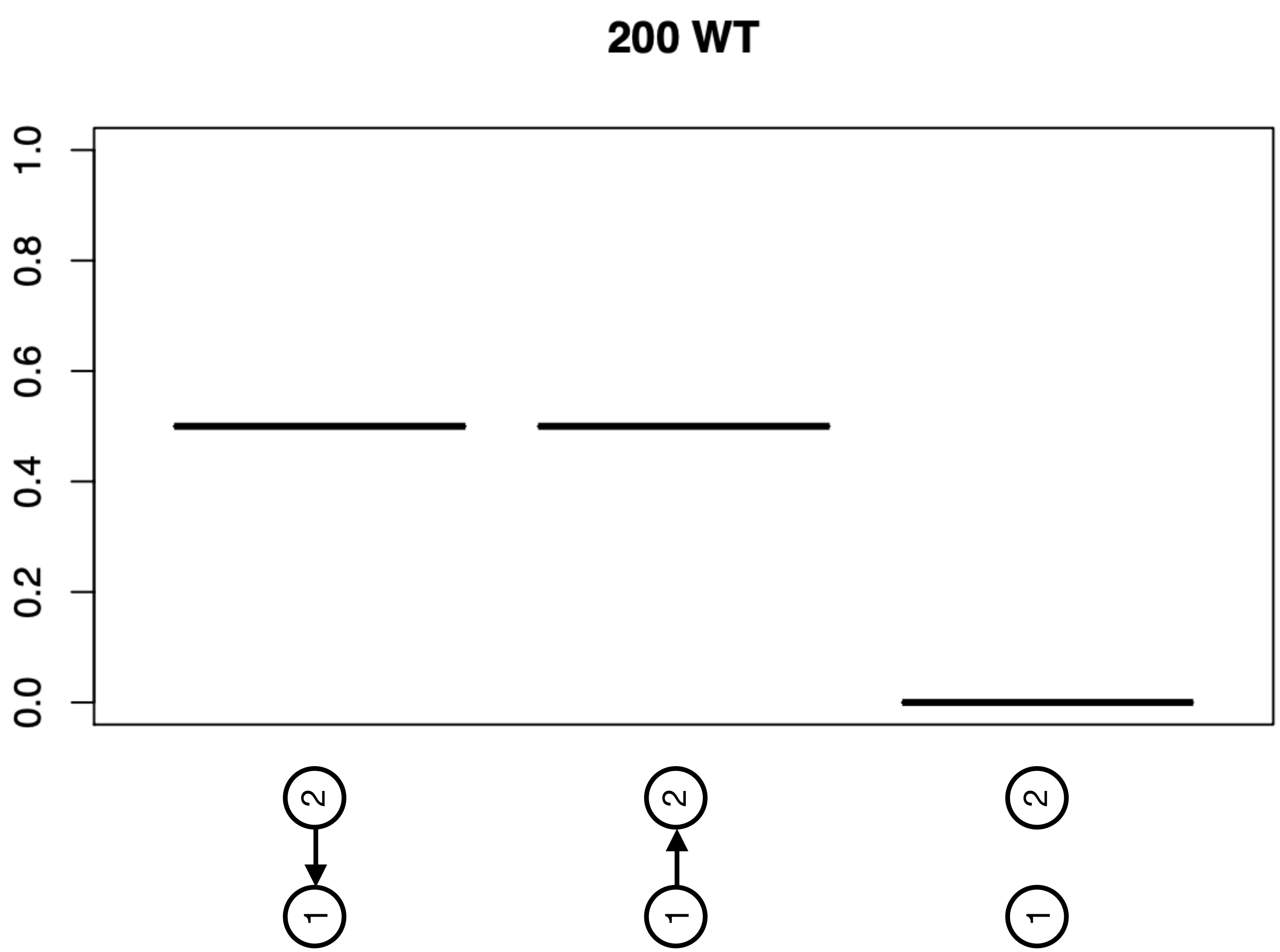
dag $=$ [ ] [ ]



$p = 2$ search over 3 DAGs

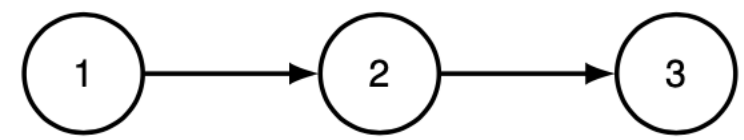
dag $=$ `[] [1]`

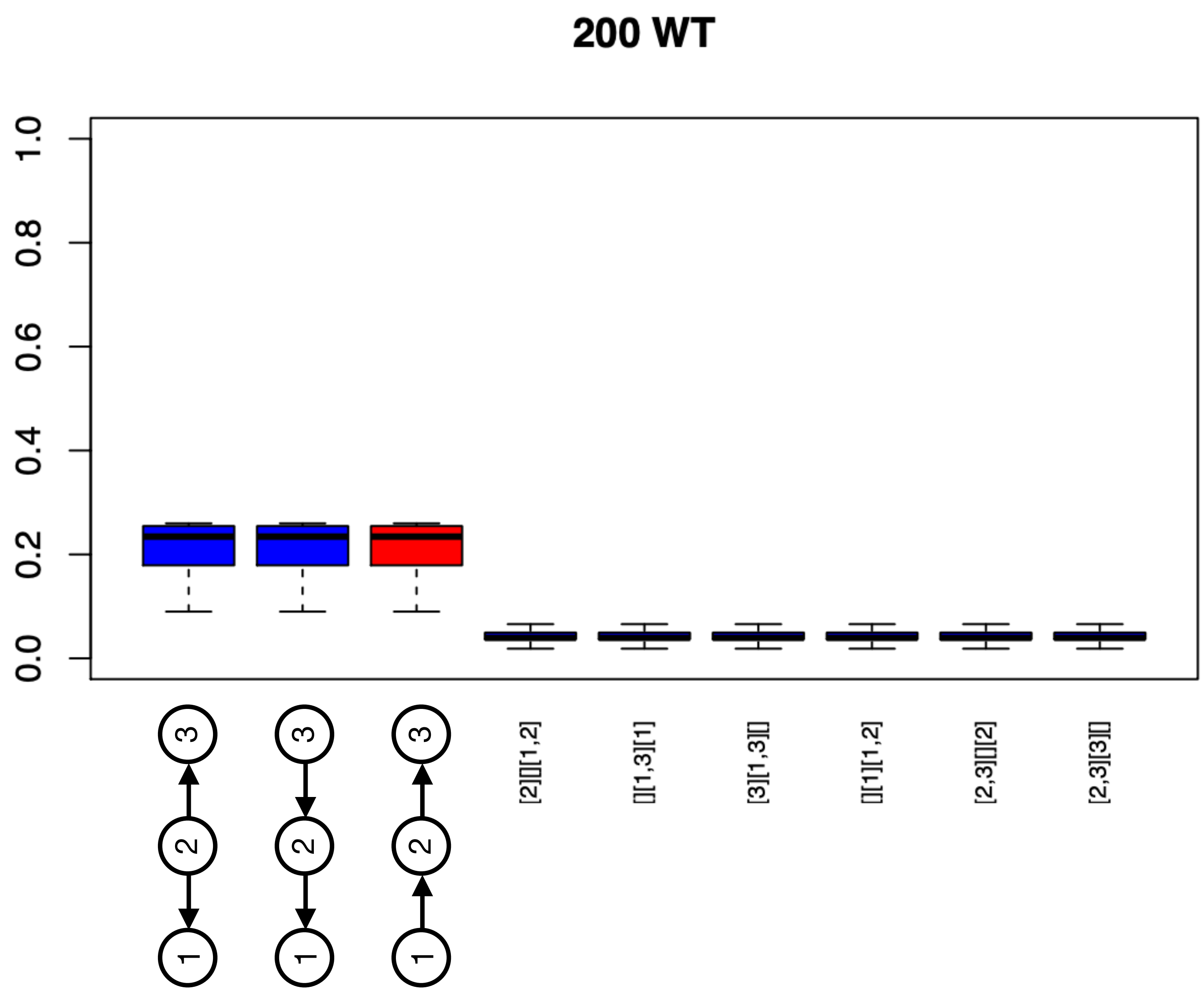




$p = 3$ search over 25 DAGs

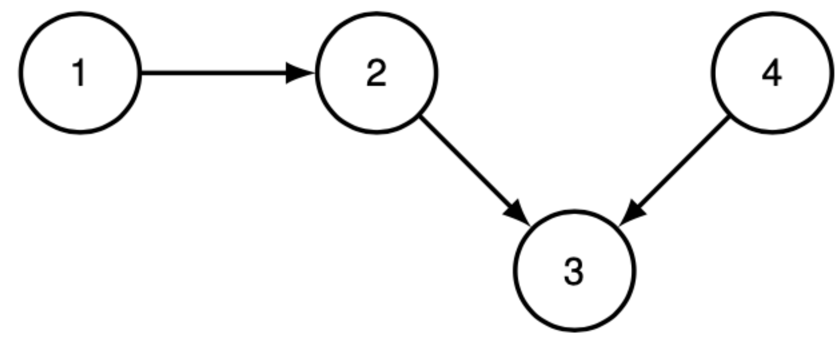
dag $=$ `[][1][2]`

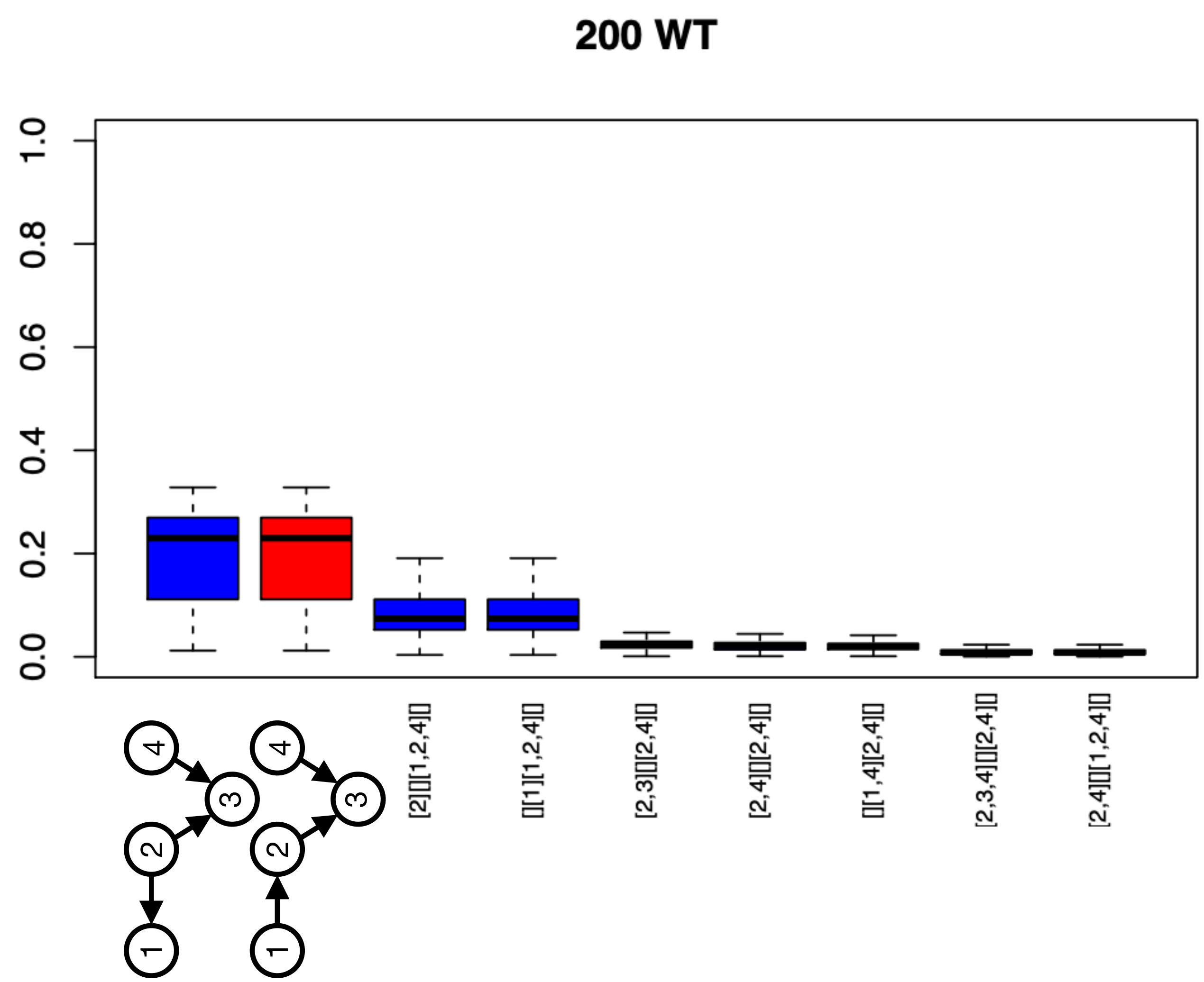




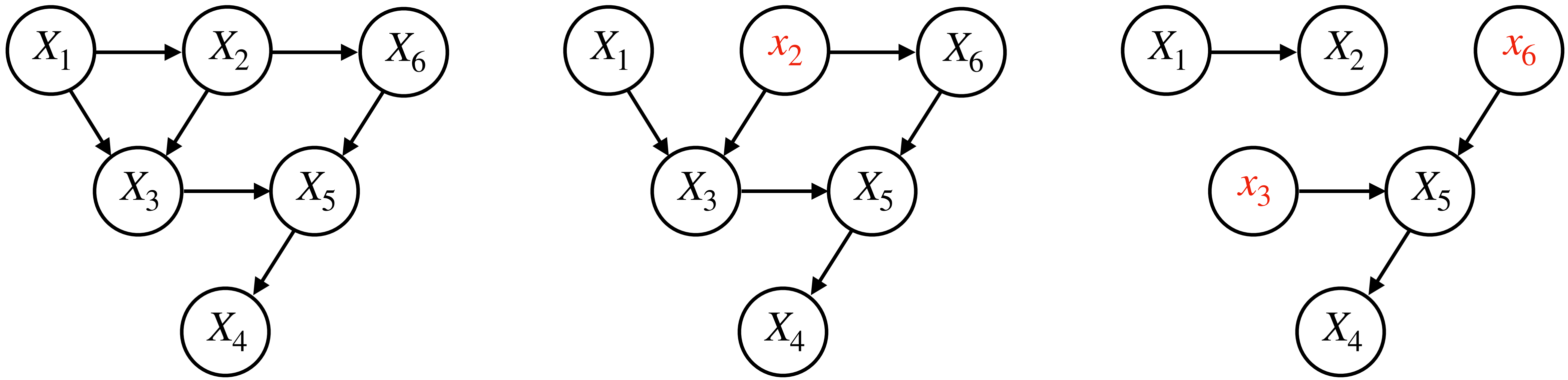
$p = 4$ search over 543 DAGs

dag $=$ `[][1][2,4][]`





Interventions: Do operator



Observation $\qquad \mathrm{Do}(X_2 = x_2) \qquad \mathrm{Do}(X_3 = x_3, X_6 = x_6)$

Example of interventions:

- Clinical randomization $\mathrm{Do}(T = t)$
- Gene knock-out $\mathrm{Do}(G = 0)$
- Knock-down/up
- Functional knock-out

$$\mathbb{P}(X \mid \mathrm{Do}(Y = y)) \neq \mathbb{P}(X \mid Y = y)$$

Causal Gaussian BN

Causal GBN with parameter $\theta = (w, m, \sigma)$: let us denote by X_j the expression of gene $j \in \{1, \ldots, p\}$ then we have:

$$X_j = m_j + \sum_{i \in \mathrm{pa}(j)} w_{i,j} X_i + \varepsilon_j \text{ with } \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$$

with $w_{i,j} \neq 0$ if and only if $i \in \mathrm{pa}(j)$. NB: with a proper *causal ordering*[1] such that $i \in \mathrm{pa}(j) \Rightarrow i < j$ $\boldsymbol{W} = (w_{i,j})$ is upper triangular. $\boldsymbol{W}$ is hence a nilpotent matrix with $\boldsymbol{W}^p = \mathbf{0}$.

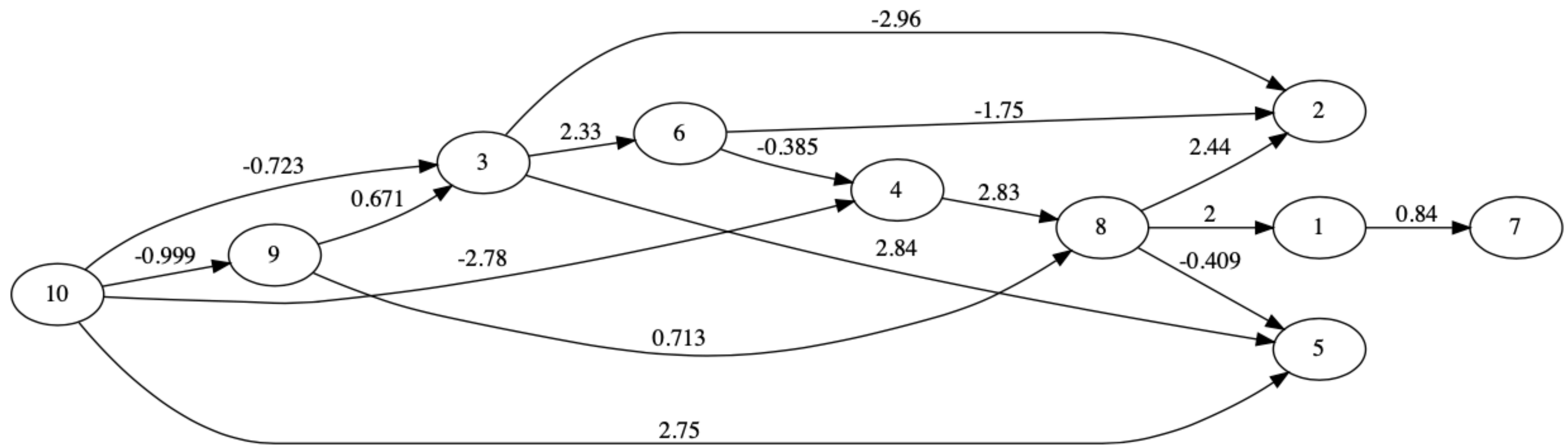
- Direct causal effects $\boldsymbol{W} = (w_{i,j})$
- Total causal effects $\boldsymbol{L} = (\ell_{i,j}) = (\boldsymbol{I} - \boldsymbol{W})^{-1} = \boldsymbol{I} + \boldsymbol{W} + \ldots + \boldsymbol{W}^{p-1}$

$$w_{i,j} = \frac{d}{dx}\mathbb{E}[X_j | X_{-j}, \mathrm{do}(X_i = x)] \quad \ell_{i,j} = \frac{d}{dx}\mathbb{E}[X_j | \mathrm{do}(X_i = x)]$$

[1] also called *topological ordering* in a DAG.

Example

A random DAG with $p = 10$ genes



j	1	2	3	4	5	6	7	8	9	10
m	-0.61	-0.41	1.14	-1.84	1.00	0.71	-1.31	-0.96	0.06	0.70
σ	1.90	1.10	0.77	1.30	0.81	0.72	0.98	1.20	0.91	0.41

Some values (a causal ordering $10, 9, 3, 6, 4, 8, 2, 5, 1, 7$):

$$\mathrm{pa}(1) = \{8\} \quad \mathrm{pa}(4) = \{6, 10\} \quad \mathrm{pa}(10) = \emptyset$$

$$w_{6,2} = -1.75 \quad \ell_{6,2} = w_{6,2} + w_{6,4} \times w_{4,8} \times w_{8,2} = -4.41$$

MLE with known DAG

For each experiment k, we denote by $\mathcal{J}_k$ the intervention set ($\emptyset$ for no intervention). Each experiment k is only informative for the genes that are *not* in the intervention set $\mathcal{J}_k$.

$$\text{loglik}(\theta) = \sum_{j=1}^{p} \underbrace{\sum_{k, j \notin \mathcal{J}_k} \log \texttt{dnorm} \left(x_{kj}, \mu_j + \sum_{i \in \text{pa}_j} w_{ij} x_{ki}, \sigma_j \right)}_{\text{loglik}_j(\theta)}$$

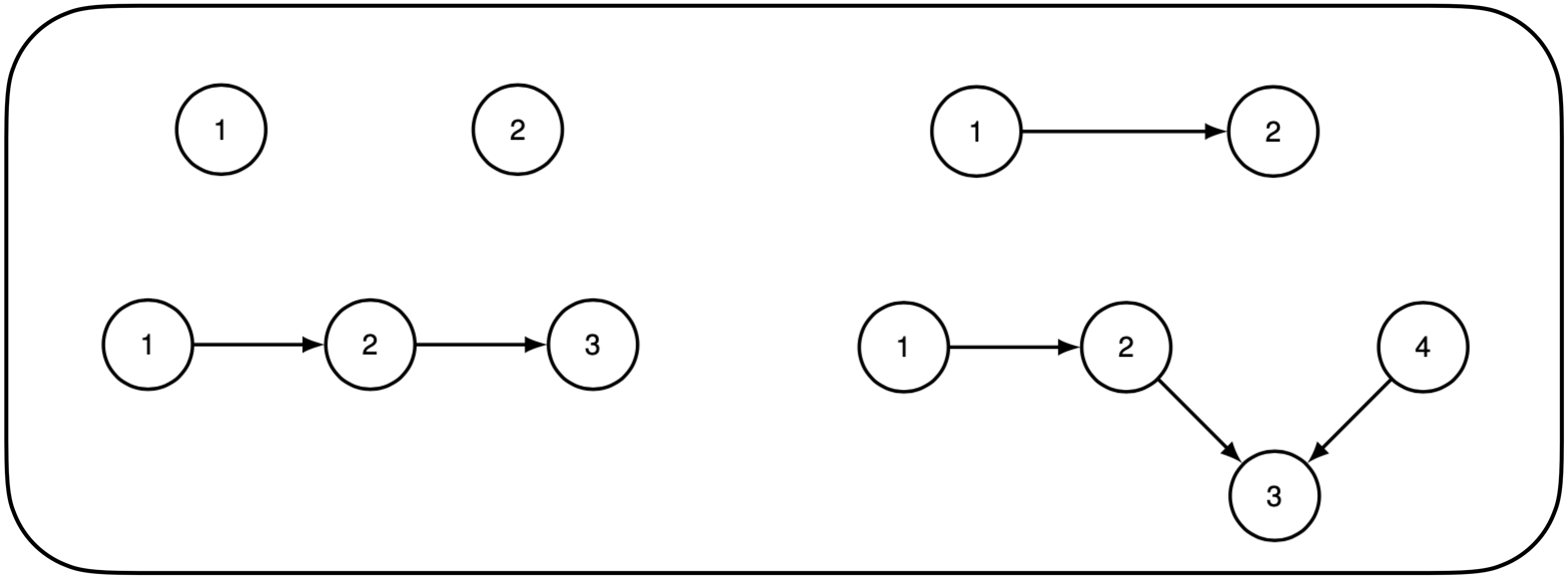
we can therefore estimate μ_j, $w_{\cdot j} = (w_{ij})_{j \in \text{pa}_j}$ and σ_j with classical regression estimators.

For example if $\text{pa}_3 = \{1, 2\}$ we simply do:

- $\texttt{fit} = \texttt{lm}(x_3 \sim 1 + x_1 + x_2, \texttt{data}[\{k, j \notin \mathcal{J}_k\},])$
- $(\hat{\mu}_3, \hat{w}_{13}, \hat{w}_{23}) = \texttt{coef}(\texttt{fit})$ and $\hat{\sigma}_3$=$\texttt{sigma}(\texttt{fit})$

Back to the Toy-examples

Four reference DAGs

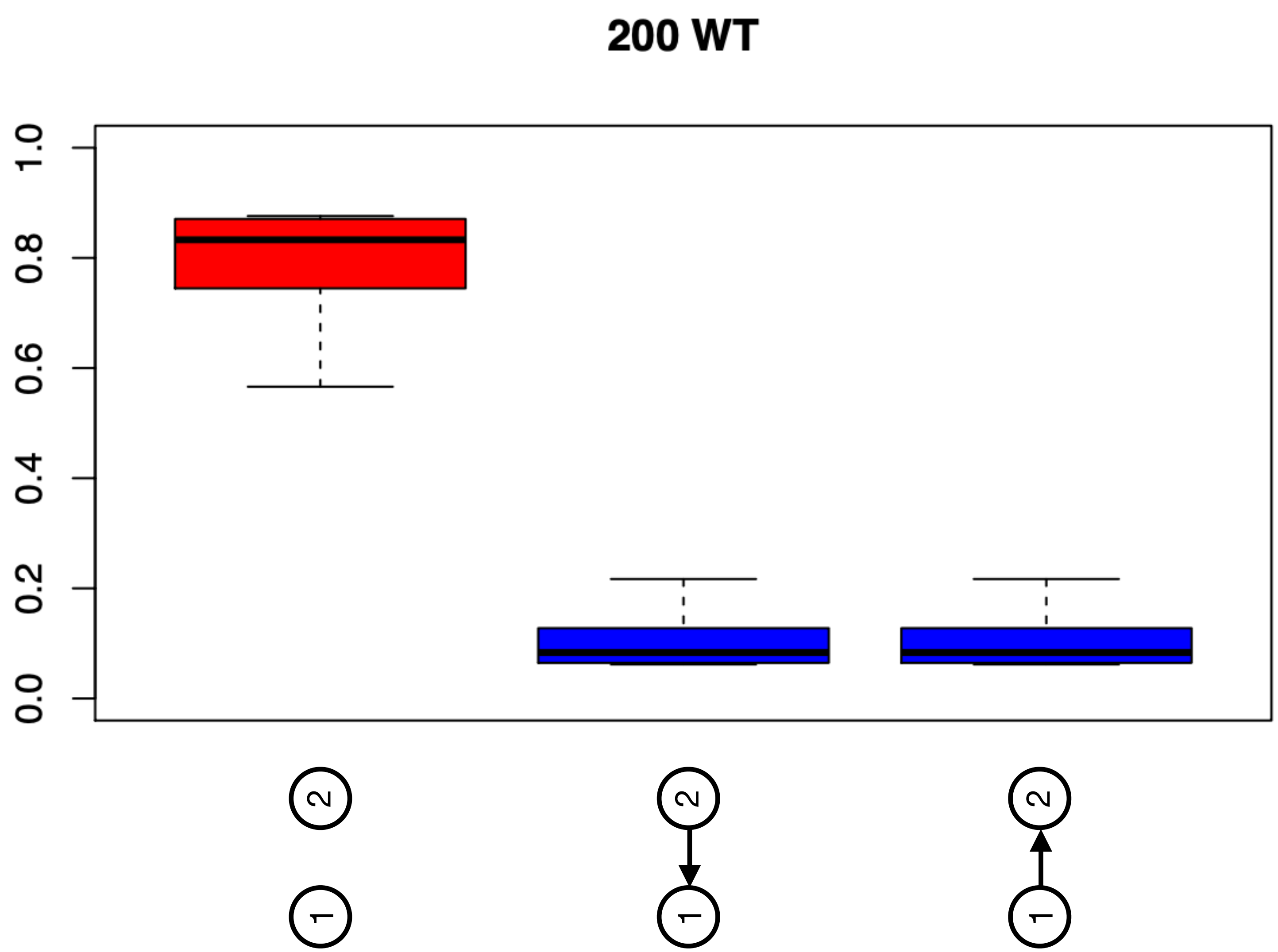


Experiments:

- Simulate 200 observations using a GBN
- Plus interventions !
- Exhaustive search over the DAG space
- Posterior $\mathbb{P}(G \,|\, \text{data})$ over 100 replicates

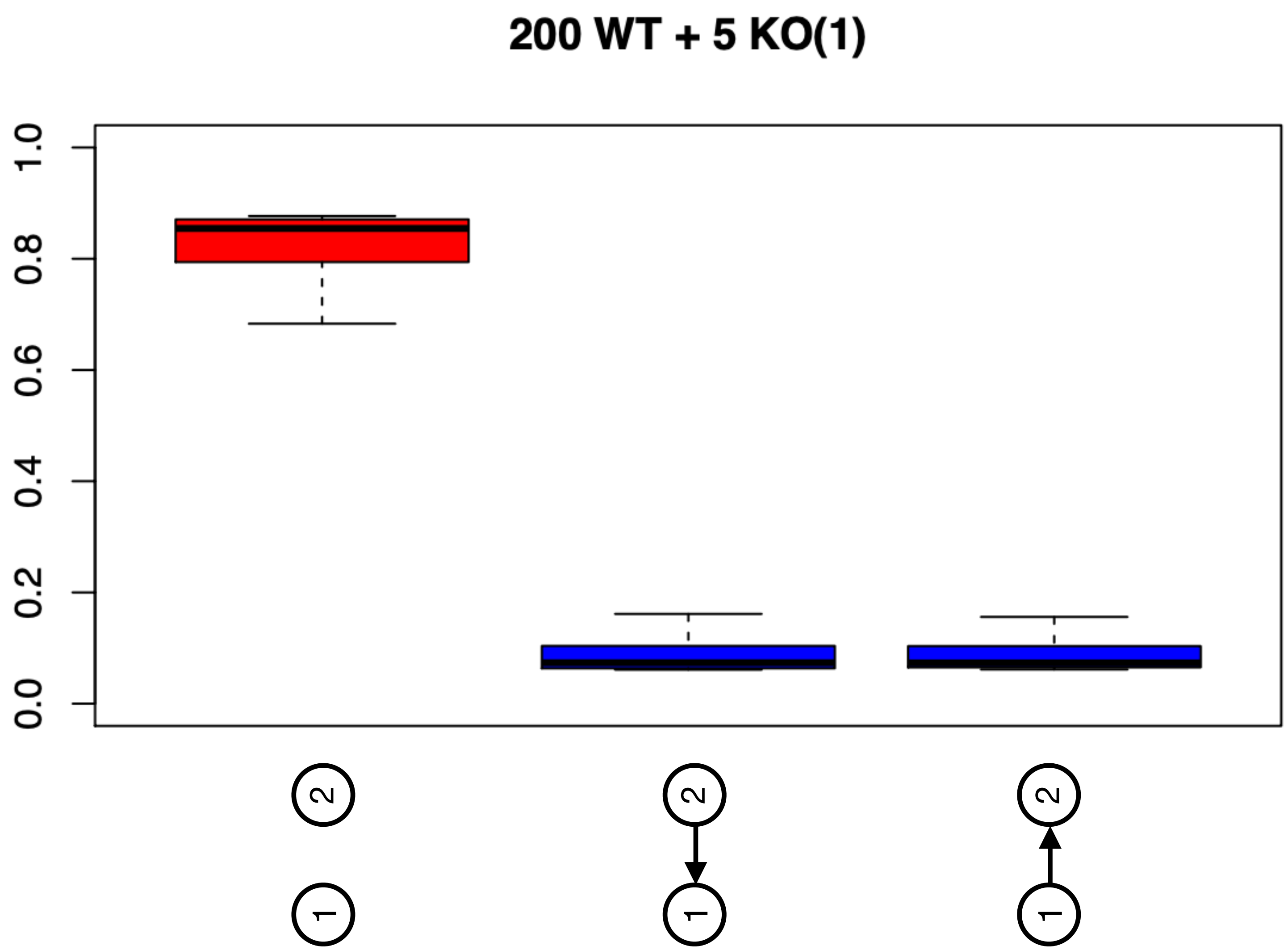
$p = 2$ search over 3 DAGs

dag = [ ] [ ]



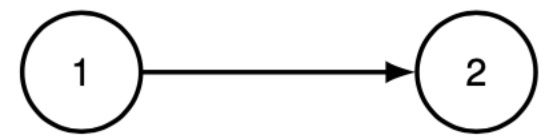
$p = 2$ search over 3 DAGs

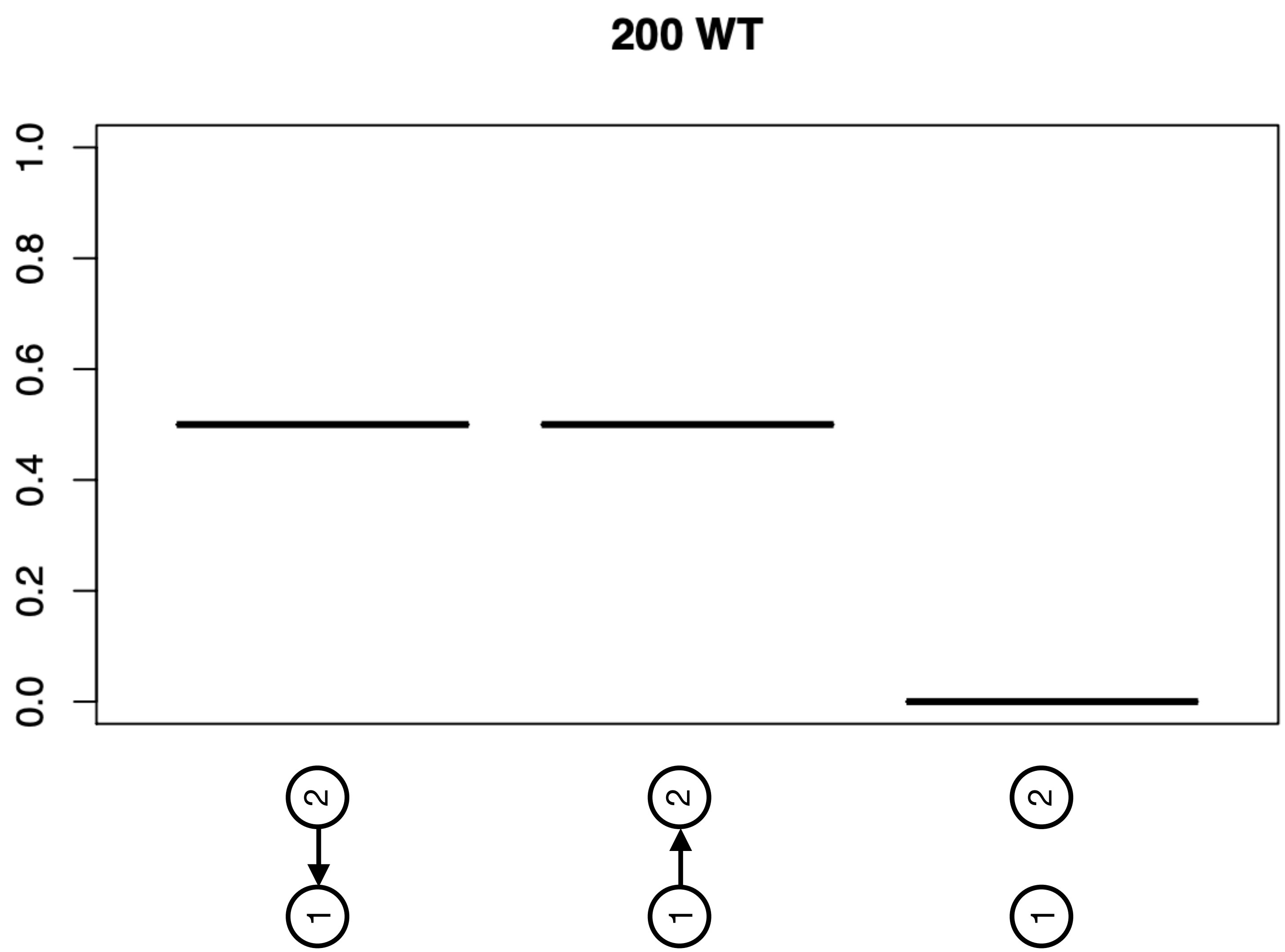
dag = [ ] [ ]



$p = 2$ search over 3 DAGs

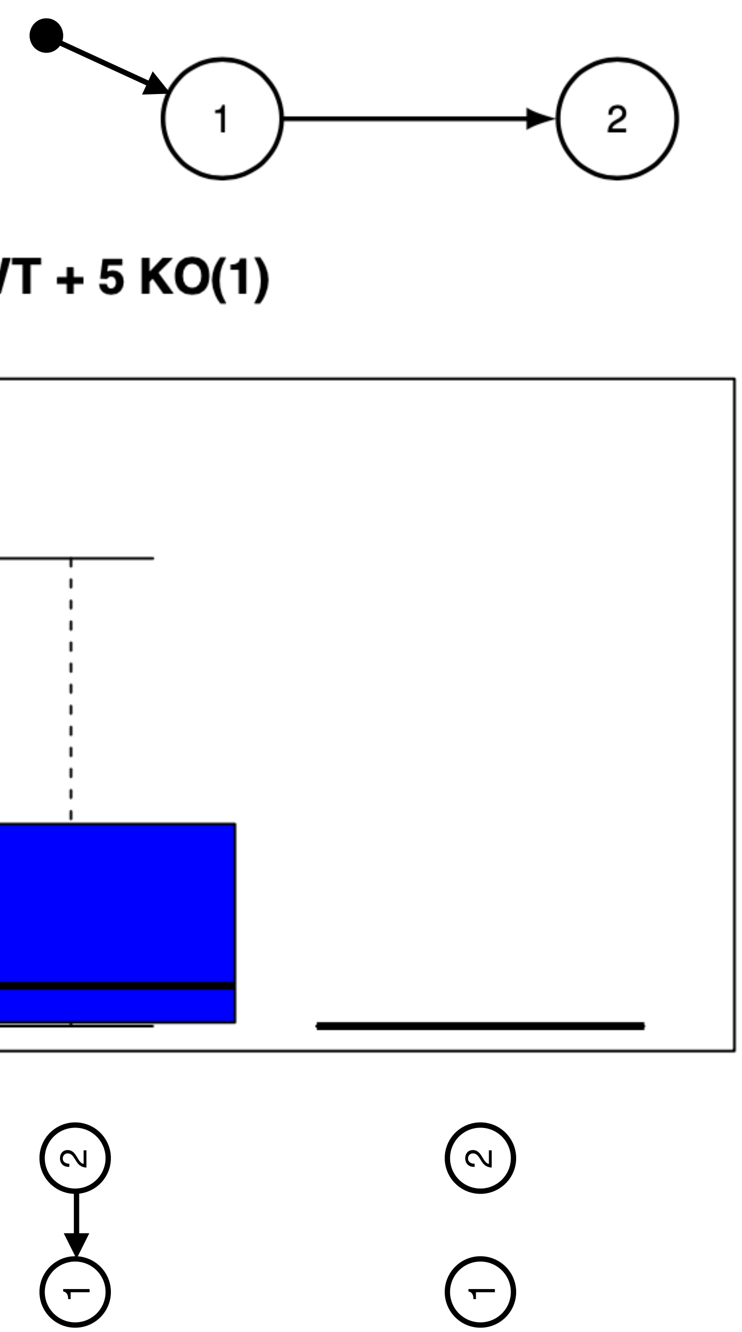
dag $=$ `[][1]`





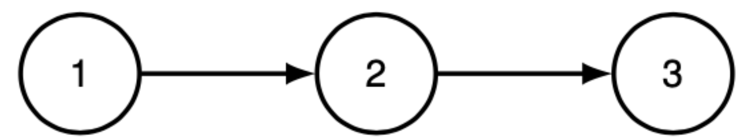
$p = 2$ search over 3 DAGs

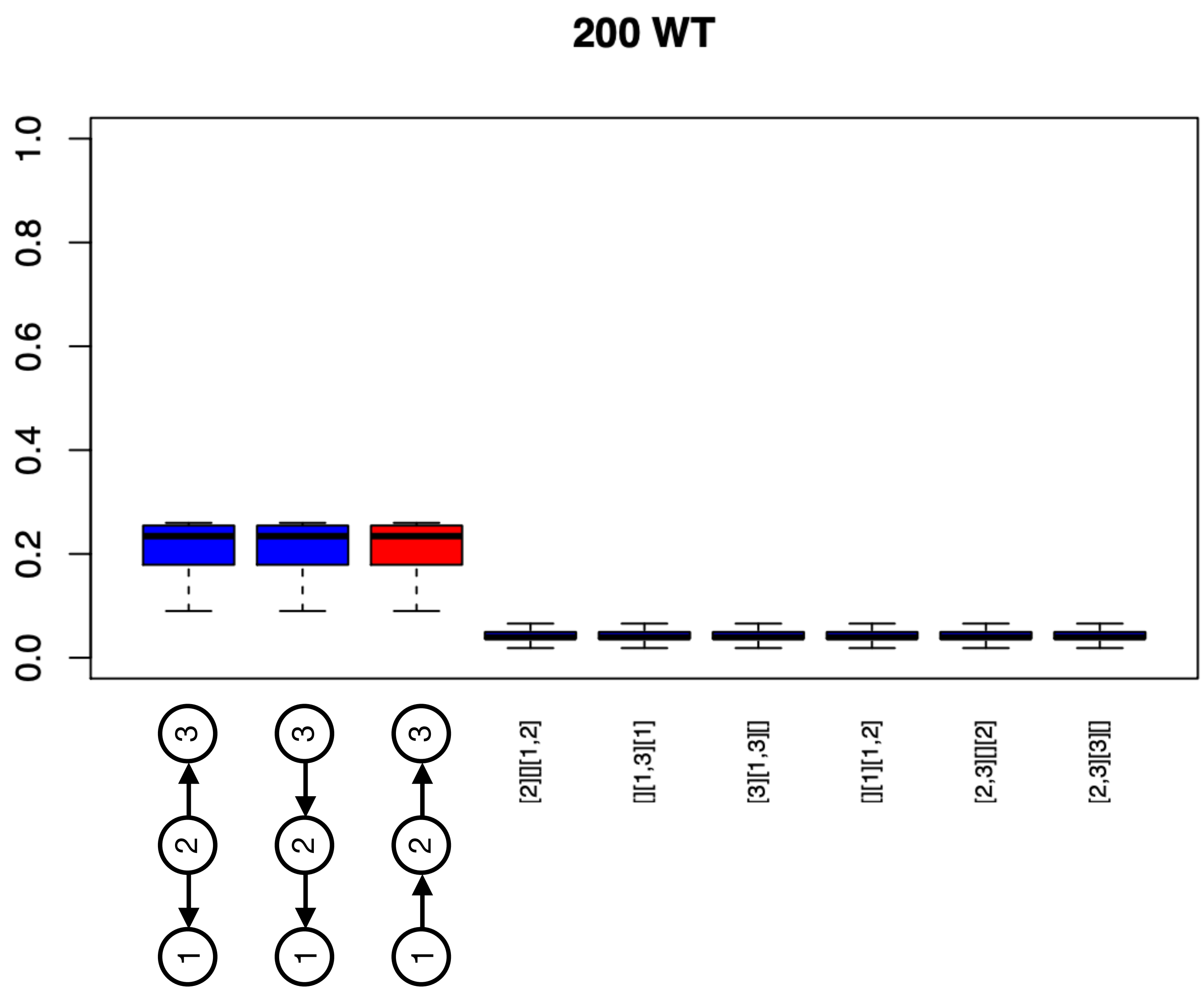
dag $=$ `[][1]`



$p = 3$ search over 25 DAGs

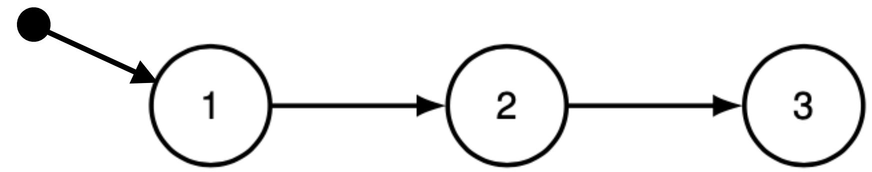
dag $=$ `[][1][2]`

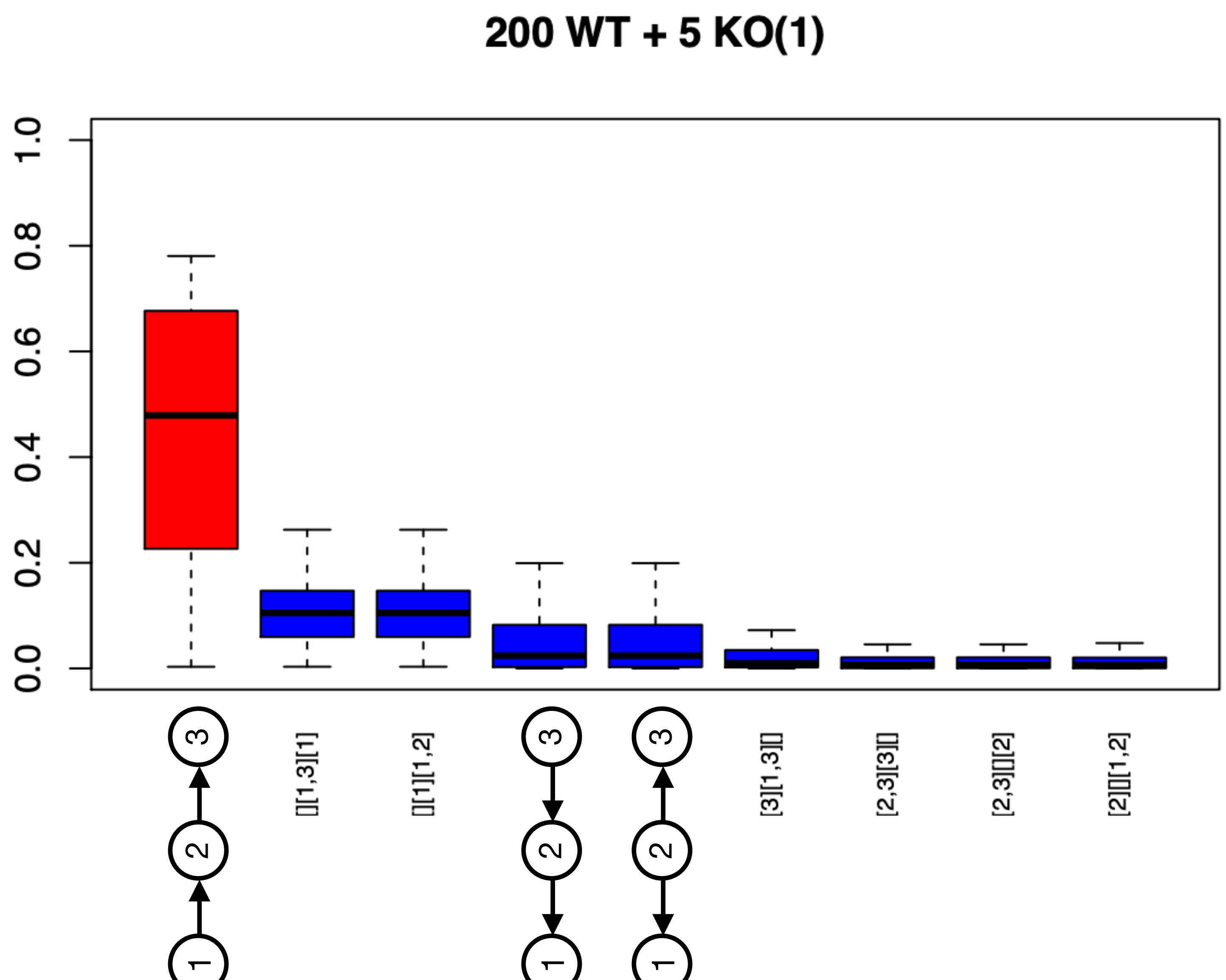




$p = 3$ search over 25 DAGs

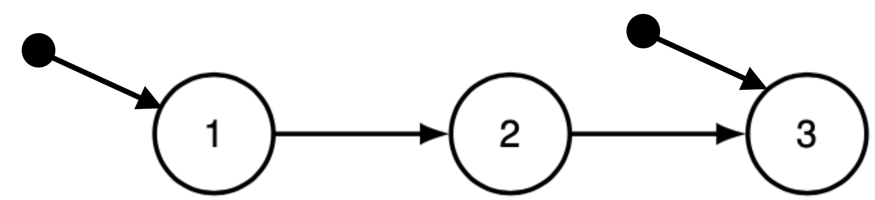
dag = [][1][2]

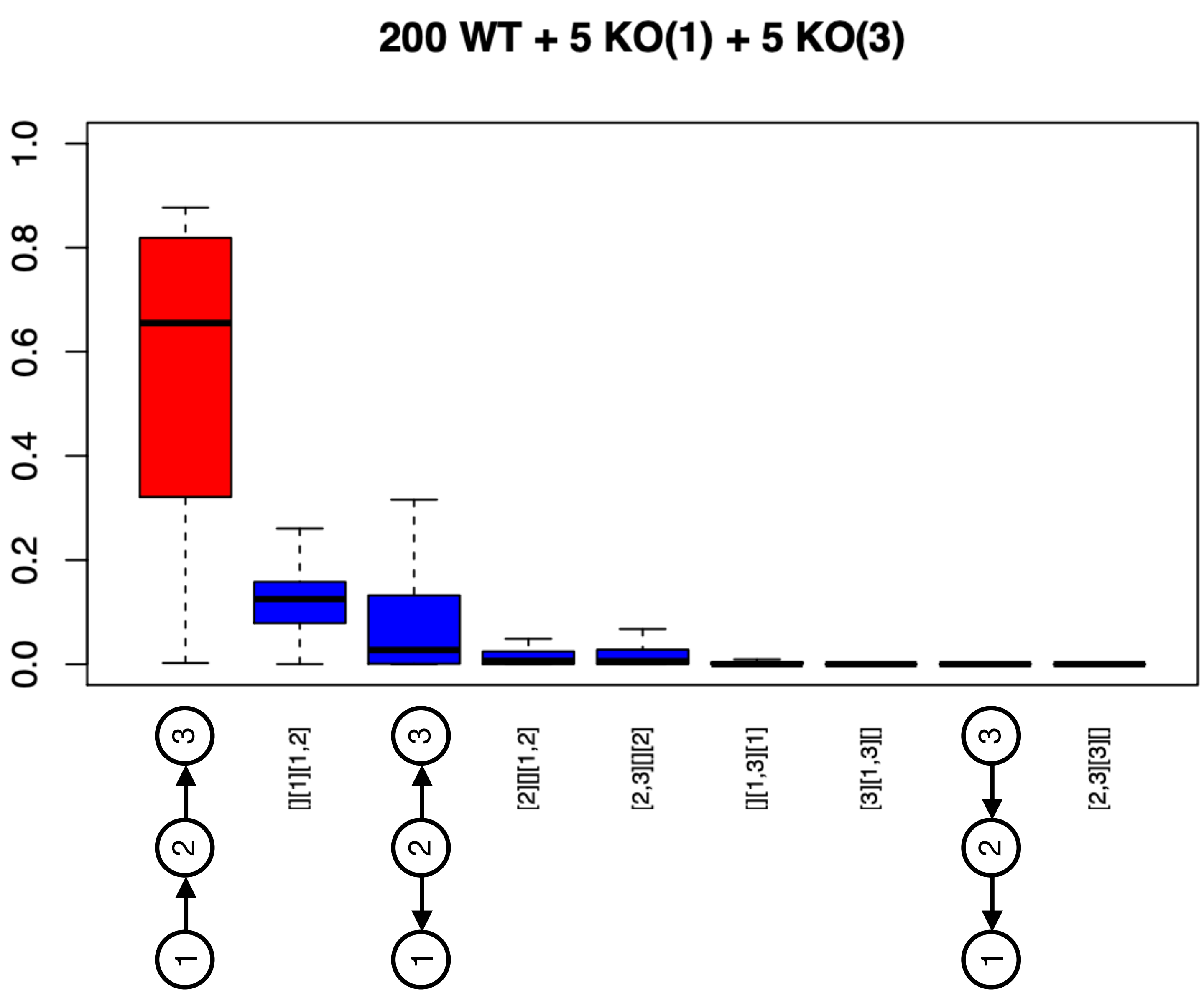




$p = 3$ search over 25 DAGs

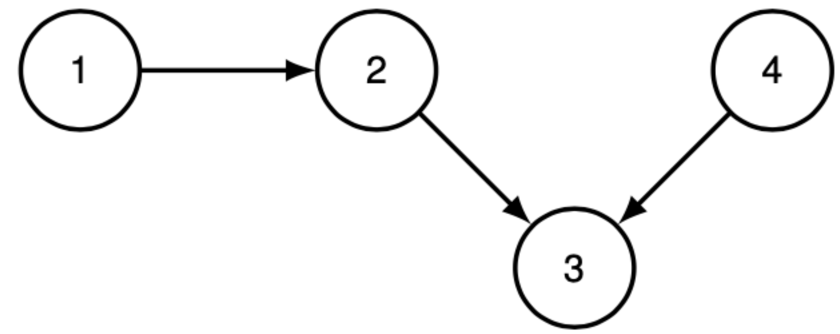
dag $=$ [][1][2]

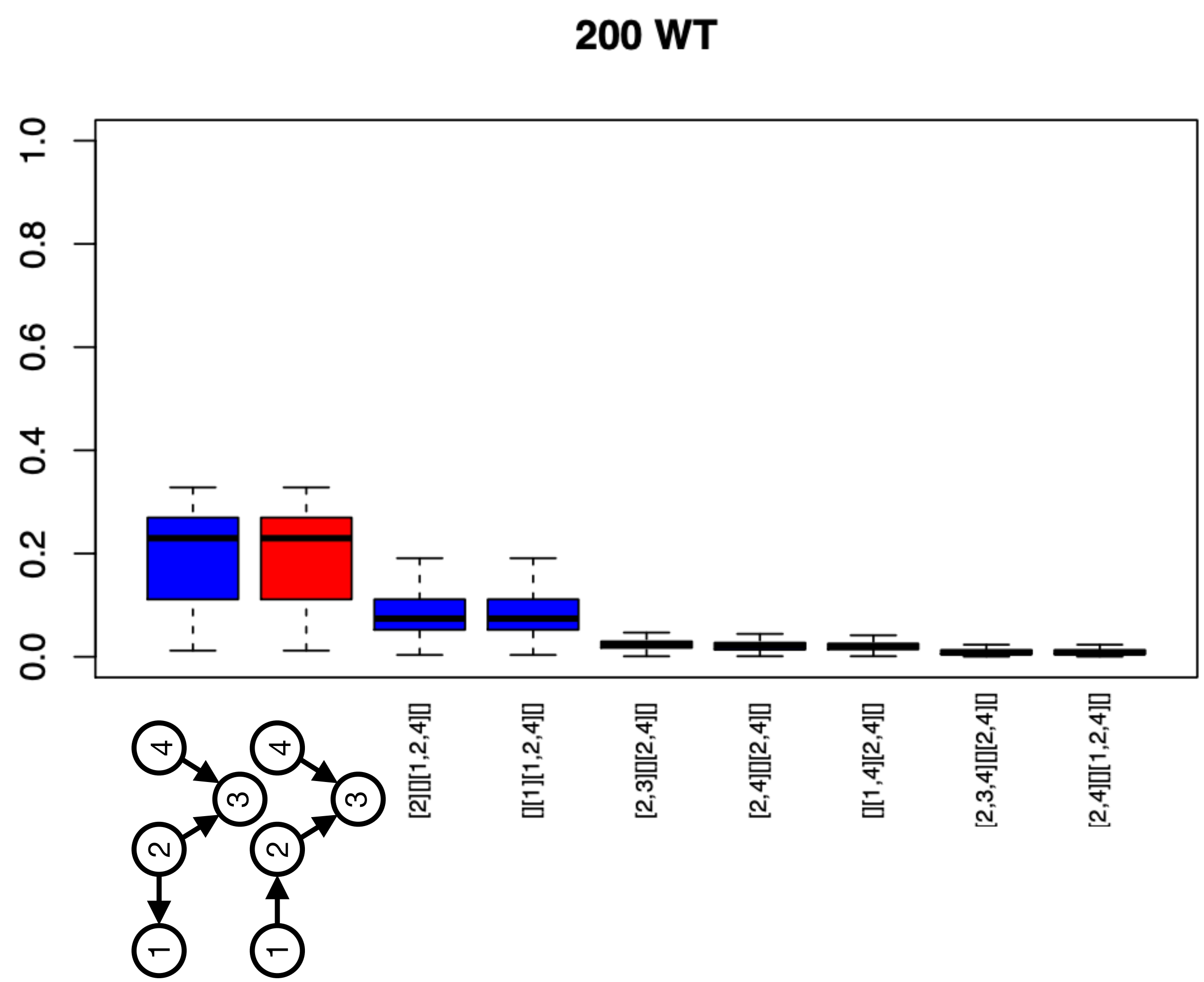




$p = 4$ search over 543 DAGs

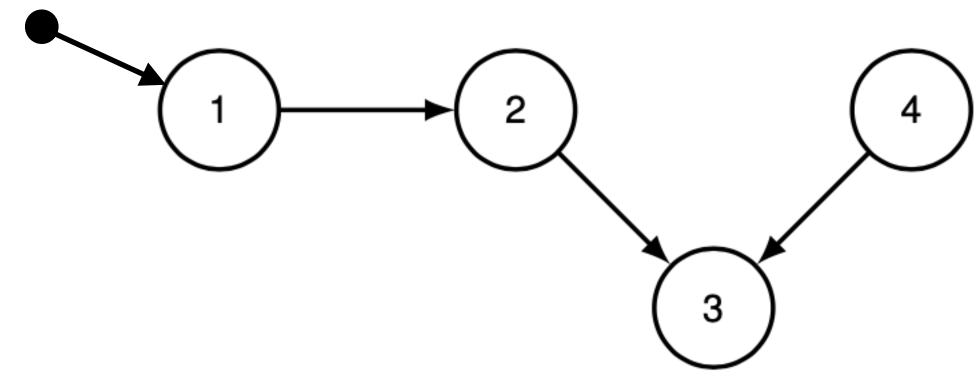
dag $=$ `[][1][2,4][]`

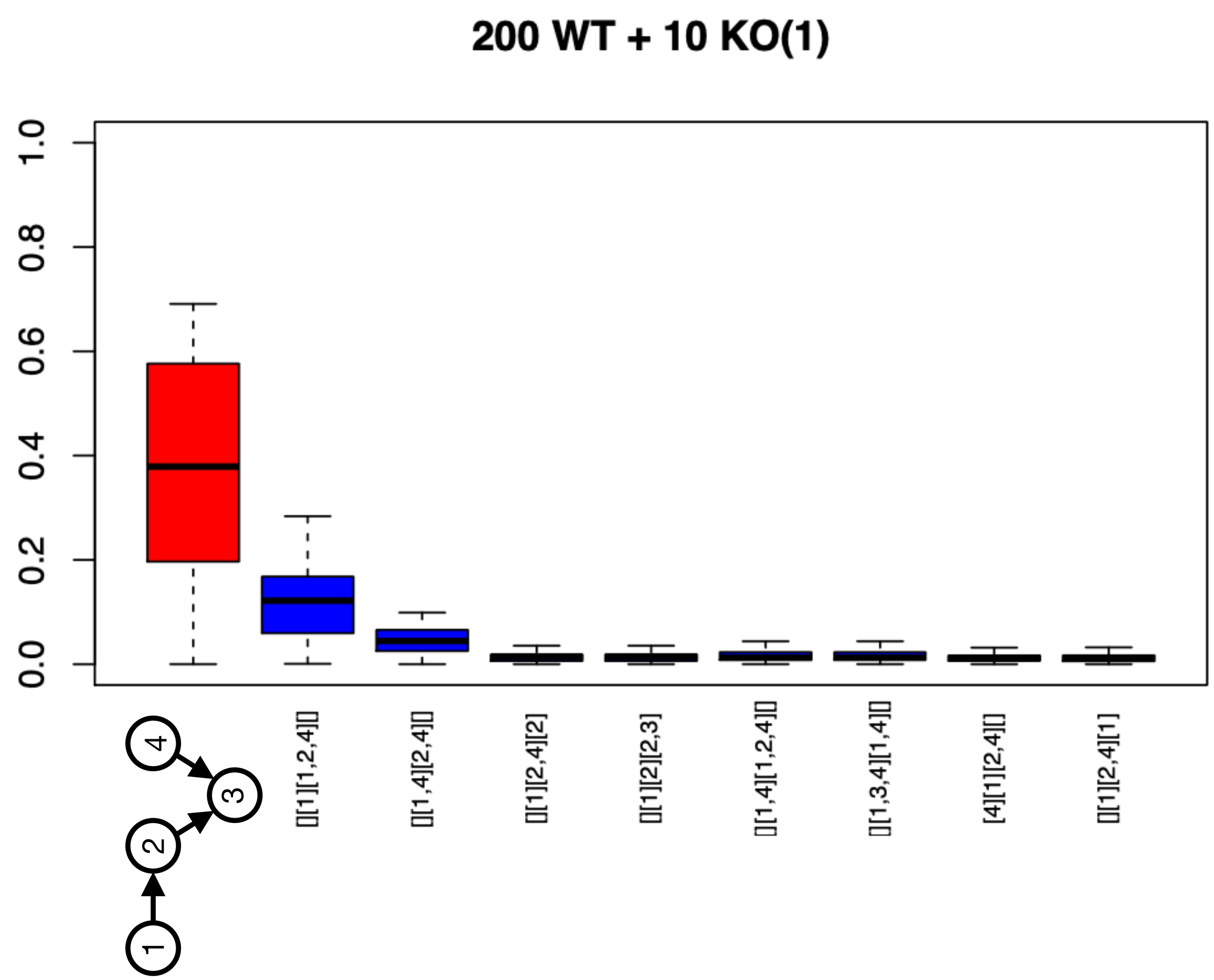




$p = 4$ search over 543 DAGs

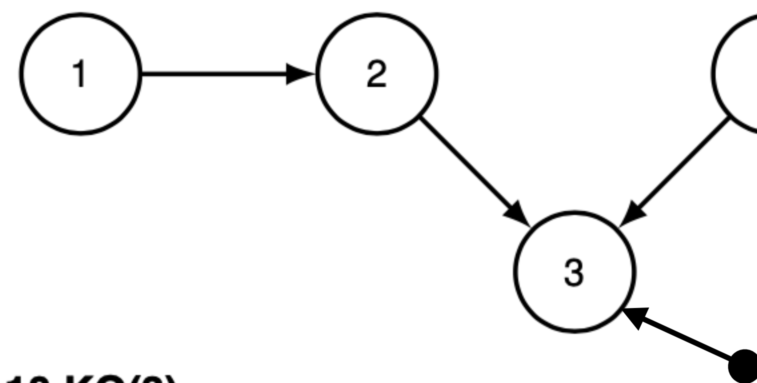
dag $=$ `[][1][2,4][]`

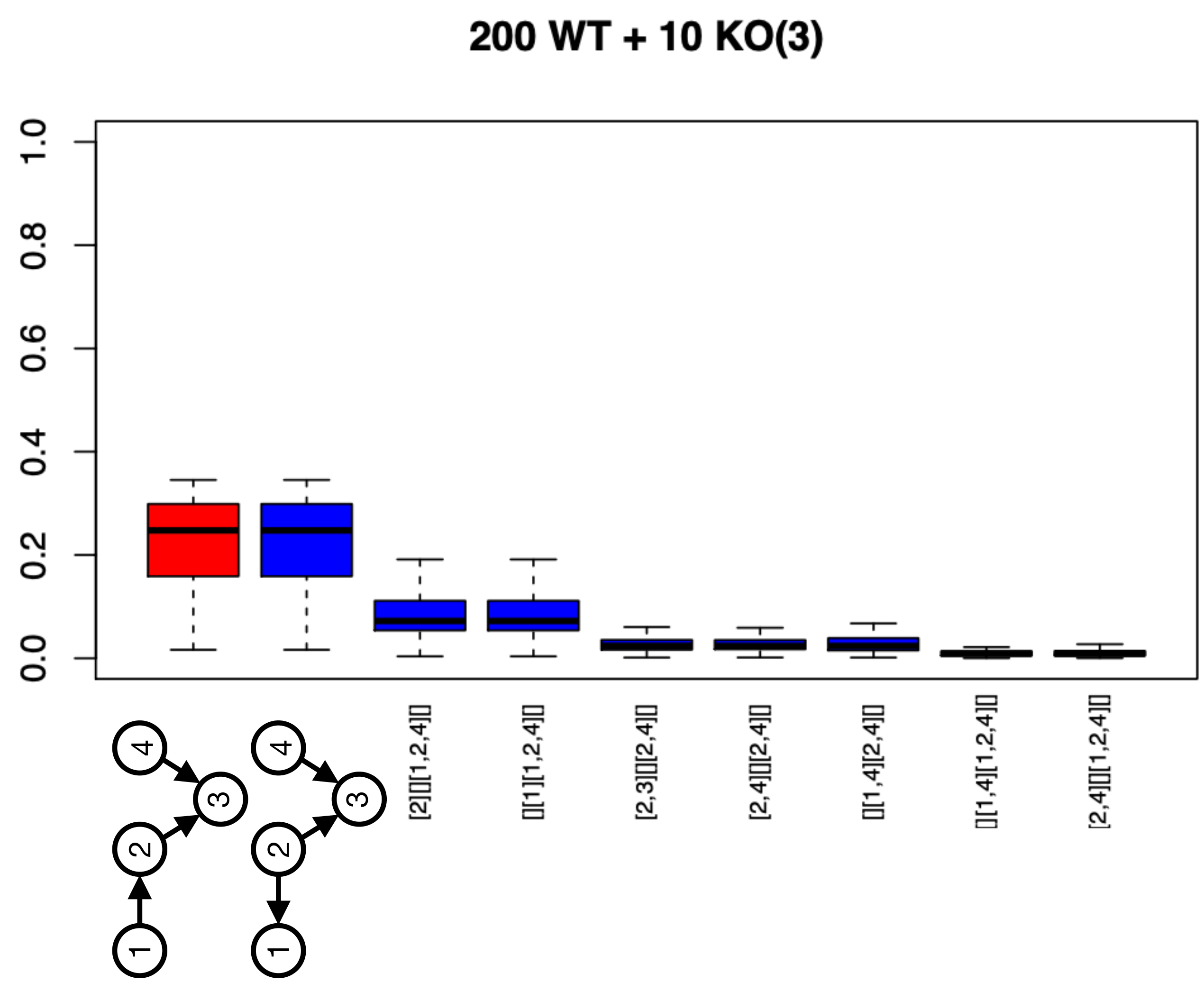




$p = 4$ search over 543 DAGs

dag $=$ `[][1][2,4][]`

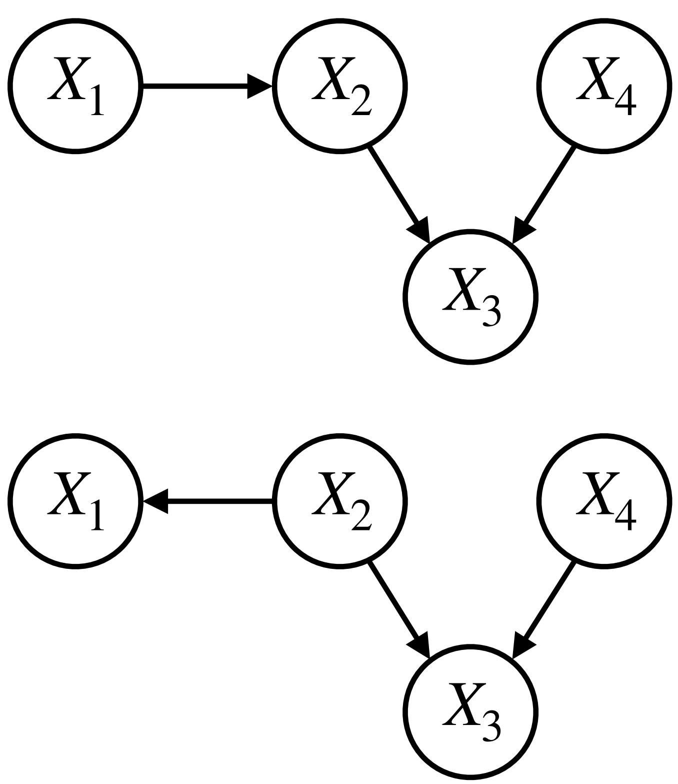




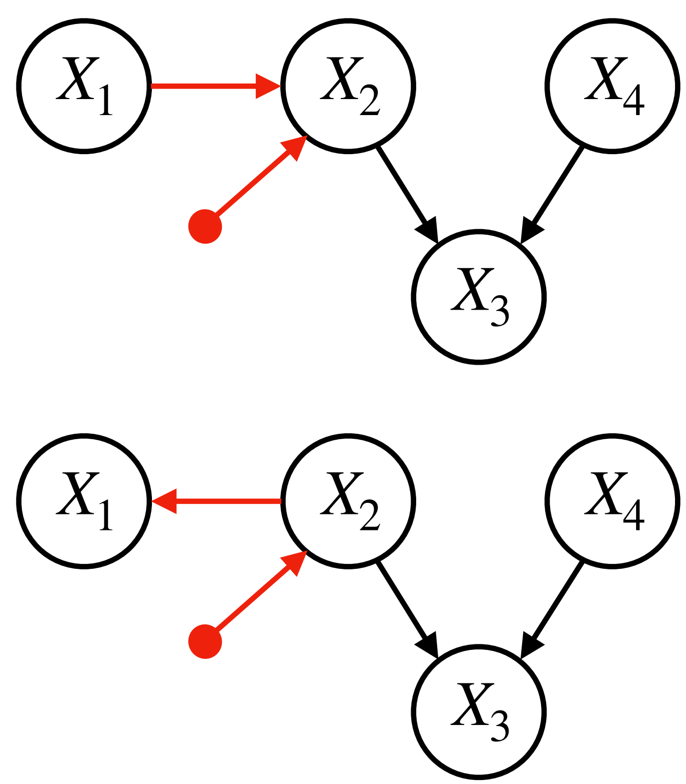
$\mathscr{I}$-Markov Equivalence Class

Theorem (3.9 in Yang *et al* 2018)**:**
Two DAGs are $\mathscr{I}$-Markov equivalent with $\emptyset \in \mathscr{I}$ if and only if they have the same *skeleton* and the same *v-structures*.

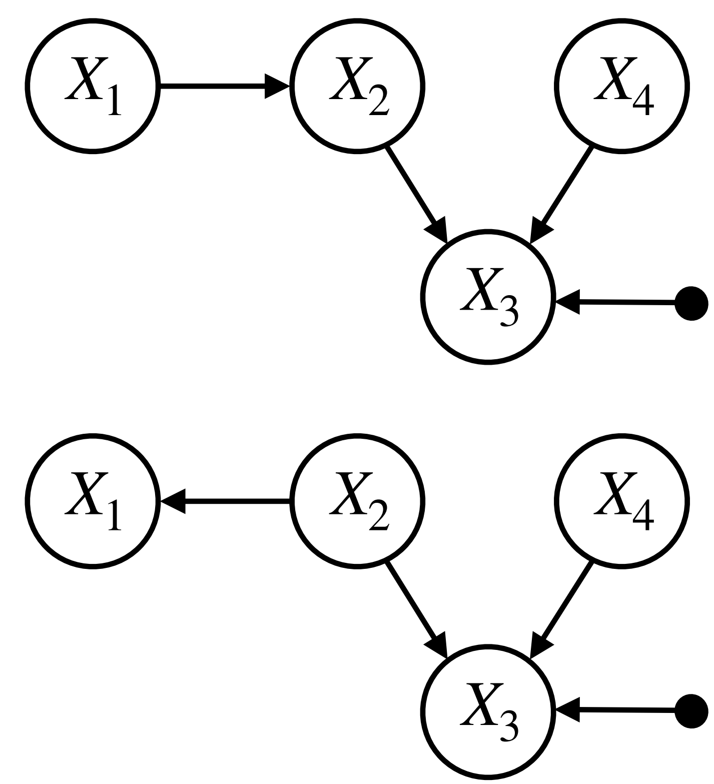
NB: strongly protected arrows (not from intervention node) of a $\mathscr{I}$-DAG with $\emptyset \in \mathscr{I}$
Are *exactly* the strongly $\mathscr{I}$-protected arrows of the DAG (Hauser & Bühlmann, 2012)



ME DAGs

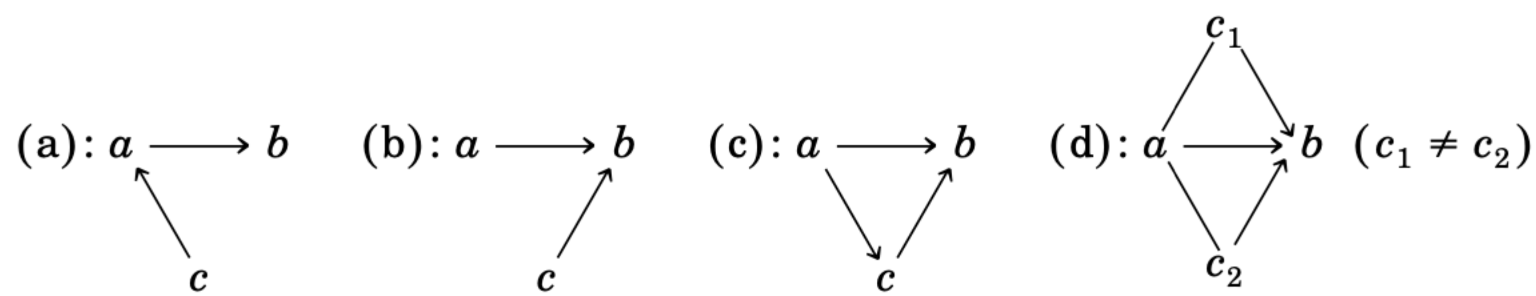


not $\mathscr{I}$-ME DAGs

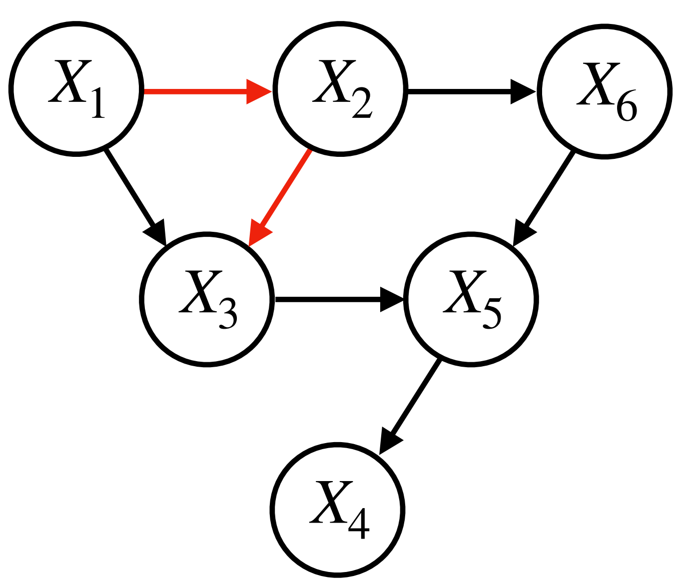


$\mathscr{I}$-ME DAGs

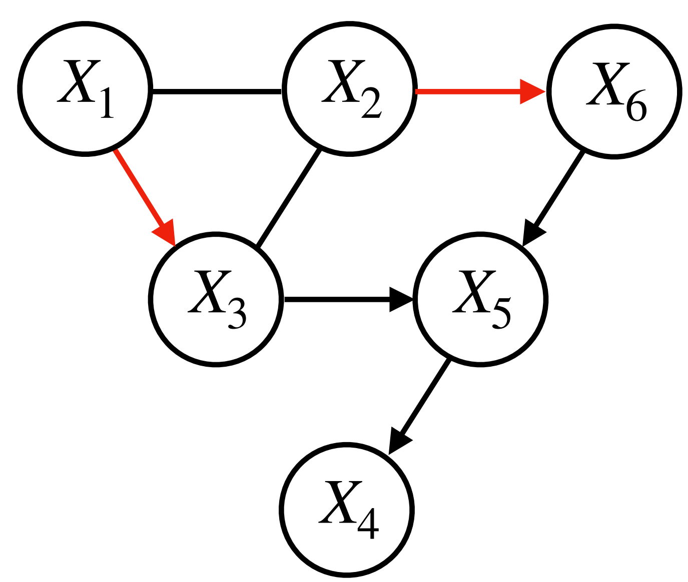
$\mathscr{I}$-CPDAG: $\mathscr{I}$-Completed Partially Directed Acyclic Graph



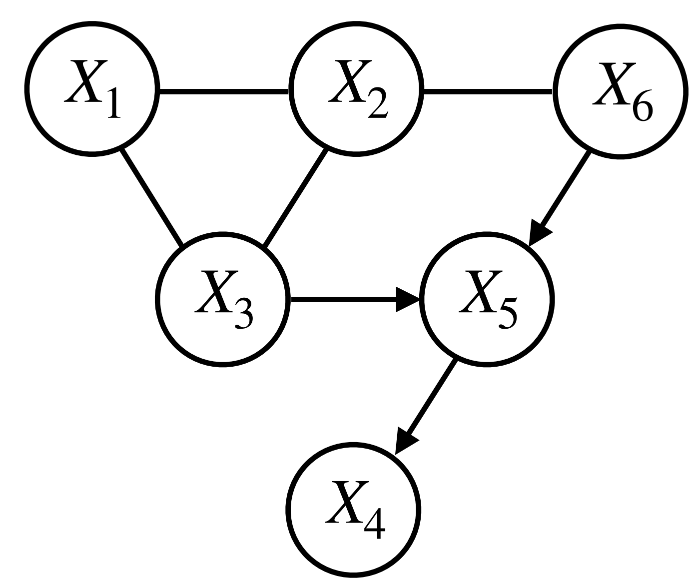
$a \rightarrow b$ strongly protected (Andersson *et al*,1997)



Initial DAG

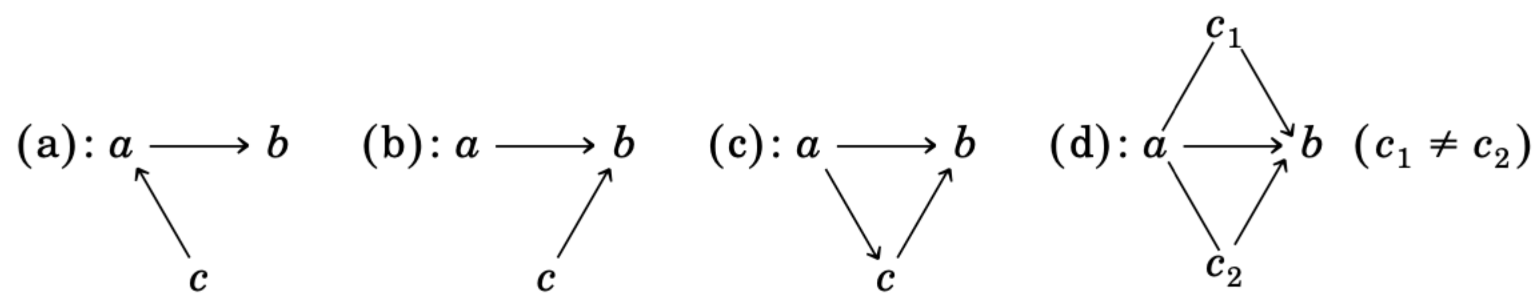


Intermediary PDAG

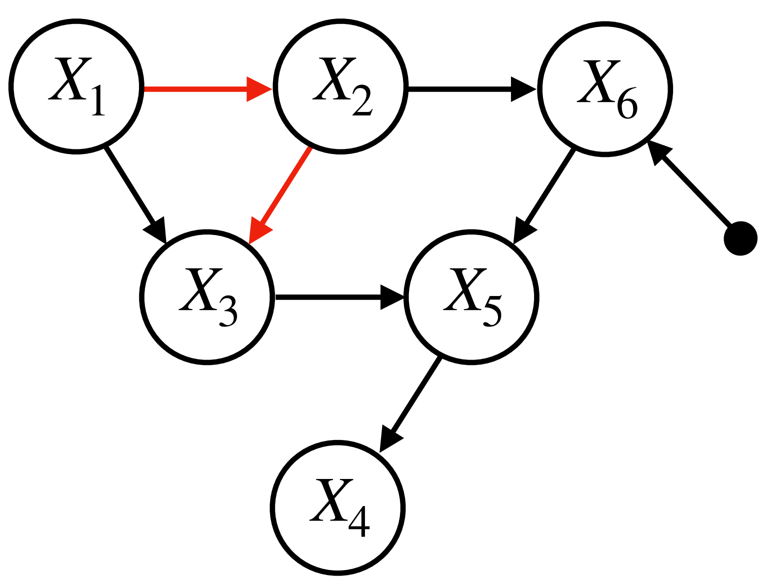


CPDAG

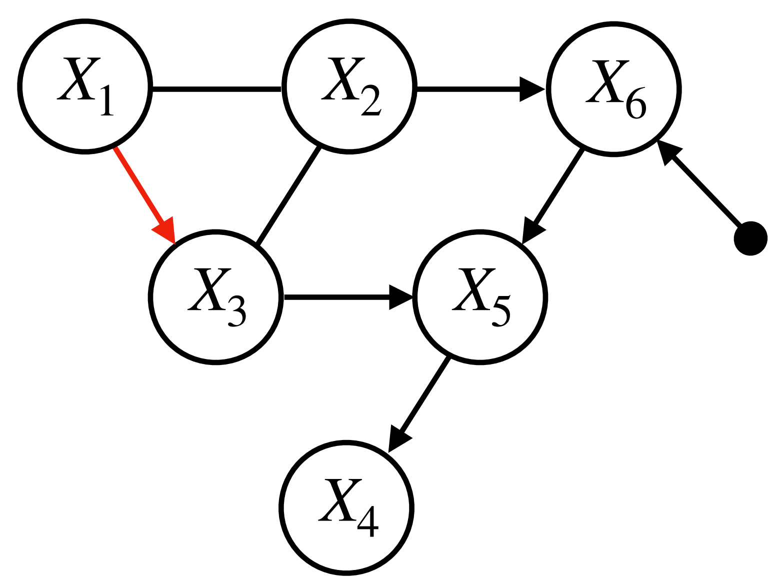
$\mathscr{I}$-CPDAG: $\mathscr{I}$-Completed Partially Directed Acyclic Graph



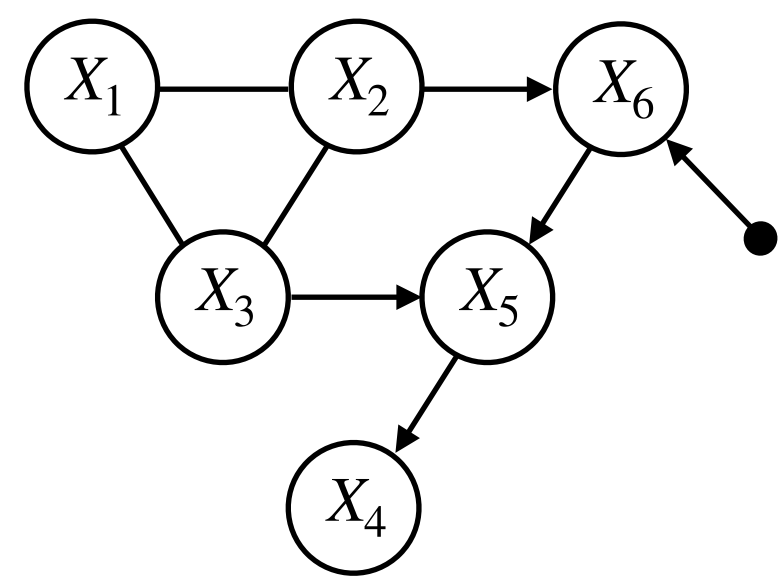
$a \to b$ strongly protected (Andersson *et al*,1997)



Initial $\mathscr{I}$-DAG

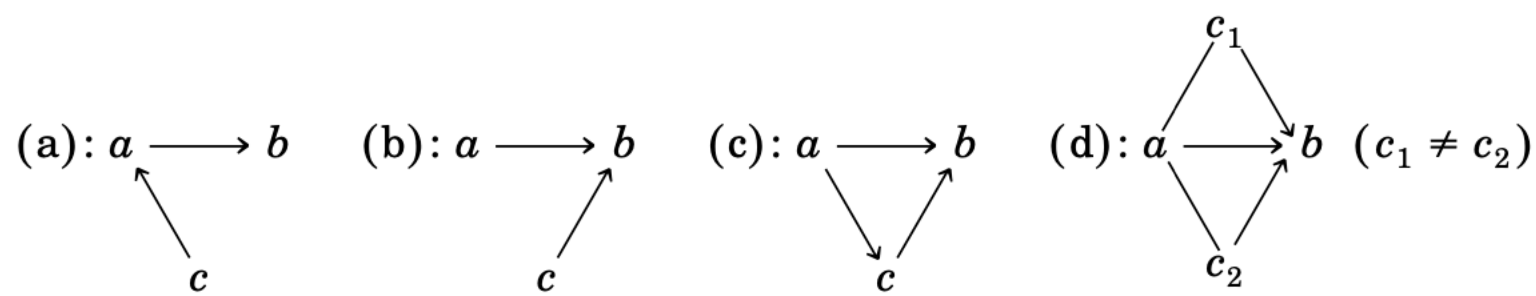


Intermediary $\mathscr{I}$-PDAG

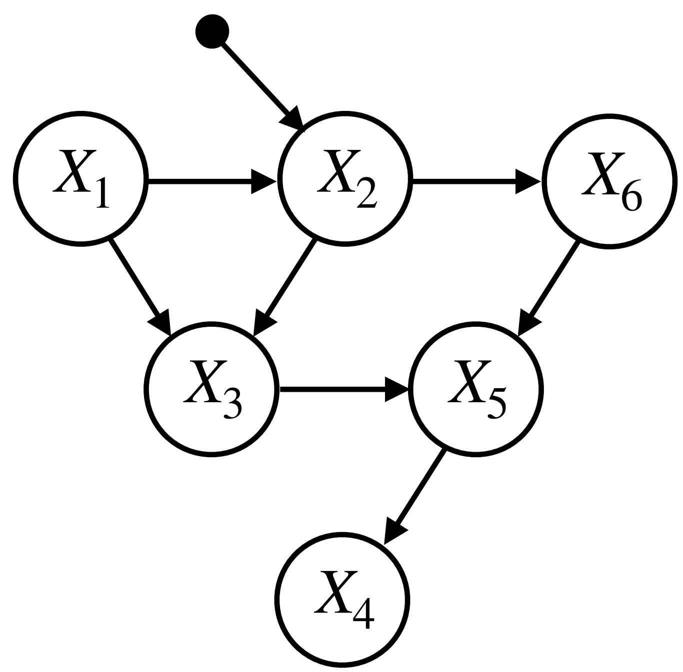


$\mathscr{I}$-CPDAG

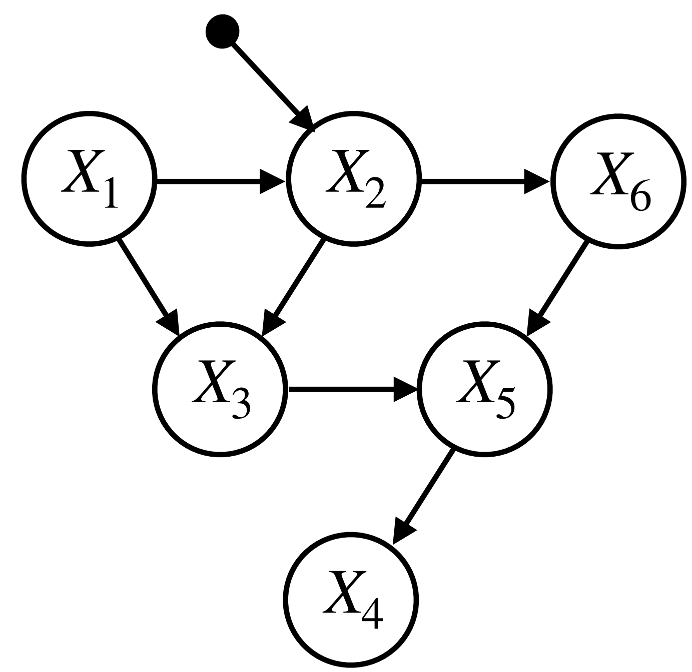
$\mathscr{I}$-CPDAG: $\mathscr{I}$-Completed Partially Directed Acyclic Graph



$a \rightarrow b$ strongly protected (Andersson *et al*,1997)

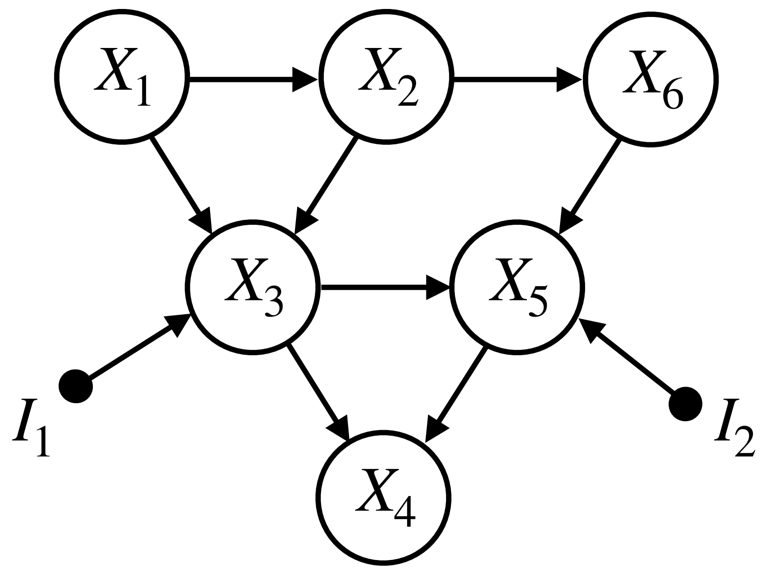


Initial $\mathscr{I}$-DAG

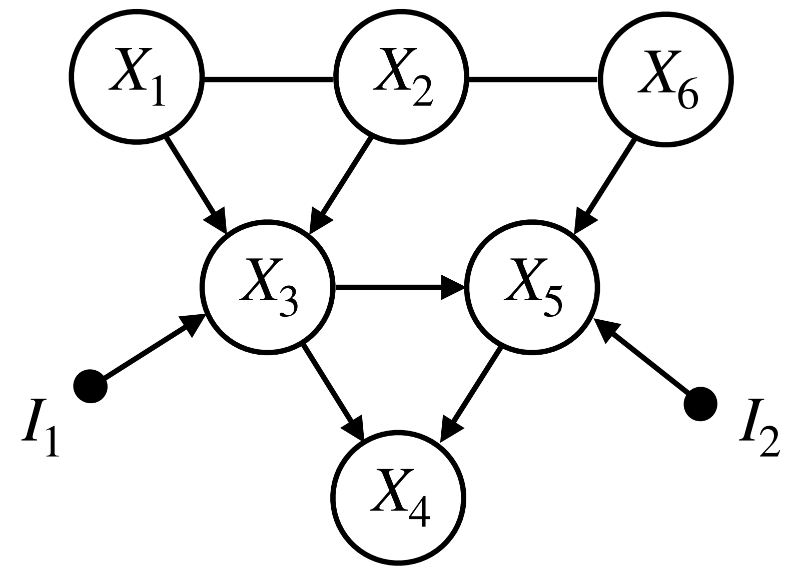


$\mathscr{I}$-CPDAG

Likelihood of a $\mathscr{I}$-CPDAG

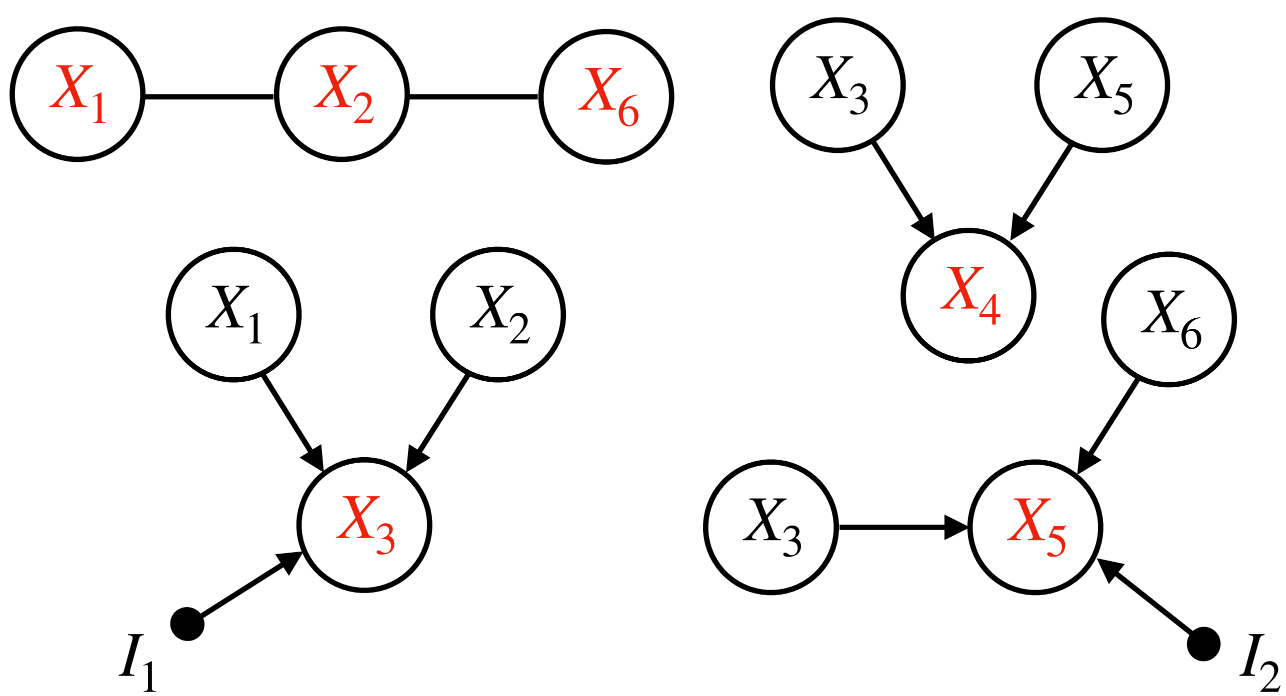


Initial $\mathscr{I}$-DAG



$\mathscr{I}$-CPDAG

Chain components



$$\mathbb{P}(\color{red}{X_1}, \color{red}{X_2}, \color{red}{X_6})$$

$$\mathbb{P}(\color{red}{X_3} \,|\, X_1, X_2, I_1)$$

$$\mathbb{P}(\color{red}{X_4} \,|\, X_3, X_5)$$

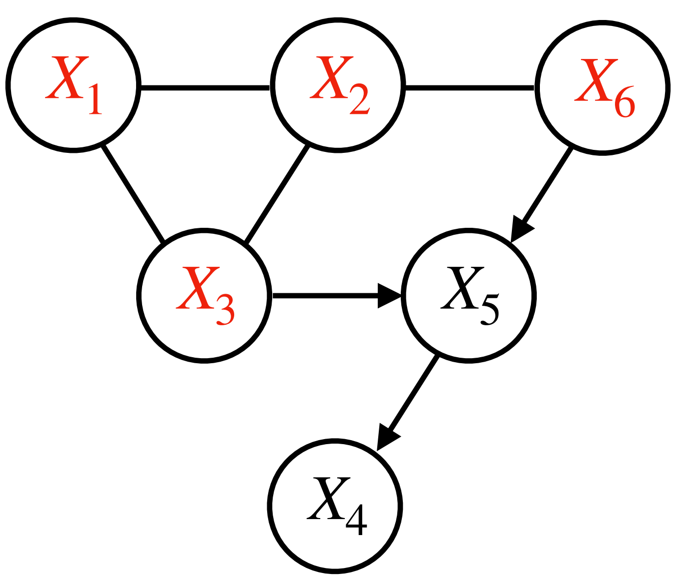
$$\mathbb{P}(\color{red}{X_5} \,|\, X_3, X_6, I_2)$$

Likelihood of a chain component

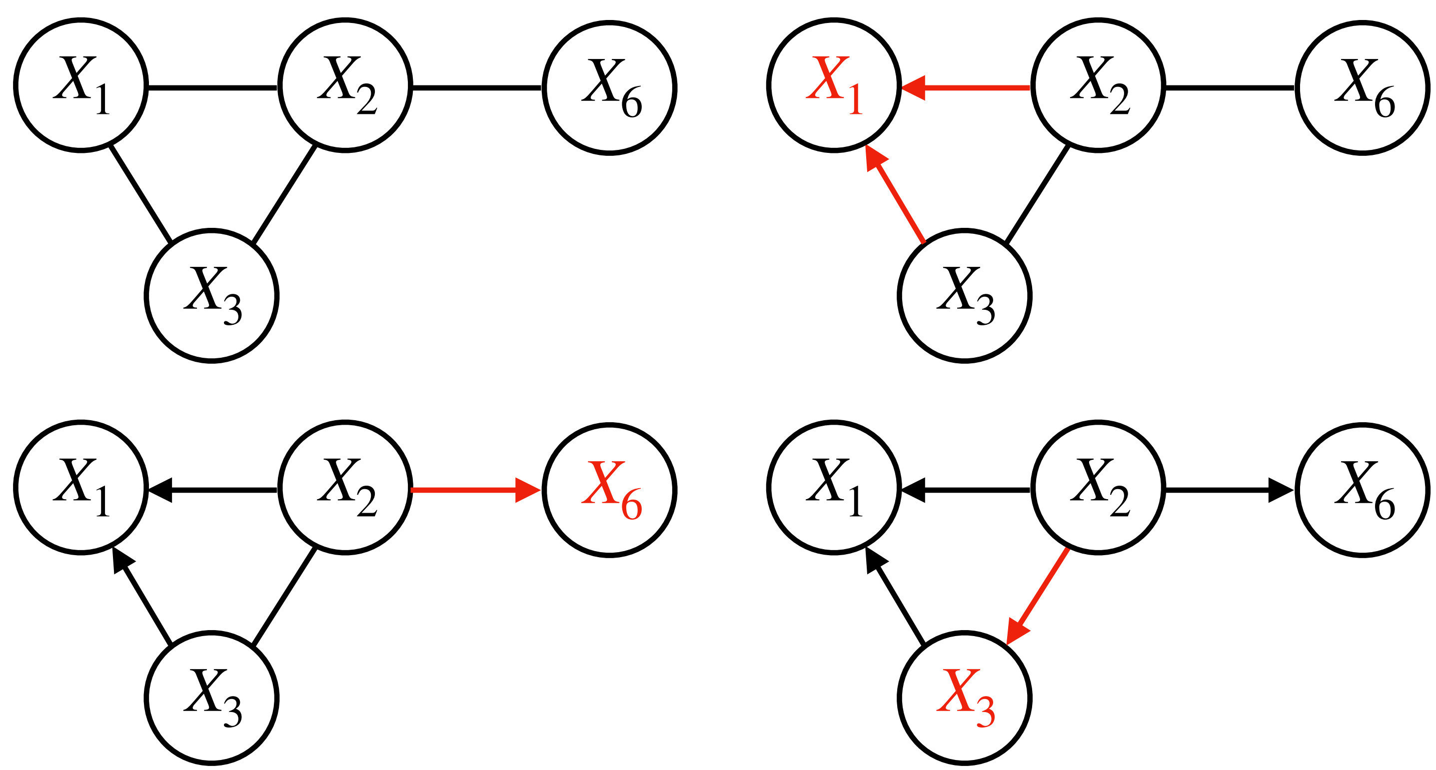
Theoretical results:

- If one intervention contains one element of the chain component it must contain all of them (Theorem 18, Hauser & Bühlmann, 2012)
- A chain component is necessary chordal, elimination order provide DAG representative (Appendix A.1, Hauser & Bühlmann, 2012)

Example: with (perfect) elimination order X_1, X_6, X_3, X_2



CPDAG
CC in Red

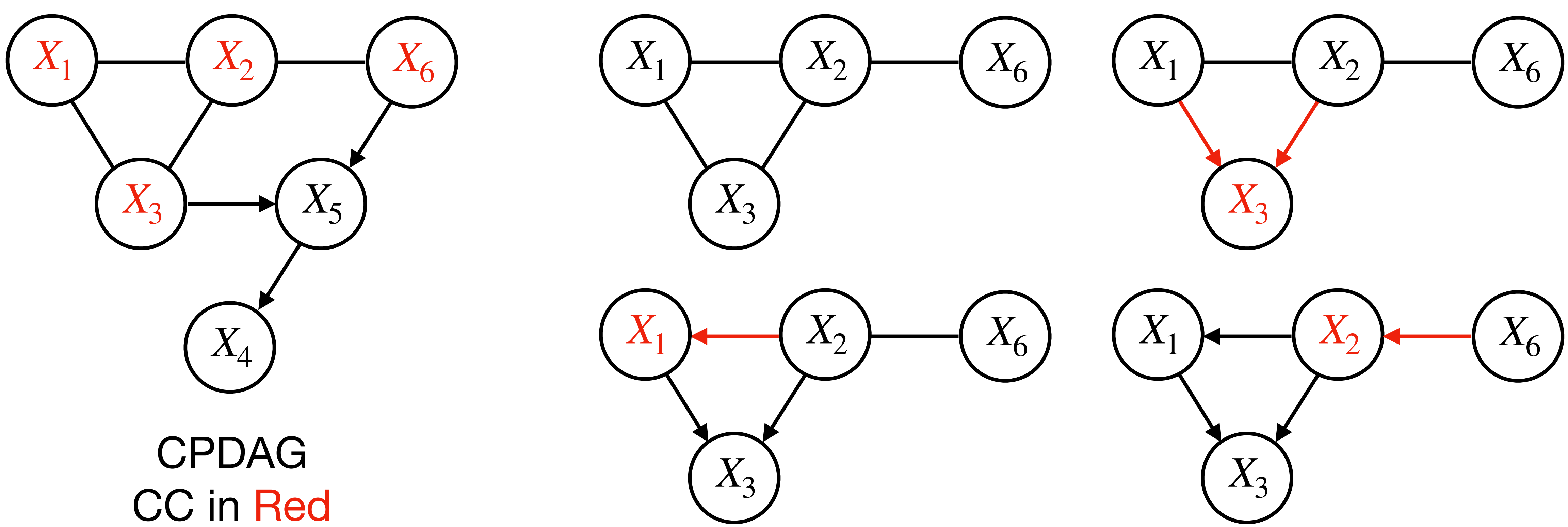


Likelihood of a chain component

Theoretical results:

- If one intervention contains one element of the chain component it must contain all of them (Theorem 18, Hauser & Bühlmann, 2012)
- A chain component is necessary chordal, elimination order provide DAG representative (Appendix A.1, Hauser & Bühlmann, 2012)

Example: with (perfect) elimination order X_3, X_1, X_2, X_6



Real life implementation

The problem: compute for all G

$$\mathbb{P}(G|\text{data}) \propto \exp\left(\log \mathbb{P}(G) + \text{loglik}(\hat{\theta}|G) - \text{pen}(G)\right)$$

A simple solution: by enumerating all DAGs.

p	number of DAGs
1	1
2	3
3	25
4	543
5	$29,281$
6	$3,781,503$
7	$1,138,779,265$
8	$783,702,329,343$

A better idea: through MCMC.

MCMC framework

The problem: sample DAG G from

$$\mathbb{P}(G|\text{data}) \propto \exp\left(\log \mathbb{P}(G) + \text{loglik}(\hat{\theta}|G) - \text{pen}(G)\right)$$

Metropolis-Hastings: perform iteratively

- propose $G' \sim q(\cdot|G)$
- accept G' with rate $\min(1, \alpha)$ with

$$\alpha = \frac{\mathbb{P}(G'|\text{data})}{\mathbb{P}(G|\text{data})} \times \frac{q(G|G')}{q(G'|G)}$$

MC output: a collection of N DAGs (default: $N = 5000$):

$$\underbrace{\text{DAG}_1, \text{DAG}_2, \ldots, \text{DAG}_B}_{\text{burn-in (default: } B = 1000)}, \underbrace{\text{DAG}_{B+1}, \text{DAG}_{B+2} \ldots, \text{DAG}_\text{N}}_{\text{exploitable DAGs}}$$

empirical posterior:

$$\mathbb{P}(G = g|\text{data}) = \frac{1}{N - B} \sum_{i=B+1}^{N} \mathbb{1}_{\text{DAG}_i = g}$$

Proposals & Implementation

DAG space: MC3 (Madigan & Raftery, 1995)

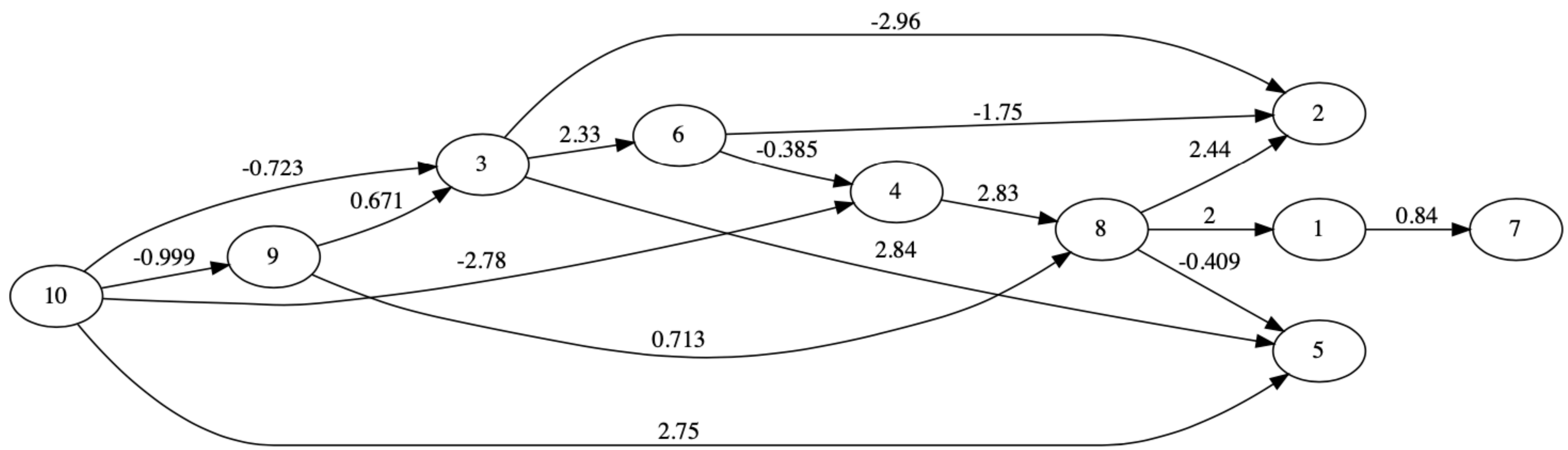
- Add/remove/flip arrow uniformly
- DAG constraint need smart update of route tables
- Available in `structmcmc` R package (Goudie, 2016)
- More constraints: max number of parents, fixed arrows

CPDAG space: He *et al* (2013), Castelleti *et al* (2018)

- Six moves: InsertU, DeleteU, InsertD, DeleteD, MakeV, RemoveV
- Plus one: ReverseD (Chickering 2002)
- Multiple theoretical conditions, asymmetric proposal
- `https://github.com/FedeCastelletti/obayes_learn_essential_graphs`

10 genes example

A random DAG with $p = 10$ genes



j	1	2	3	4	5	6	7	8	9	10
m	-0.61	-0.41	1.14	-1.84	1.00	0.71	-1.31	-0.96	0.06	0.70
σ	1.90	1.10	0.77	1.30	0.81	0.72	0.98	1.20	0.91	0.41

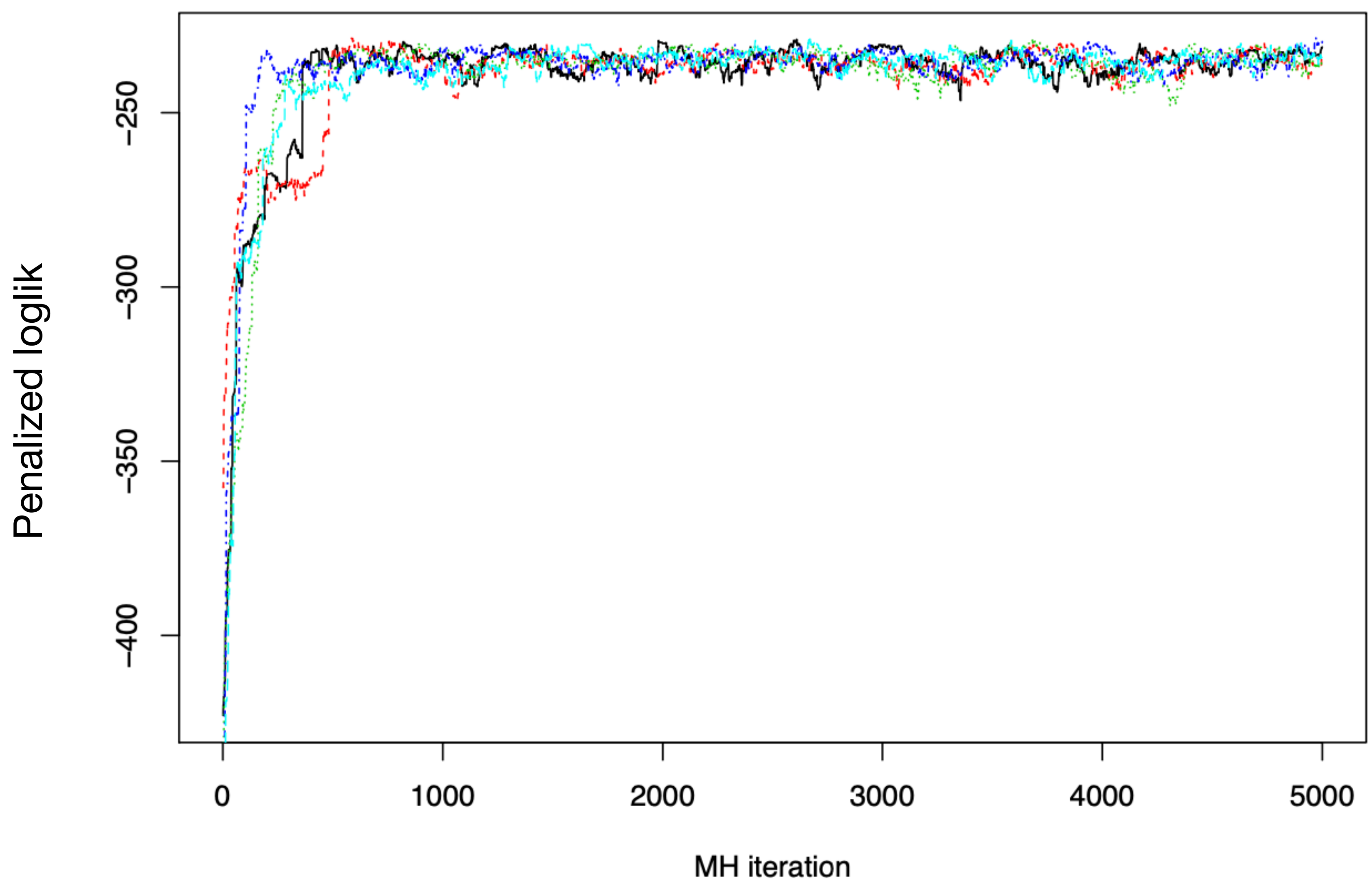
Some values (a causal ordering $10, 9, 3, 6, 4, 8, 2, 5, 1, 7$):

$$\mathrm{pa}(1) = \{8\} \quad \mathrm{pa}(4) = \{6, 10\} \quad \mathrm{pa}(10) = \emptyset$$

$$w_{6,2} = -1.75 \quad \ell_{6,2} = w_{6,2} + w_{6,4} \times w_{4,8} \times w_{8,2} = -4.41$$

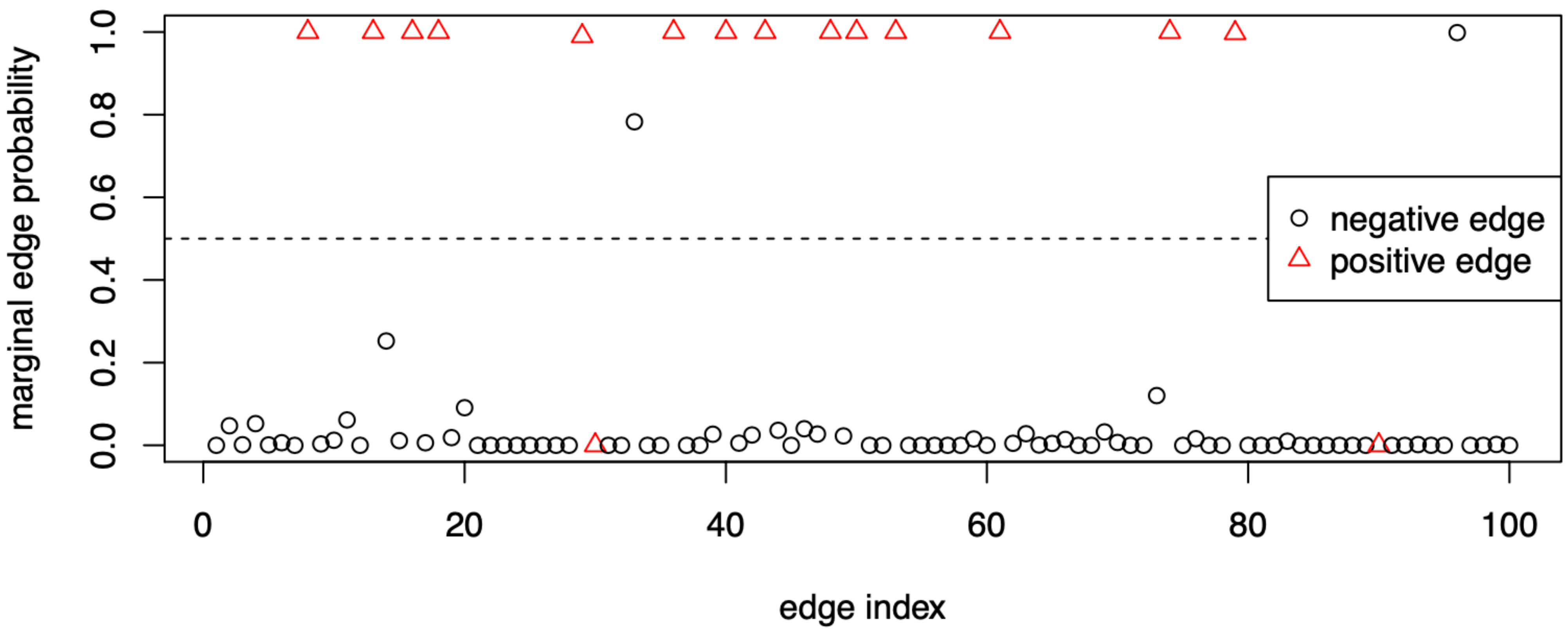
MCMC convergence

Design: fixed DAG and data (50 WT + 50 KO), 5000 MCMC iterations, unconstrained search, acceptance rate $\simeq 40\,\%$



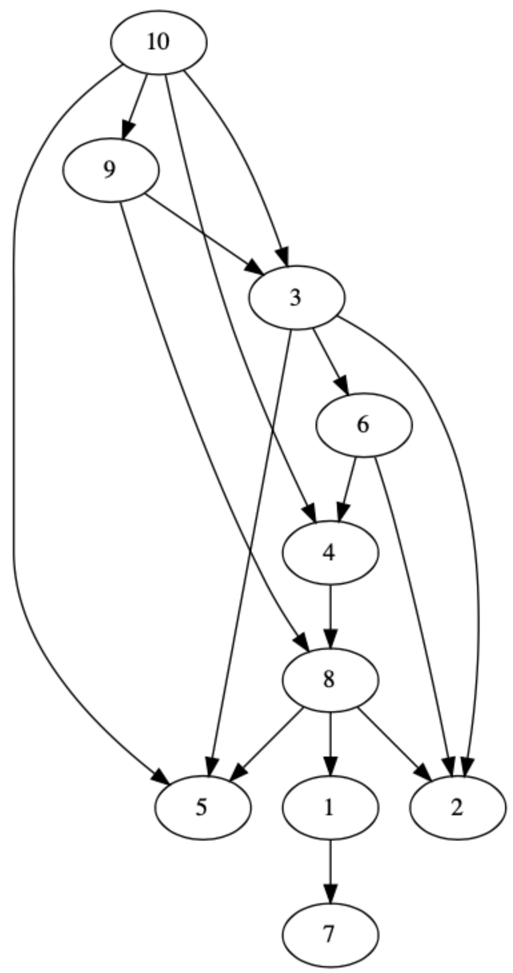
5 replications

Marginal edge probability

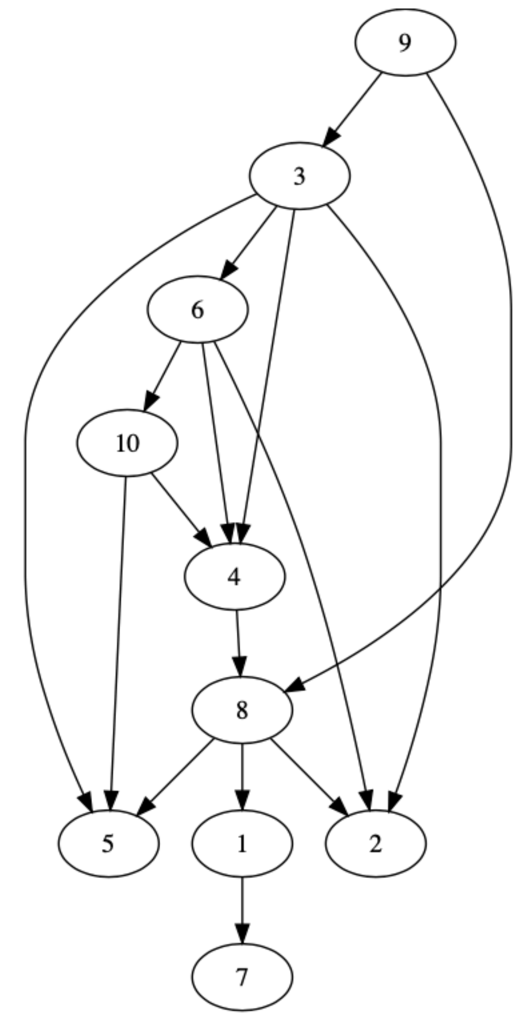


marginal posterior distribution on edges
threshold at 0.5 – AUROC=0.91 [0.78-1.00]

Consensus DAG



reference



consensus MC3

Direct effects

i	j	$w^*_{i,j}$	mean	sd
3	2	−2.96	−2.59	0.03
10	4	−2.78	−2.39	0.04
6	2	−1.75	−1.84	0.03
10	9	−1.00	0.00	0.00
10	3	−0.72	0.00	0.00
8	5	−0.41	−0.41	0.01
6	4	−0.39	−0.75	0.13
4	2	0.00	0.06	0.10
3	4	0.00	0.63	0.33
6	10	0.00	−0.06	0.00

i	j	$w^*_{i,j}$	mean	sd
9	3	0.67	0.82	0.08
9	8	0.71	0.62	0.07
1	7	0.84	0.84	0.00
8	1	2.00	2.03	0.02
3	6	2.33	2.32	0.01
8	2	2.44	2.42	0.03
10	5	2.75	2.93	0.02
4	8	2.83	2.67	0.01
3	5	2.84	2.82	0.03

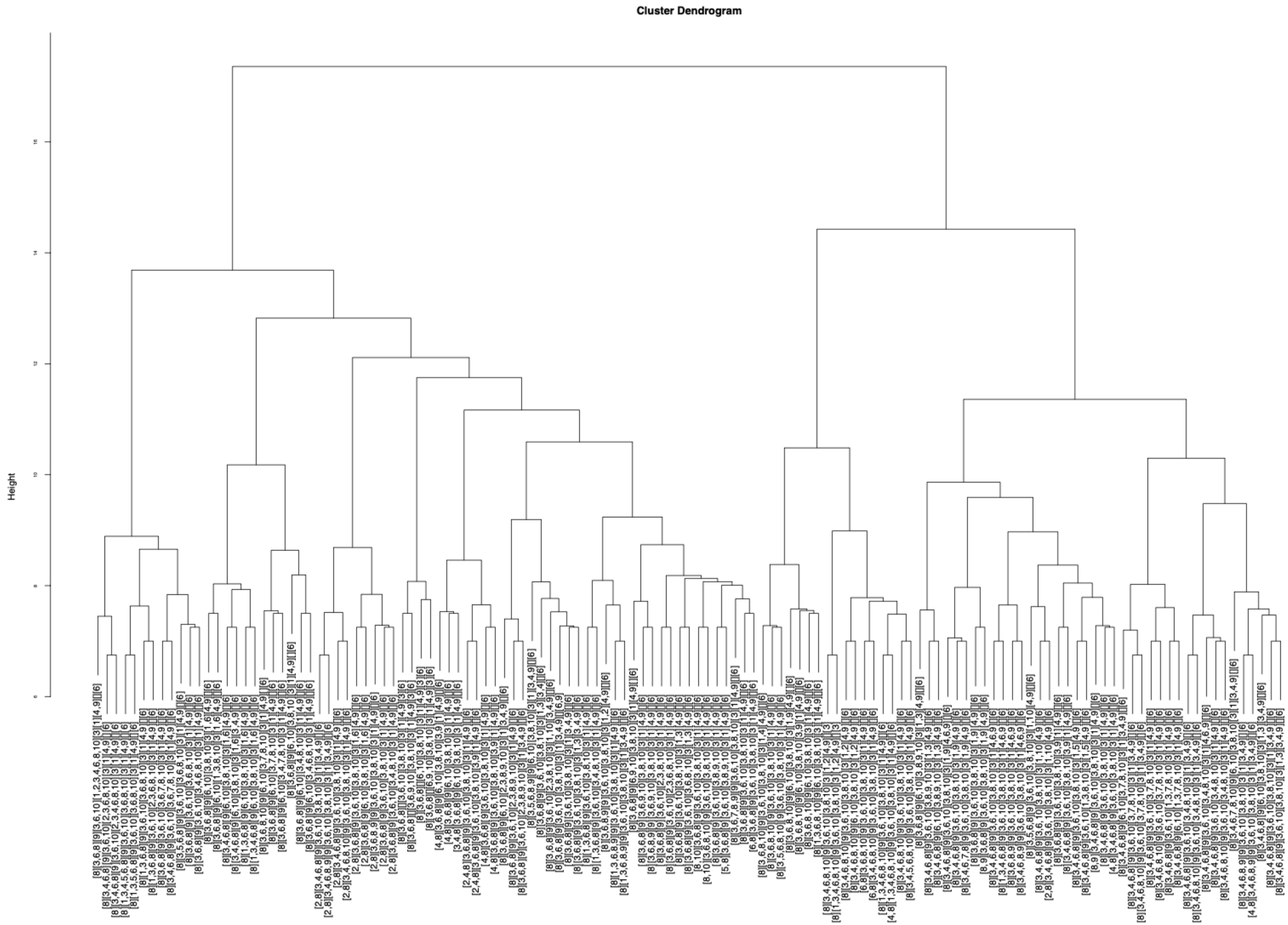
In/out degrees

i	$\mathrm{in}_i^{g^*}$	mean	sd
1	1	1.12	0.36
2	3	3.44	0.67
3	2	0.99	0.10
4	2	2.81	0.43
5	3	3.16	0.42
6	1	1.02	0.12
7	1	1.09	0.29
8	2	2.13	0.34
9	1	0.01	0.10
10	0	1.00	0.05

i	$\mathrm{out}_i^{g^*}$	mean	sd
1	1	1.07	0.25
2	0	0.08	0.27
3	3	3.94	0.61
4	1	1.34	0.52
5	0	0.02	0.13
6	2	3.07	0.27
7	0	0.03	0.18
8	3	3.00	0.00
9	2	2.11	0.35
10	4	2.11	0.32

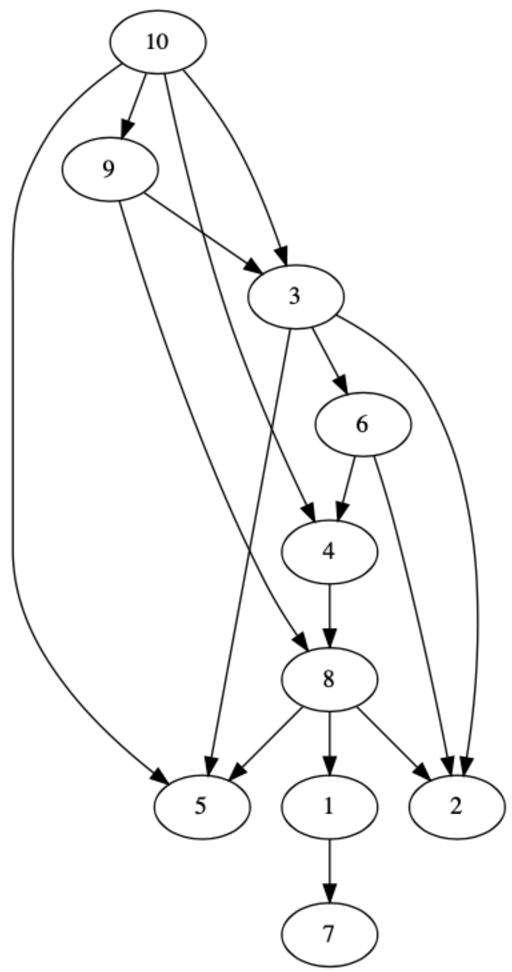
DAG clustering

Literature: Kendall tau and greedy centroids (Malmi, 2015).

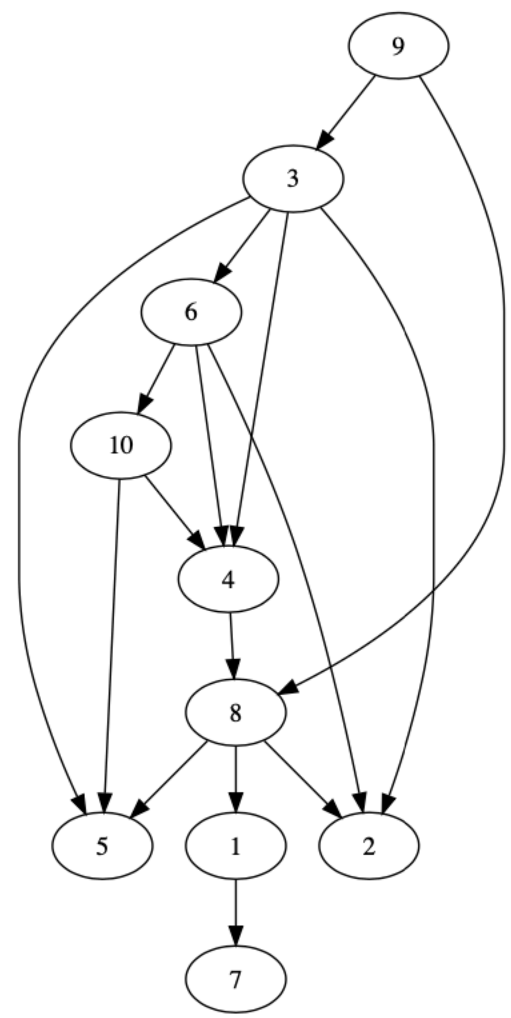


Centroid DAG

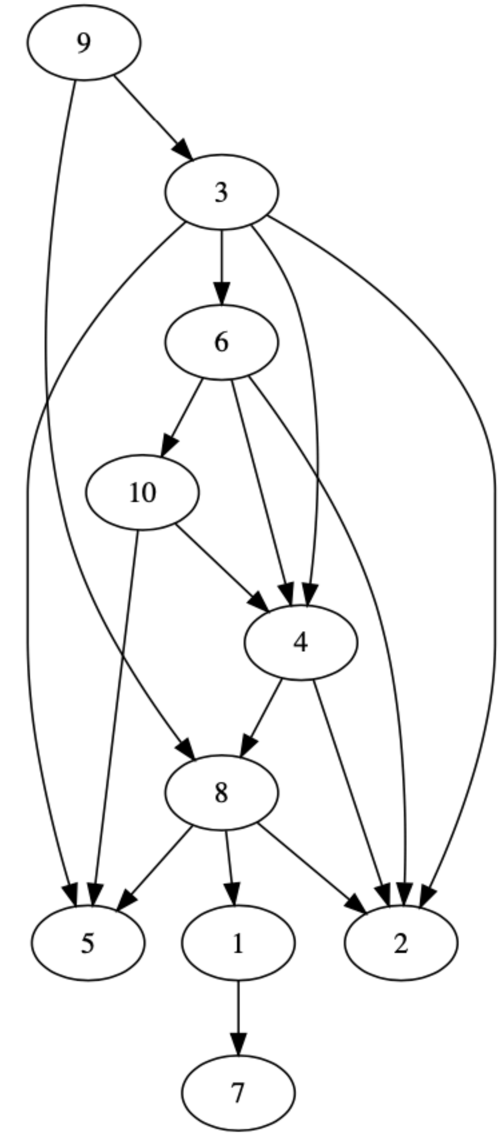
Literature: Kendall tau and greedy centroids (Malmi, 2015).



reference



consensus

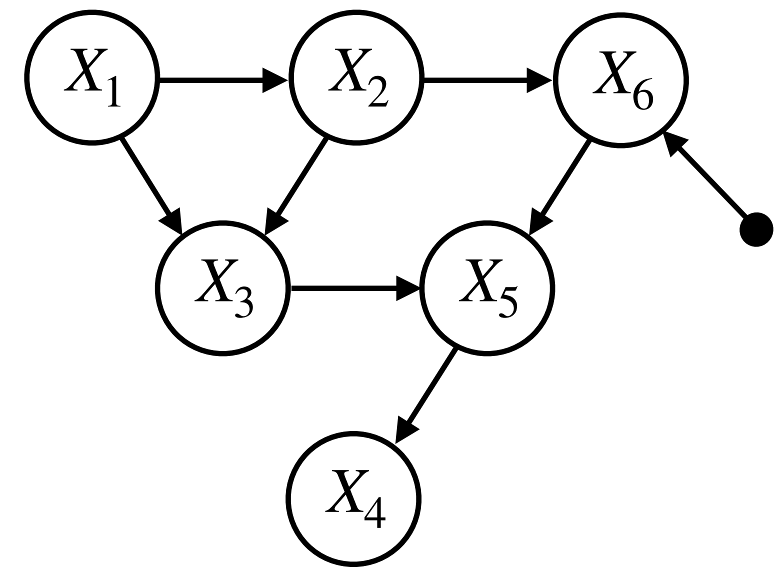


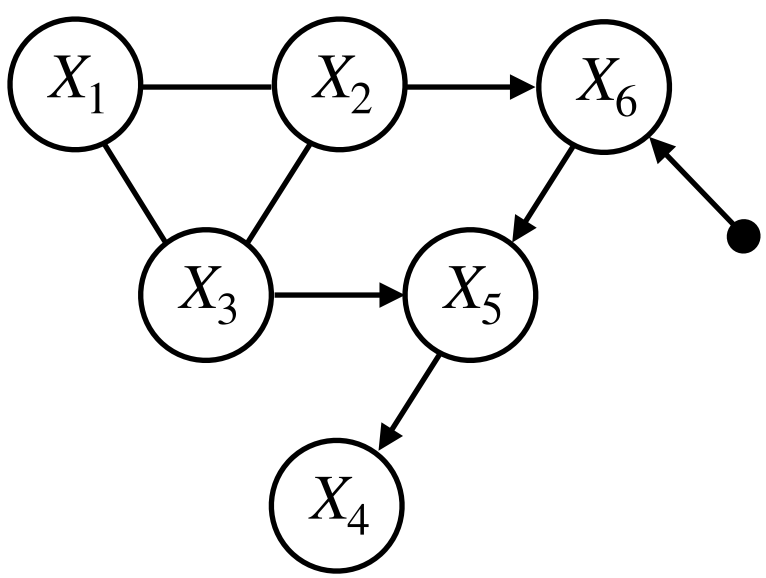
centroid

Conclusions & Perspectives

Take-home messages:

- Correlation is not causation
- CPDAGs = Markov equivalence class of DAGs
- Extension with interventions $\mathscr{I}$-CPDAGs
- Relatively « simple » with intervention nodes
- MCMC over DAG or CPDAG spaces





What Next ?

- MCMC over CPDAG not trivial
- What to do with a collection of DAGs or CPDAGs ?
- Clinical trials: mixing observations and interventions ?
- Gene regulation networks: best interventions ?

Few references

- Andersson, S. A., Madigan, D., & Perlman, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, *25*(2), 505-541.

- Castelletti, F., Consonni, G., Della Vedova, M. L., & Peluso, S. (2018). Learning Markov equivalence classes of directed acyclic graphs: an objective Bayes approach. *Bayesian Analysis*, *13*(4), 1235-1260.

- Hauser, A., & Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, *13*(1), 2409-2464.

- He, Y., Jia, J., & Yu, B. (2013). Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs. *The Annals of Statistics*, *41*(4), 1742-1779.

- Yang, K., Katcoff, A., & Uhler, C. (2018, July). Characterizing and learning equivalence classes of causal DAGs under interventions. In *International Conference on Machine Learning* (pp. 5541-5550). PMLR.