Machines Climbing Pearl's Ladder of Causation

Lecture II - Causal Discovery

Adèle H. Ribeiro

adele.ribeiro@uni-marburg.de
https://adele.github.io/
GitHub: @adele | Youtube: @adelehelena | X: @adelehr

Faculty of Mathematics and Computer Science, Philipps-University of Marburg
Institute of Medical Informatics, University of Münster

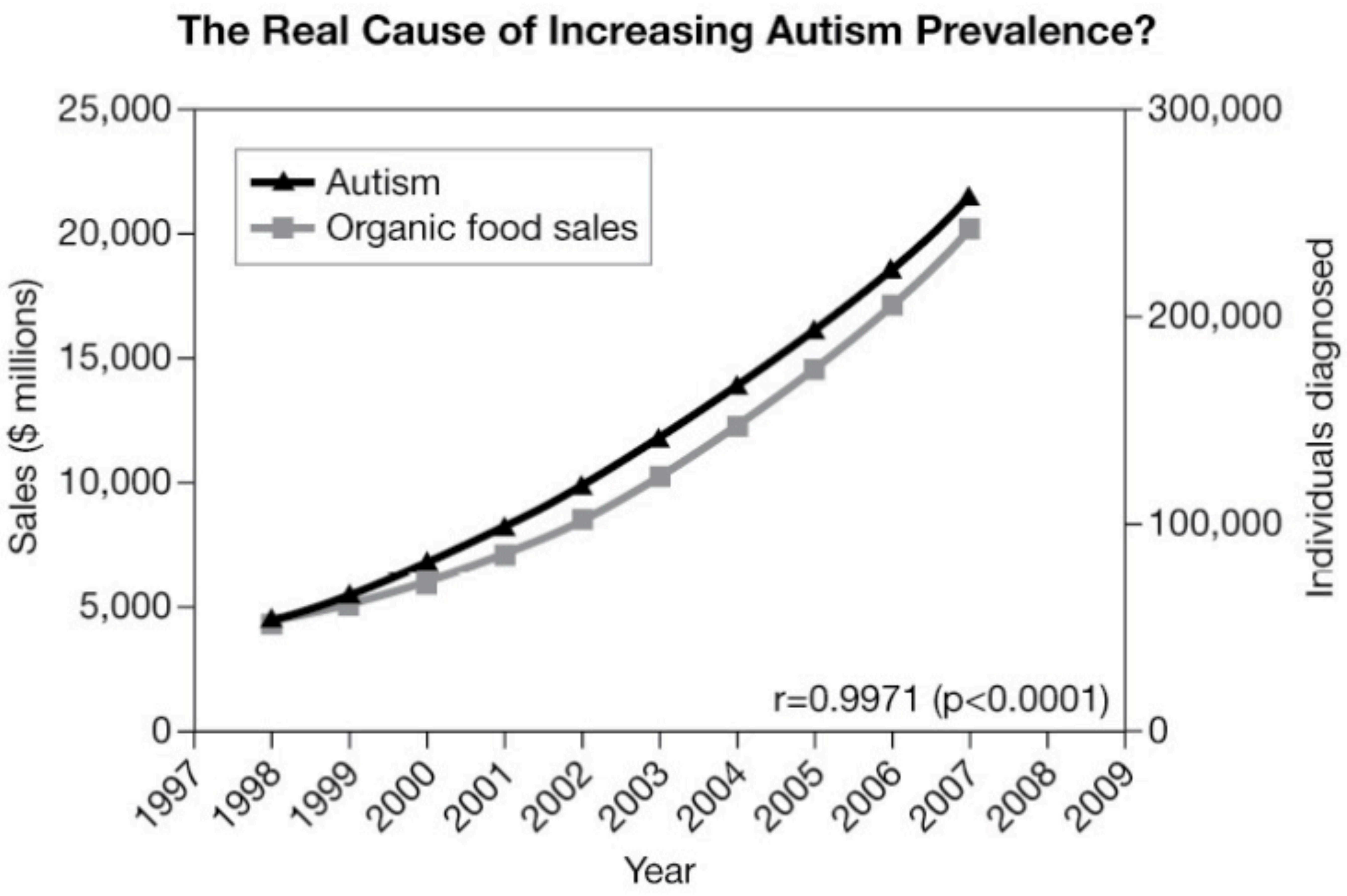
ESSAI & ACAI 2024
ATHENS - GREECE

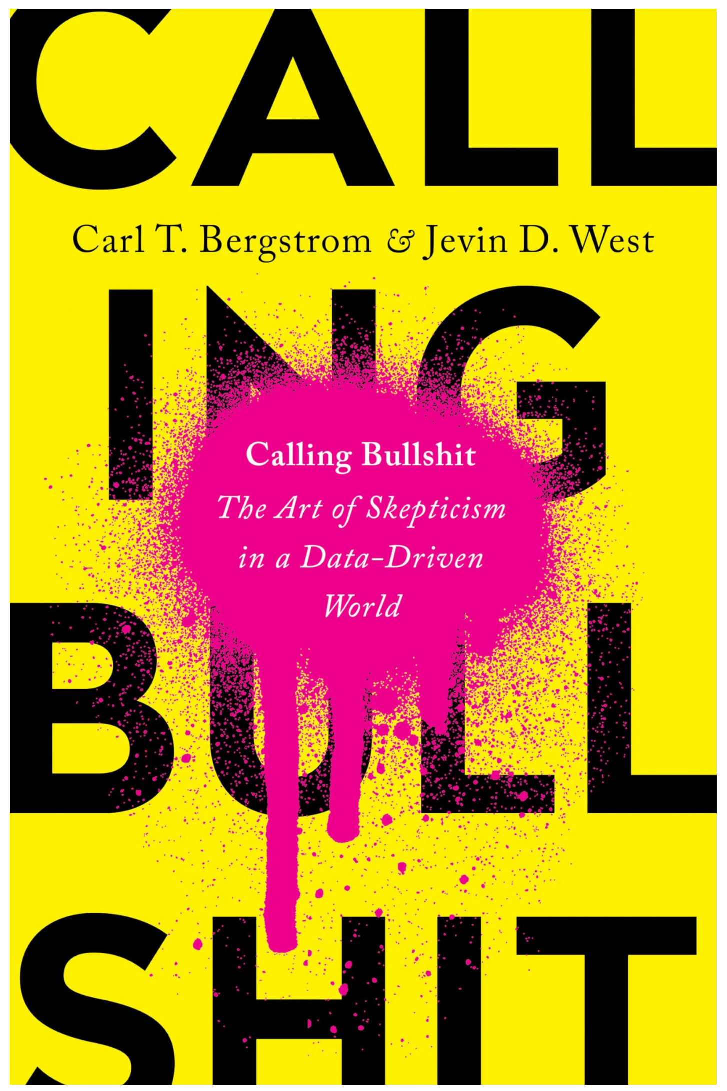
2nd European Summer School on Artificial Intelligence (ESSAI)
July 22-26, 2024

Outline

- Bayesian Networks — Encoders of Conditional Independencies
- Markov Equivalence Class
- D-Separation
- Causal Discovery — Score-Based & Constraint-Based Algorithms
  - Fast Causal Inference (FCI) Algorithm
- Advances in Causal Discovery under Latent Confounding
  - From Observational & Interventional Data / Multiple Environments
  - Integration of Background Knowledge
  - Probabilistic Approach for Modeling Uncertainty
  - Parametric Approaches - Linear + Non-Gaussian / Additive Noise Models
  - Dynamic Systems: Cycles and Time-Series Data
- Current Challenges and Open Problems

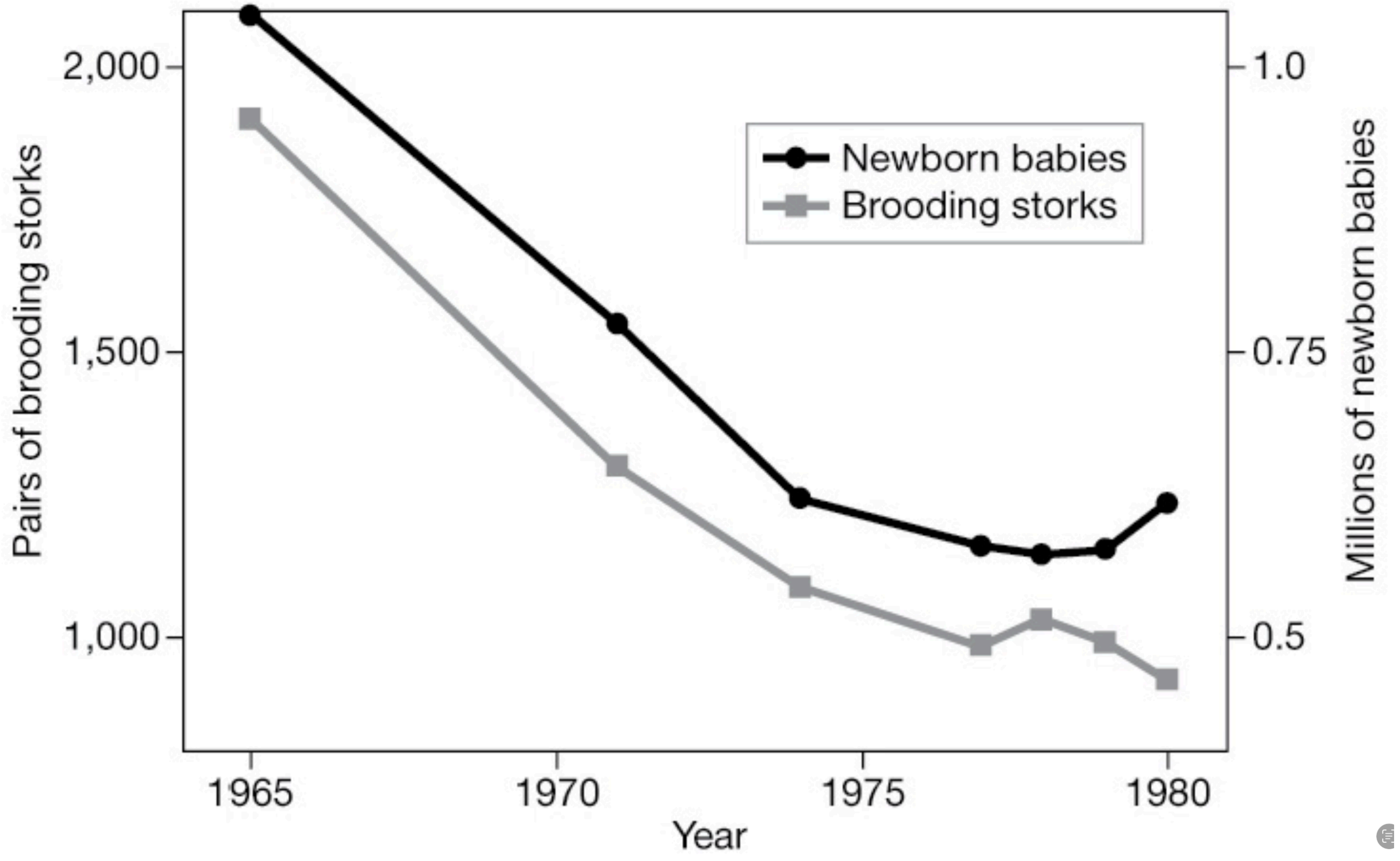
What Do Statistical Associations Reveal?



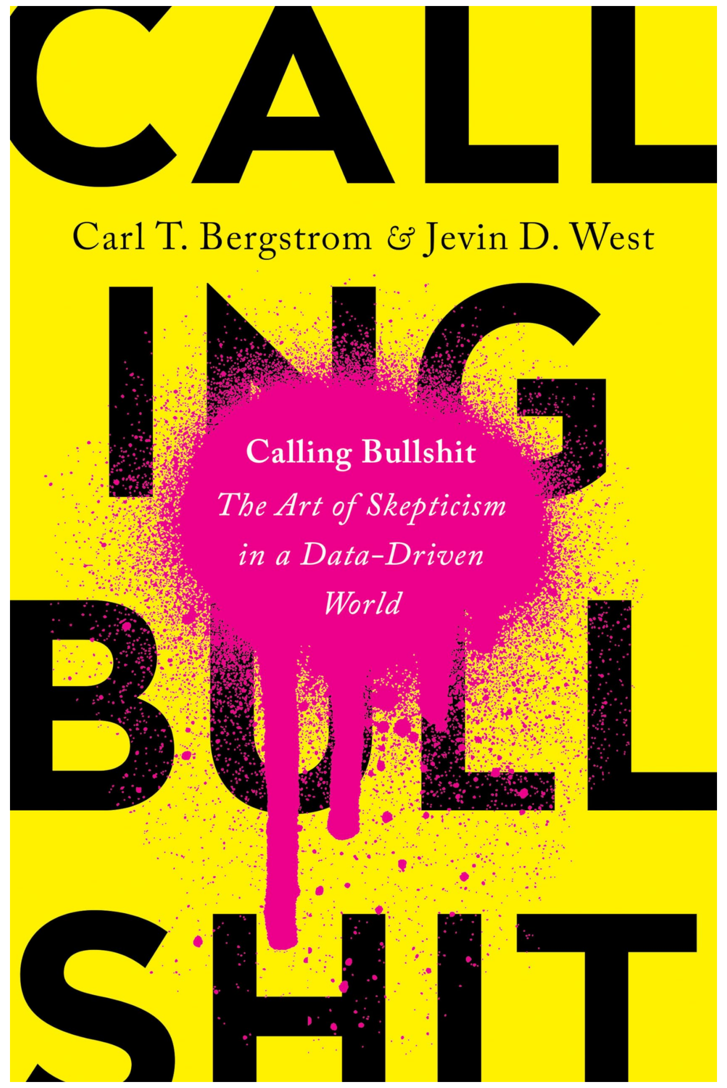


“one user of the Reddit website posted the following graph”

What Do Statistical Associations Reveal?



"Pairs of brooding storks in West Germany and the number of newborn human babies."



The graph, titled "A New Parameter for Sex Education," appeared in a humorous publication in Nature.

"Perhaps the old tall tale is right: Perhaps storks do bring babies after all."

Correlation does not imply causation!

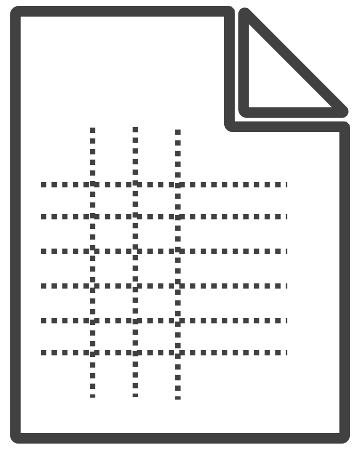
$$\mathcal{M}_1 = \begin{cases} \mathbf{V} = \{X, Y\} \\ \mathbf{U} = \{U_x, U_Y\} \\ \mathscr{F} = \begin{cases} f_X(U_X) \\ f_Y(X, U_Y) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

$\vdots$

$$\mathcal{M}_{N-1} = \begin{cases} \mathbf{V} = \{X, Y\} \\ \mathbf{U} = \{U_x, U_Y, U_{X,Y}\} \\ \mathscr{F} = \begin{cases} f_X(Y, U_X, U_{X,Y}) \\ f_Y(U_Y, U_{X,Y}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

$$\mathcal{M}_N = \begin{cases} \mathbf{V} = \{X, Y\} \\ \mathbf{U} = \{U_x, U_Y\} \\ \mathscr{F} = \begin{cases} f_X(U_X) \\ f_Y(U_Y) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Data



Conditional (in)dependencies

$P(\mathbf{v})$

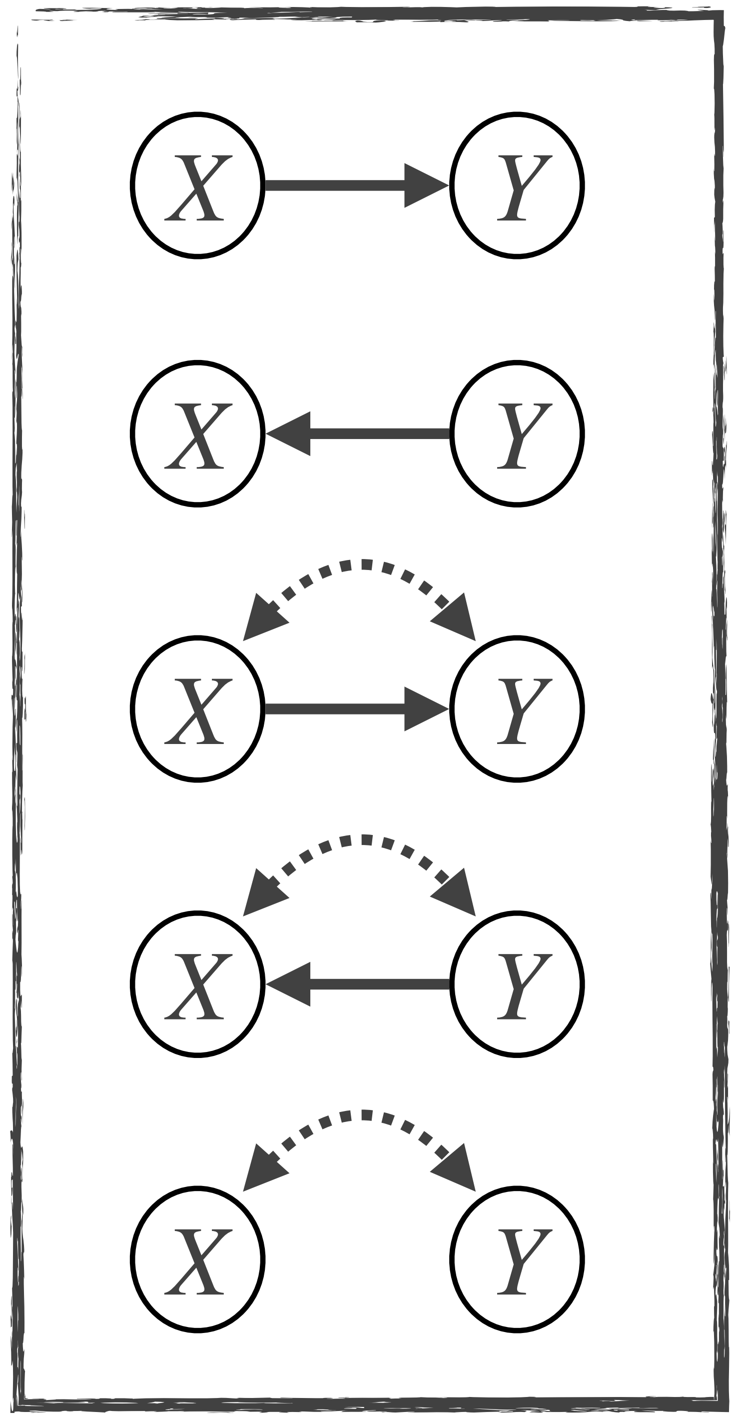
$X \not\perp\!\!\!\perp Y$

$$P(x, y) = \sum_{u_x, u_y} P(x \mid y) P(y) P(u_x, u_y)$$

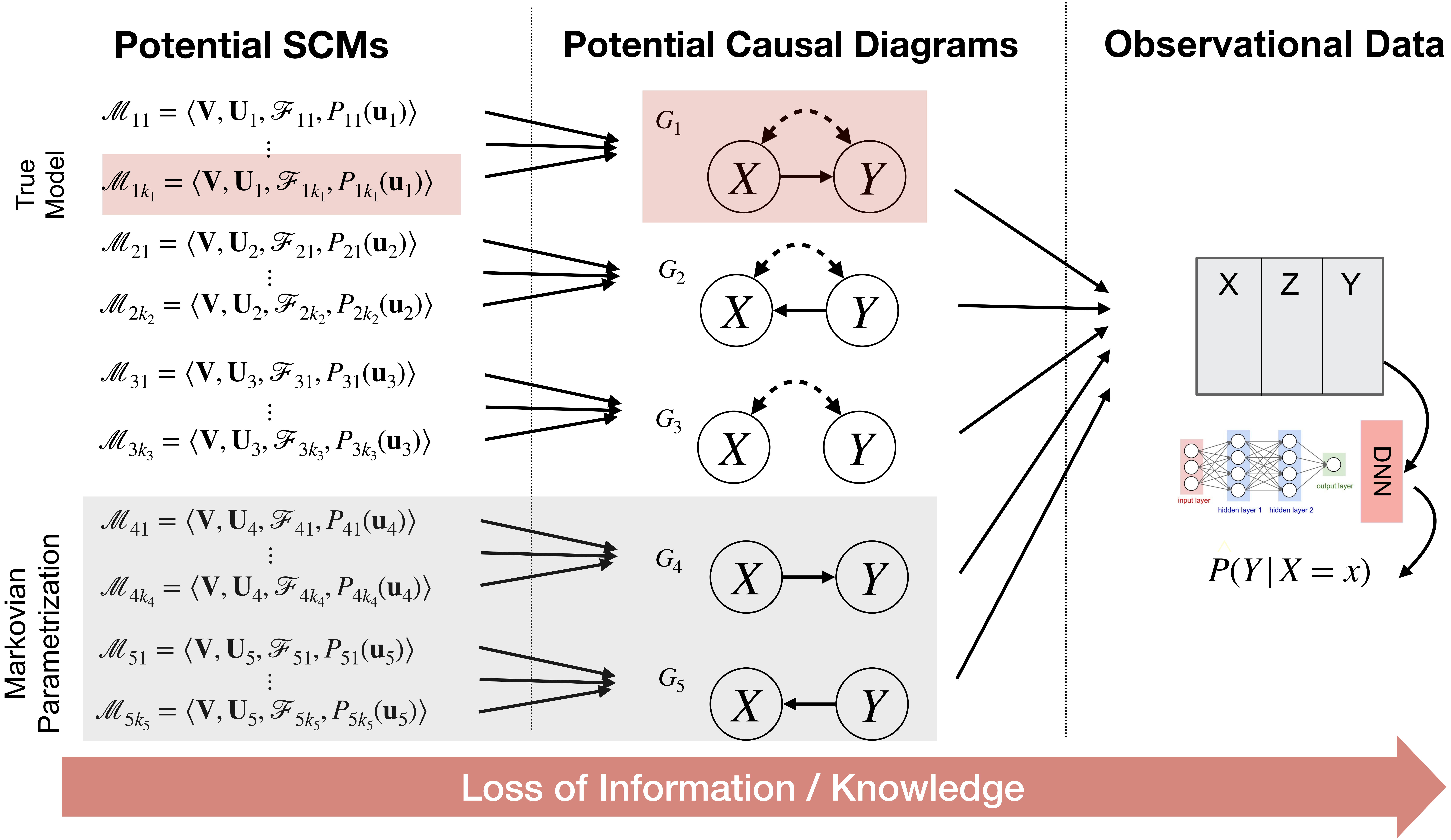
$$P(x, y) = \sum_{u_x, u_y} P(y \mid x) P(x) P(u_x, u_y)$$

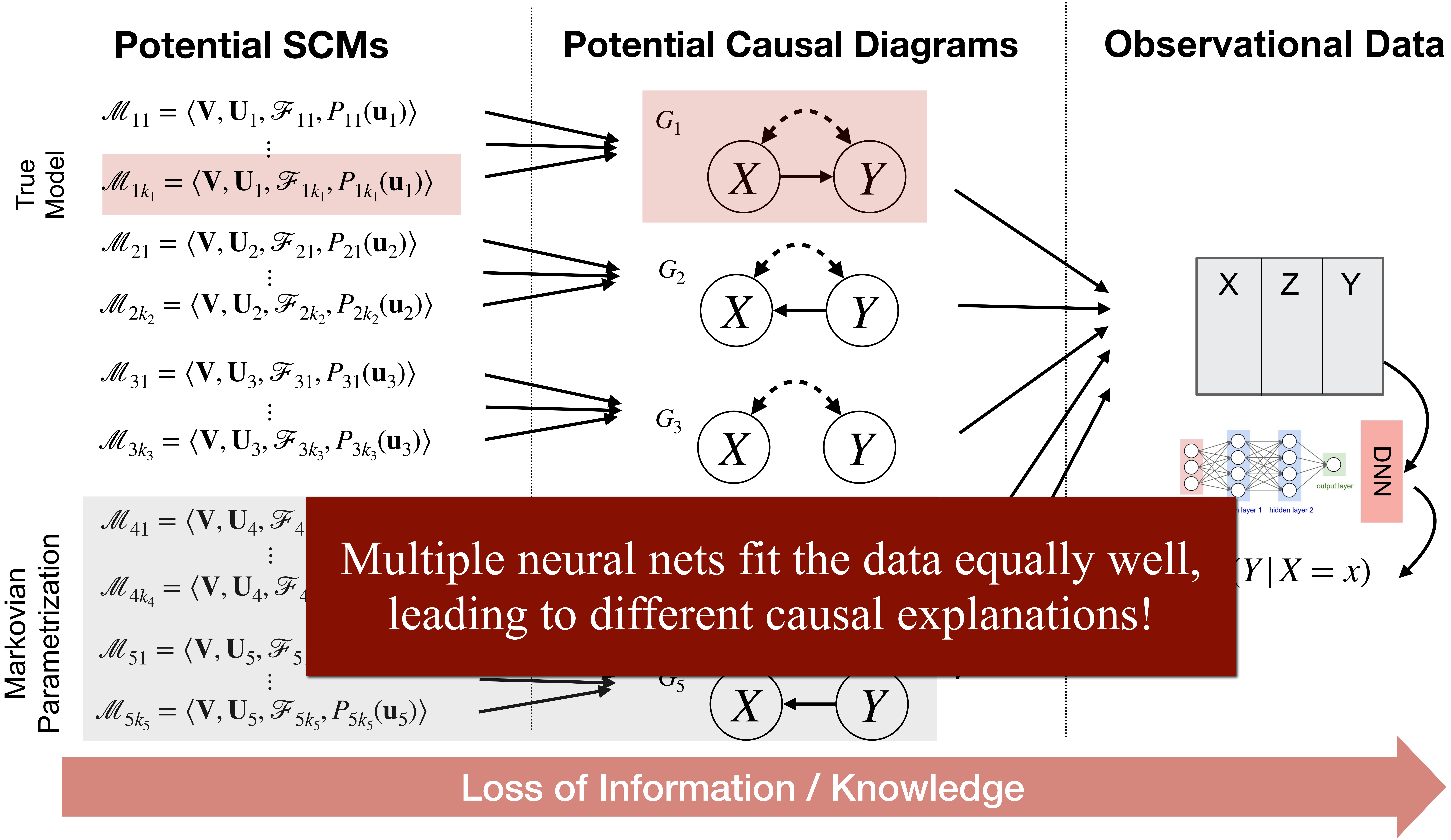
Markov Equivalence Class

(class of models implying the same set of conditional independencies)



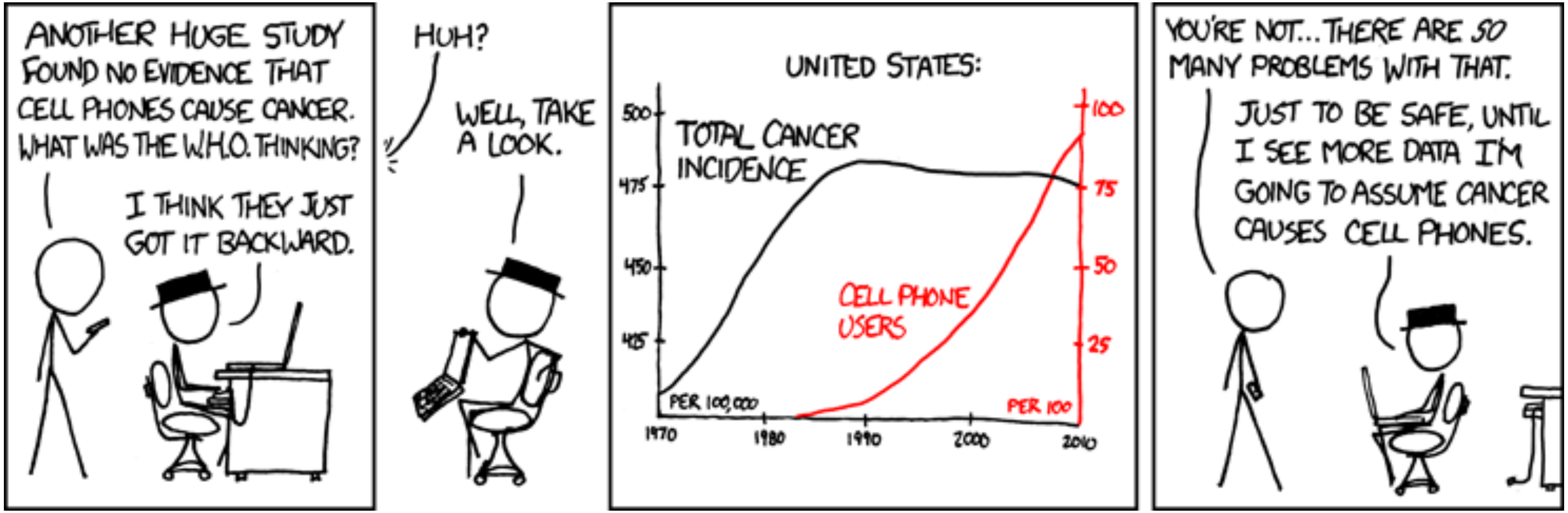
Correlation does not imply causation!


Potential SCMs
Potential Causal Diagrams
Observational Data
True Model
Markovian Parametrization
$\mathcal{M}_{11} = \langle \mathbf{V}, \mathbf{U}_1, \mathcal{F}_{11}, P_{11}(\mathbf{u}_1) \rangle$
⋮
$\mathcal{M}_{1k_1} = \langle \mathbf{V}, \mathbf{U}_1, \mathcal{F}_{1k_1}, P_{1k_1}(\mathbf{u}_1) \rangle$
$\mathcal{M}_{21} = \langle \mathbf{V}, \mathbf{U}_2, \mathcal{F}_{21}, P_{21}(\mathbf{u}_2) \rangle$
⋮
$\mathcal{M}_{2k_2} = \langle \mathbf{V}, \mathbf{U}_2, \mathcal{F}_{2k_2}, P_{2k_2}(\mathbf{u}_2) \rangle$
$\mathcal{M}_{31} = \langle \mathbf{V}, \mathbf{U}_3, \mathcal{F}_{31}, P_{31}(\mathbf{u}_3) \rangle$
⋮
$\mathcal{M}_{3k_3} = \langle \mathbf{V}, \mathbf{U}_3, \mathcal{F}_{3k_3}, P_{3k_3}(\mathbf{u}_3) \rangle$
$\mathcal{M}_{41} = \langle \mathbf{V}, \mathbf{U}_4, \mathcal{F}_{41}, P_{41}(\mathbf{u}_4) \rangle$
⋮
$\mathcal{M}_{4k_4} = \langle \mathbf{V}, \mathbf{U}_4, \mathcal{F}_{4k_4}, P_{4k_4}(\mathbf{u}_4) \rangle$
$\mathcal{M}_{51} = \langle \mathbf{V}, \mathbf{U}_5, \mathcal{F}_{51}, P_{51}(\mathbf{u}_5) \rangle$
⋮
$\mathcal{M}_{5k_5} = \langle \mathbf{V}, \mathbf{U}_5, \mathcal{F}_{5k_5}, P_{5k_5}(\mathbf{u}_5) \rangle$
G_1
X
Y
G_2
X
Y
G_3
X
Y
G_4
X
Y
G_5
X
Y
X
Z
Y
input layer
hidden layer 1
hidden layer 2
output layer
DNN
$P(Y \mid X = x)$
Loss of Information / Knowledge

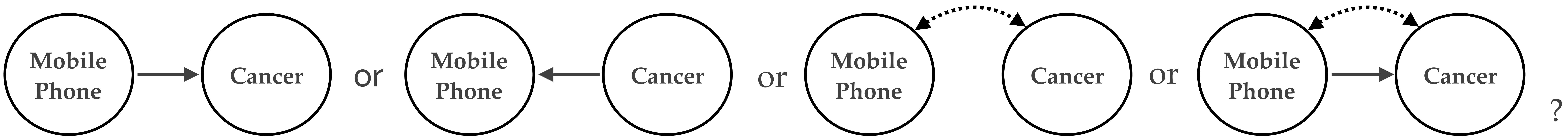


Potential SCMs
Potential Causal Diagrams
Observational Data
True Model
$\mathcal{M}_{11} = \langle \mathbf{V}, \mathbf{U}_1, \mathcal{F}_{11}, P_{11}(\mathbf{u}_1) \rangle$
$\mathcal{M}_{1k_1} = \langle \mathbf{V}, \mathbf{U}_1, \mathcal{F}_{1k_1}, P_{1k_1}(\mathbf{u}_1) \rangle$
$\mathcal{M}_{21} = \langle \mathbf{V}, \mathbf{U}_2, \mathcal{F}_{21}, P_{21}(\mathbf{u}_2) \rangle$
$\mathcal{M}_{2k_2} = \langle \mathbf{V}, \mathbf{U}_2, \mathcal{F}_{2k_2}, P_{2k_2}(\mathbf{u}_2) \rangle$
$\mathcal{M}_{31} = \langle \mathbf{V}, \mathbf{U}_3, \mathcal{F}_{31}, P_{31}(\mathbf{u}_3) \rangle$
$\mathcal{M}_{3k_3} = \langle \mathbf{V}, \mathbf{U}_3, \mathcal{F}_{3k_3}, P_{3k_3}(\mathbf{u}_3) \rangle$
Markovian Parametrization
$\mathcal{M}_{41} = \langle \mathbf{V}, \mathbf{U}_4,$
$\mathcal{M}_{4k_4} = \langle \mathbf{V}, \mathbf{U}_4,$
$\mathcal{M}_{51} = \langle \mathbf{V}, \mathbf{U}_5,$
$\mathcal{M}_{5k_5} = \langle \mathbf{V}, \mathbf{U}_5, \mathcal{F}_{5k_5}, P_{5k_5}(\mathbf{u}_5) \rangle$
G_1
G_2
G_3
X
Y
X
Z
Y
DNN
output layer
hidden layer 2
$Y | X = x)$
Multiple neural nets fit the data equally well,
leading to different causal explanations!
Loss of Information / Knowledge

Association vs Causation



Will we be able to decide the true relationship just by **seeing** more data?



Which **type of data** would helps us to derive more definite conclusions?

How is then possible to learn causal relations solely from observational data?

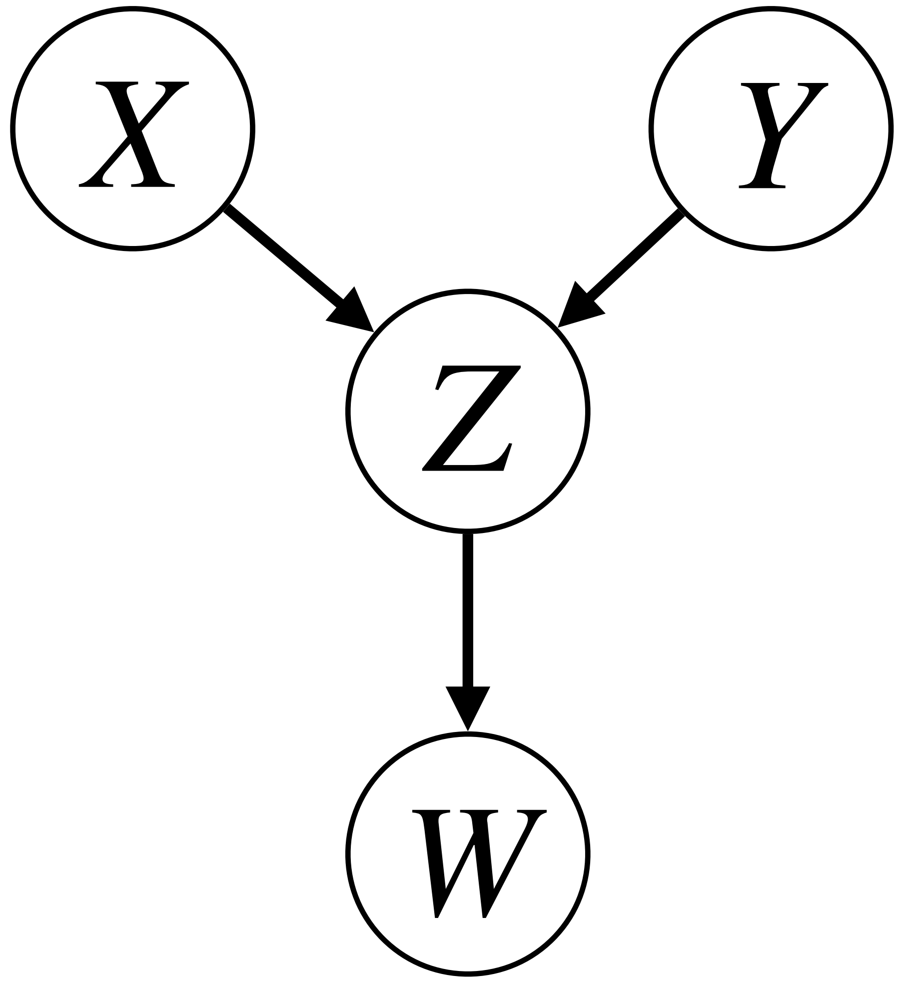
Bayesian Network

A DAG, possibly with latent confounders (ADMG), representing the **conditional independences** implied by an SCM

Directed Acyclic Graph

Acyclic Directed Mixed Graph

Graphical Kinship Notation



Directed Acyclic Graph (DAG)

X and Y are parents of Z, i.e., $X, Y \in Pa(Z)$

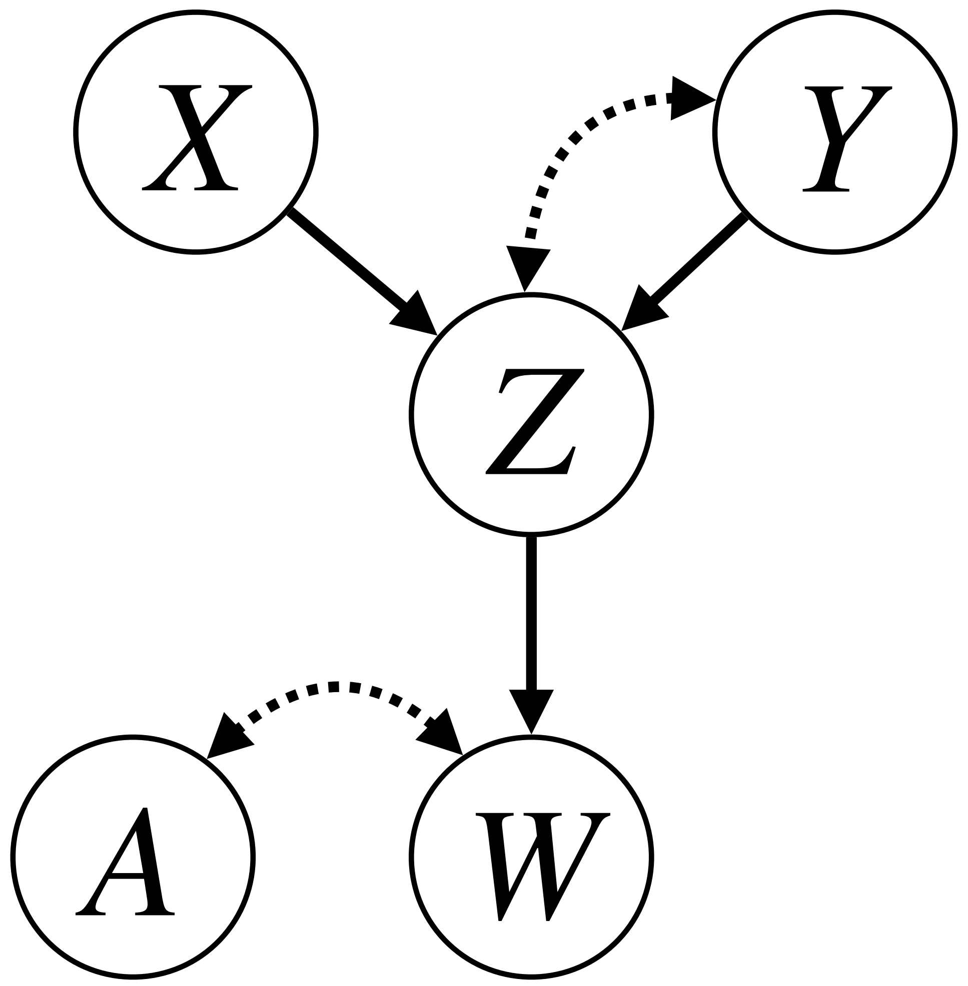
Z is a child of Y, i.e., $Z \in Ch(Y)$

W is a descendent of X, i.e., $W \in De(X)$

Y is ancestor of W, i.e., $Y \in An(W)$

Y is non-descendant of X, i.e., $Y \in NDesc(X)$

Graphical Kinship Notation



Acyclic Directed Mixed Graph (ADMG)

X and Y are parents of Z, i.e., $X, Y \in Pa(Z)$

Z is a child of Y, i.e., $Z \in Ch(Y)$

W is a descendent of X, i.e., $W \in De(X)$

Y is ancestor of W, i.e., $Y \in An(W)$

Y is non-descendant of X, i.e., $Y \in NDesc(X)$

A is spouse of W

Bayesian Networks & Markov Condition

A DAG G over $\mathbf{V}$ is a *Bayesian Network* for a joint probability distribution $P(\mathbf{V})$ if, for every $V_i \in \mathbf{V}$, it holds that $V_i \perp\!\!\!\perp NDesc_i \mid Pa_i$ and, therefore, $P(\mathbf{v})$ factorizes as follows:

P is satisfies the ***Markov Condition*** *w.r.t.* G

$$P(\mathbf{v}) = \prod_{V_i \in \mathbf{V}} P(v_i \mid v_{i-1}, \ldots, v_1)$$

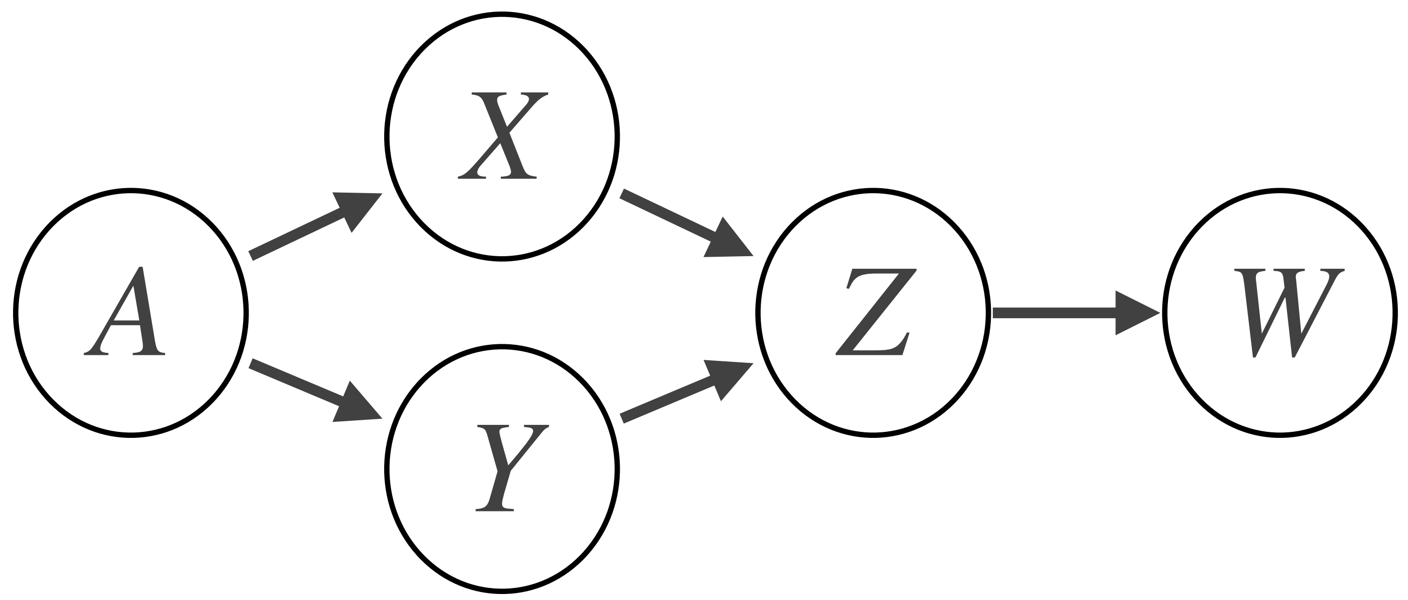
$$= \prod_{V_i \in \mathbf{V}} P(v_i \mid pa_i)$$

Chain Rule:

It holds for any topological order of G

$V_i \perp\!\!\!\perp NDesc_i \mid Pa_i, U_i$

Edges have no causal semantics!



$$P(\mathbf{v}) = P(w \mid z, x, y, a)\ P(z \mid x, y, a)\ P(x \mid y, a)\ P(y \mid a)\ P(a)$$
$$= P(w \mid z)\ P(z \mid x, y)\ P(x \mid a)\ P(y \mid a)\ P(a)$$

$W \perp\!\!\!\perp X, Y, A \mid Z$ $\quad$ $A \perp\!\!\!\perp Z \mid X, Y$ $\quad$ $Y \perp\!\!\!\perp X \mid A$

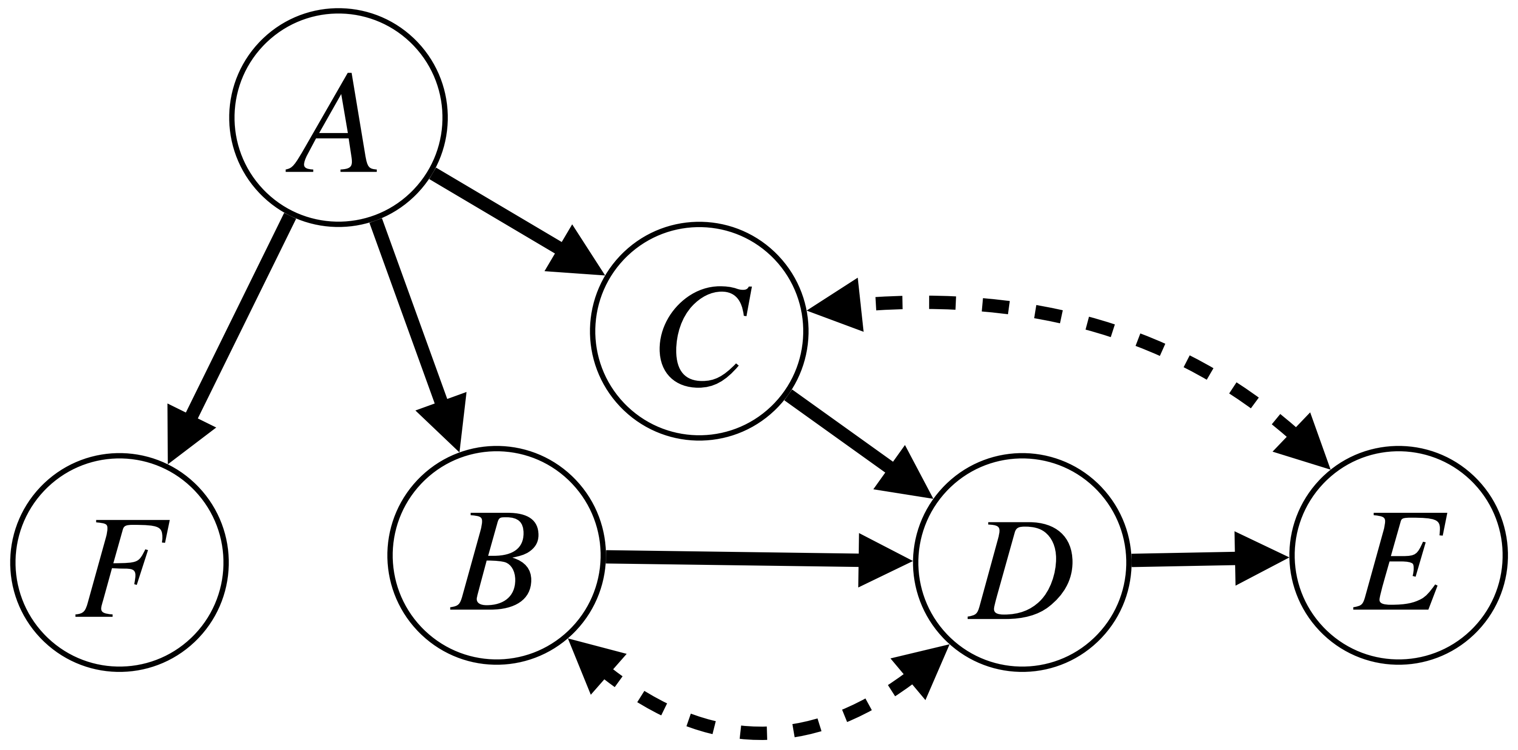
Bayesian Networks & Semi-Markov Condition

An ADMG G over $\mathbf{V}$ is a *Bayesian Network* for a joint probability distribution $P(\mathbf{V})$ if, for every $V_i \in \mathbf{V}$, it holds that $V_i \perp\!\!\!\perp NDesc_i \mid Pa_i^+$ and, therefore, $P(\mathbf{v})$ factorizes as follows:

P is satisfies the **Semi-*Markov* Condition** *w.r.t.* G

$$P(\mathbf{v}) = \prod_{V_i \in \mathbf{V}} P(v_i \mid pa_i^+).$$

The extended parents of V_i is defined as $Pa_i^+ = Pa^1(\{V \in \mathbf{C}(V_i) : V \leq V_i\}) \setminus \{V_i\}$, where $Pa^1(V) = Pa(V) \cup V$ and $\mathbf{C}(V_i)$ is a maximal path entirely made of bidirected edges.



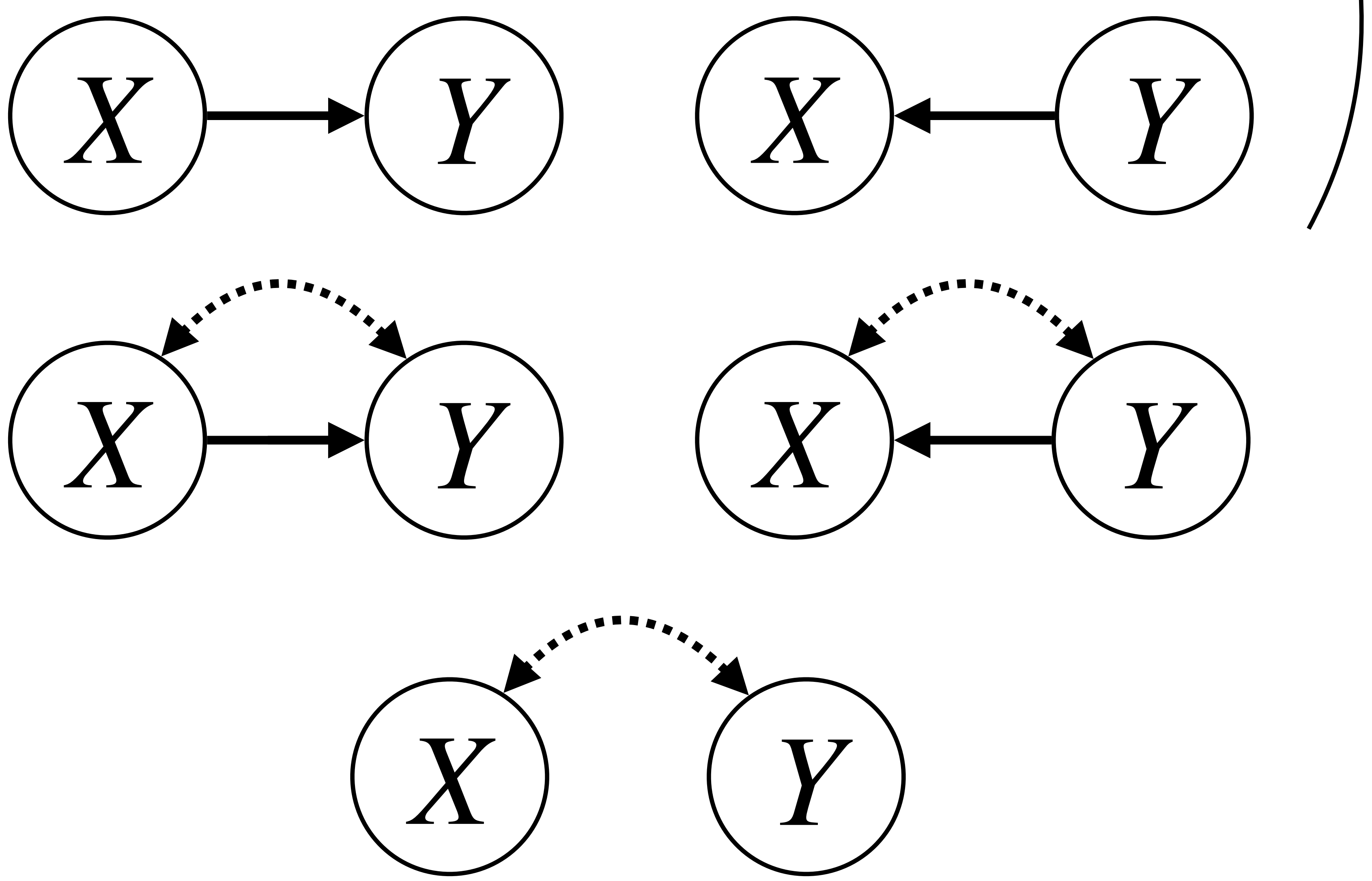
$$P(\mathbf{v}) = P(e \mid d, c, b, a, f)\ P(d \mid c, b, a, f)\ P(c \mid b, a, f)\ P(b \mid a, f)\ P(f \mid a)\ P(a)$$
$$= P(e \mid d, c, a)\ P(d \mid c, b, a)\ P(c \mid a)\ P(b \mid a)\ P(f \mid a)\ P(a)$$

$E \perp\!\!\!\perp F, B \mid D, C, A$ $\quad D \perp\!\!\!\perp F \mid B, C, A$ $\quad C \perp\!\!\!\perp F, B \mid A$ $\quad B \perp\!\!\!\perp F \mid A$

Markov Equivalence Class

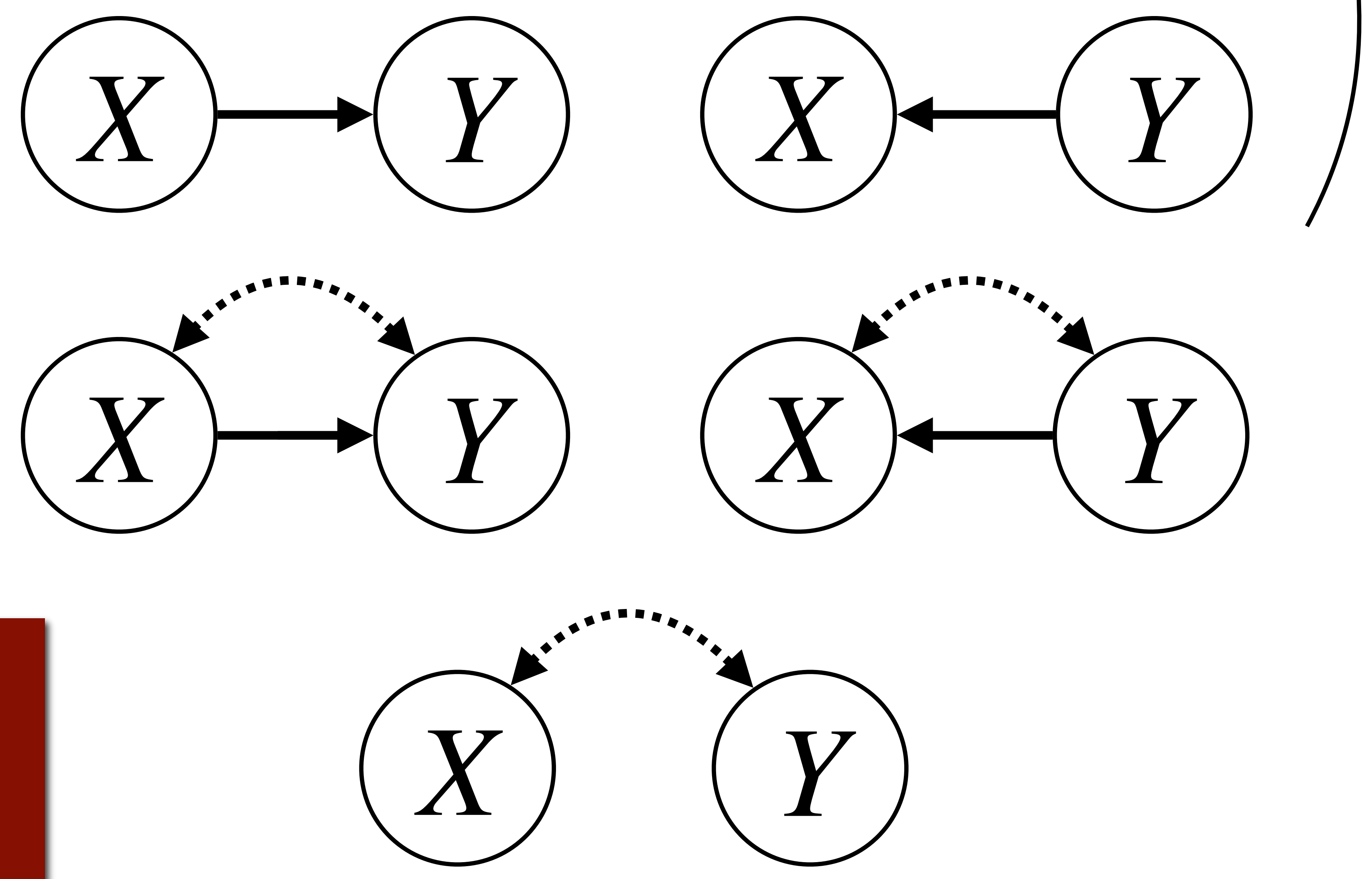
Definition (Markov Equivalence Class, MEC for short): A Markov Equivalence Class is a set of models that encode the same set of conditional independencies.

Markov Equivalent BNs

Distribution	**Factorization**	**Bayesian Networks**
$P(X, Y)$ with $P(Y\|X) \neq P(Y)$ i.e., $X \not\perp\!\!\!\perp Y$	$P(x, y) = P(y\|x)P(x)$ $P(x, y) = P(x\|y)P(y)$	

Markov Equivalence Class

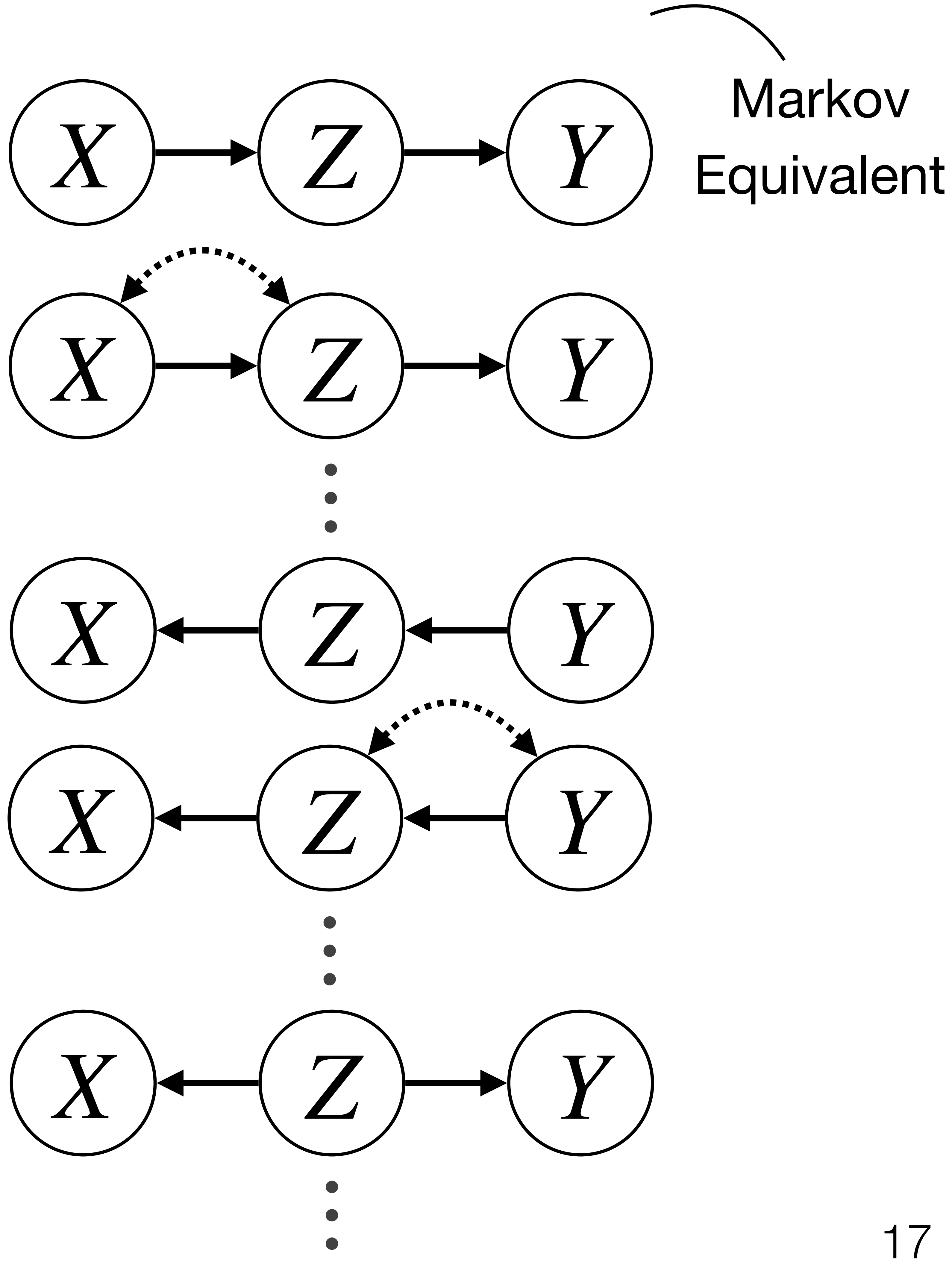
Definition (Markov Equivalence Class, MEC for short): A Markov Equivalence Class is a set of models that encode the same set of conditional independencies.

Distribution	Factorization	Bayesian Networks
$P(X, Y)$ with $P(Y \mid X) \neq P(Y)$ i.e., $X \not\perp\!\!\!\perp Y$	$P(x, y) = P(y \mid x)P(x)$ $P(x, y) = P(x \mid y)P(y)$	

All models imply no independence and no other invariance

Markov Equivalence Class

Distribution	Factorization	Bayesian Networks
$P(X, Y, Z)$ with $P(Y \mid X, Z) = P(Y \mid X)$ i.e., $X \perp\!\!\!\perp Y \mid Z$	$P(x, y, z) = P(y \mid x, z)P(z \mid x)P(x) = P(y \mid z)P(z \mid x)P(x)$	
	$P(x, y, z) = P(x \mid y, z)P(y \mid z)P(z) = P(x \mid z)P(z \mid y)P(y)$	
	$P(x, y, z) = P(y \mid x, z)P(x \mid z)P(z) = P(y \mid z)P(x \mid z)P(z)$	



Markov Equivalence Class

Distribution | **Factorization** | **Bayesian Networks**

$P(X, Y, Z)$

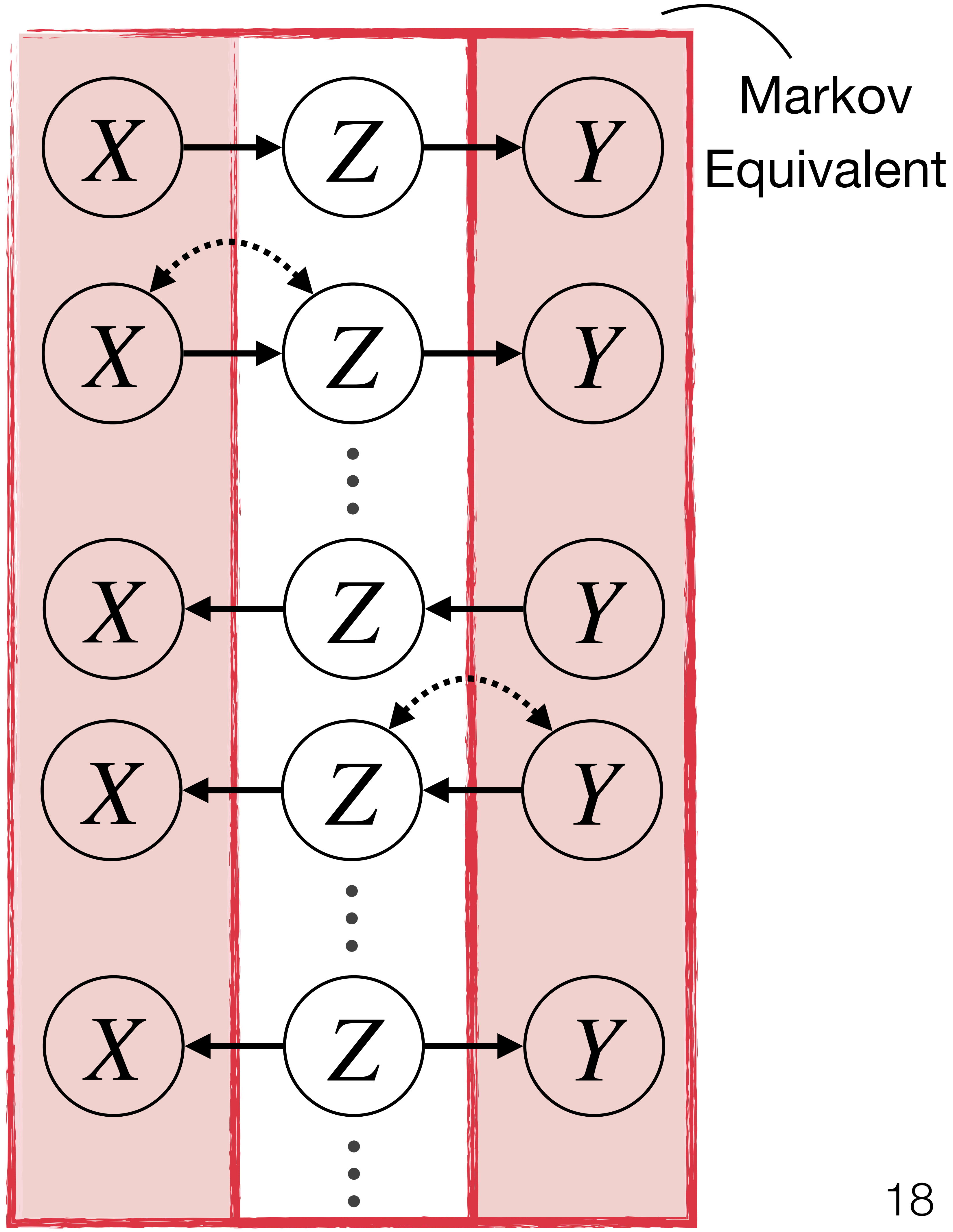
with $P(Y|X, Z) = P(Y|X)$

i.e., $X \perp\!\!\!\perp Y | Z$

$$P(x, y, z) = P(y|x, z)P(z|x)P(x) = P(y|z)P(z|x)P(x)$$

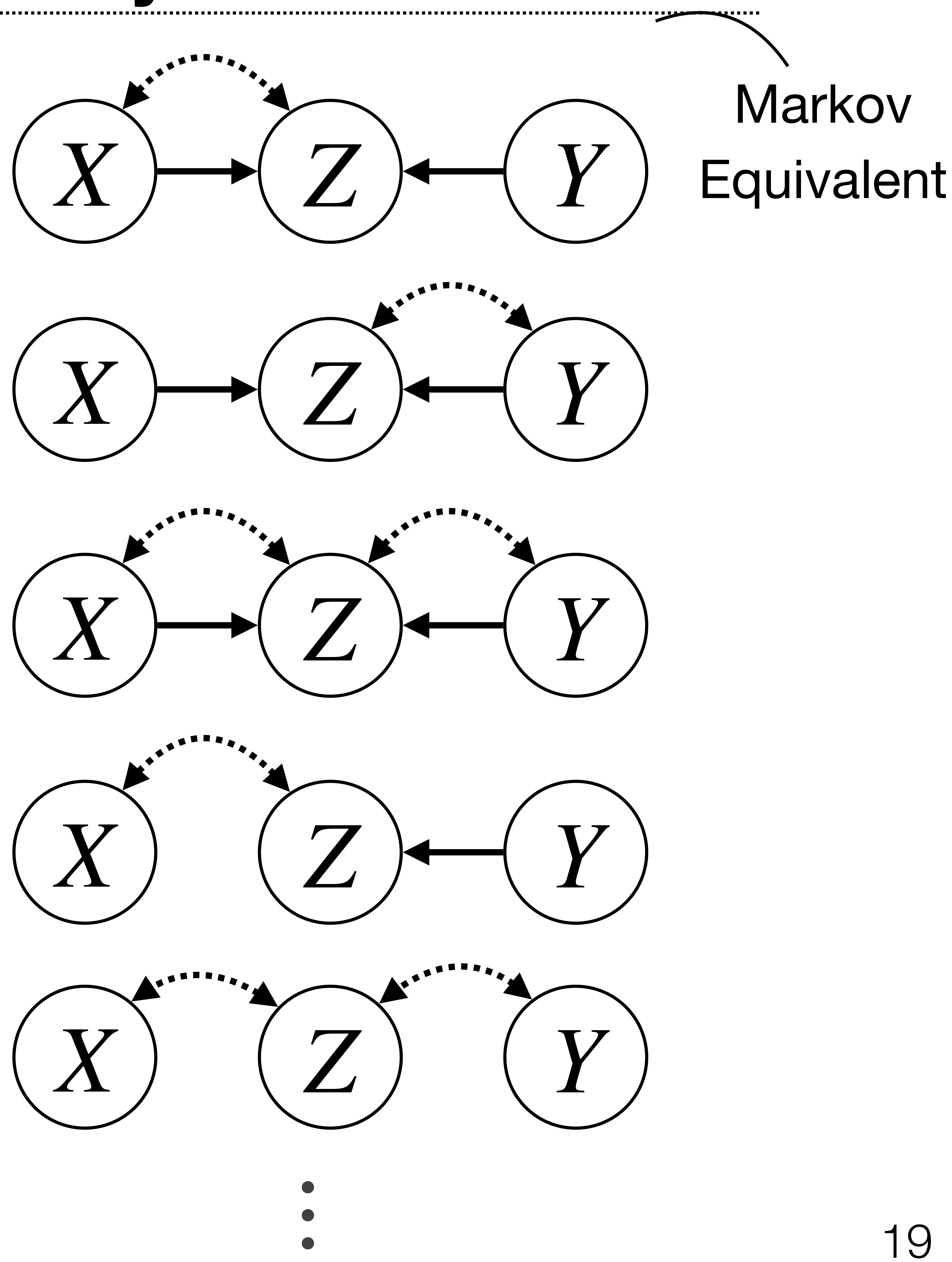
$$P(x, y, z) = P(x|y, z)P(y|z)P(z) = P(x|z)P(z|y)P(y)$$

$$= P(y|z)P(x|z)P(z)$$

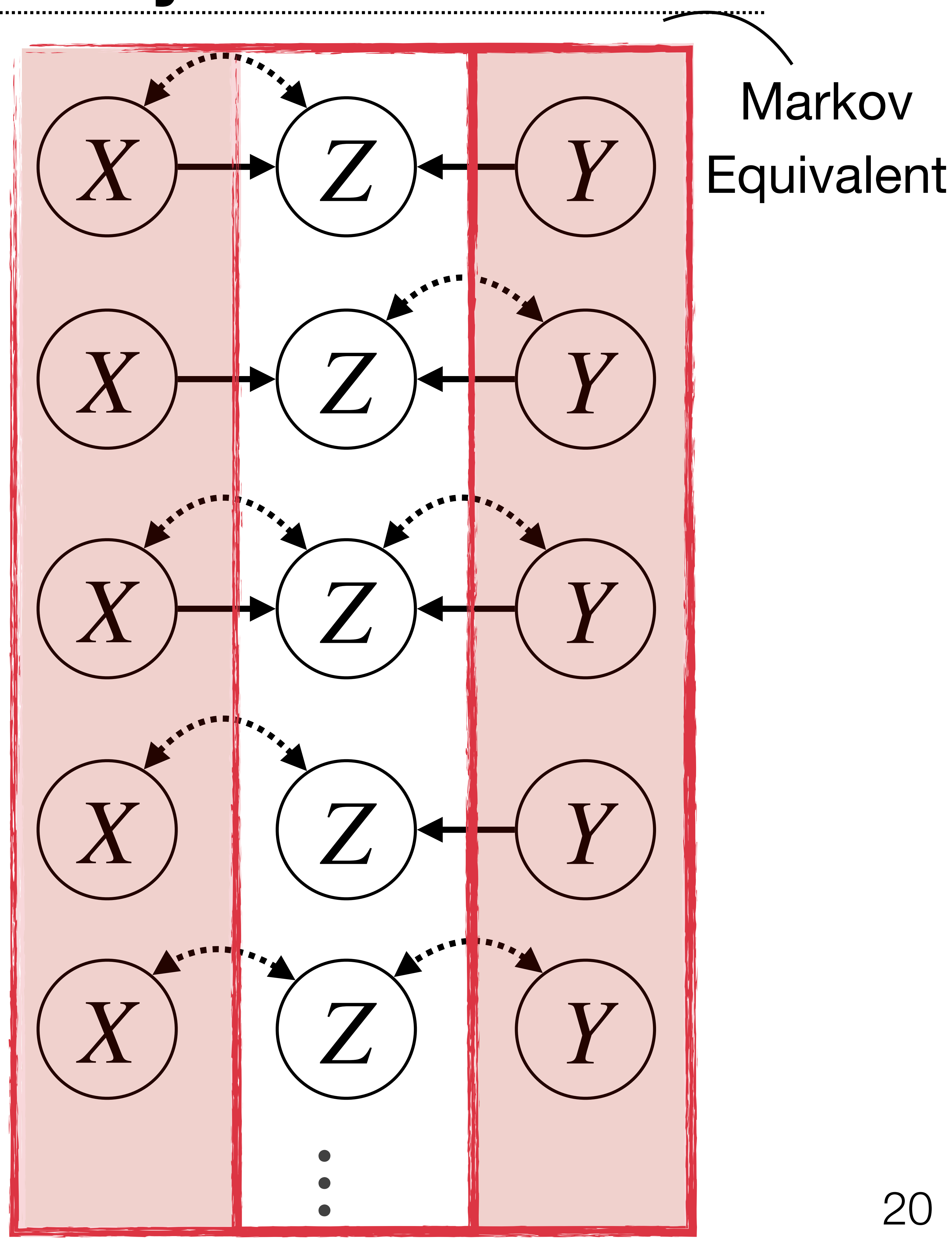


All models imply *only* $X \perp\!\!\!\perp Y | Z$ and
Z is always a *non-collider* in such models.

Markov Equivalence Class

Distribution	Factorization	Bayesian Networks
$P(X, Y, Z)$ with $P(Y \mid X) = P(Y)$ i.e., $X \perp\!\!\!\perp Y$	$P(x, y, z) = P(z \mid x, y)P(x \mid y)P(y)$ $= P(z \mid x, y)P(x)P(y)$	

Markov Equivalence Class

Distribution	Factorization	Bayesian Networks
$P(X, Y, Z)$ with $P(Y \mid X) = P(Y)$ i.e., $X \perp\!\!\!\perp Y$	$P(x, y, z) = P(z \mid x, y)P(x \mid y)P(y)$ $= P(z \mid x, y)P(x)P(y)$	

All models imply *only* $X \perp\!\!\!\perp Y$ and
Z is always a *collider* in such models,

Note: Z is never an ancestor of X or Y

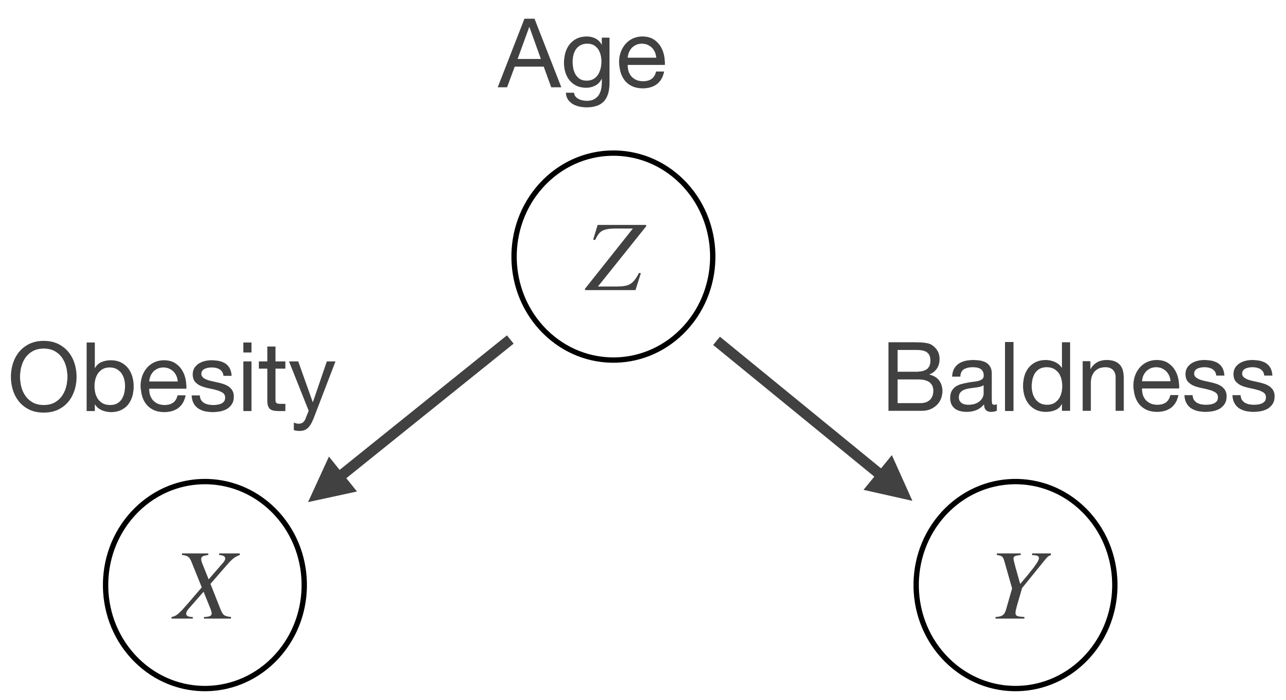
D-Separation

Graphical Tool for Identifying Conditional Independencies implied by Bayesian Networks

Implied Conditional independencies

Fork

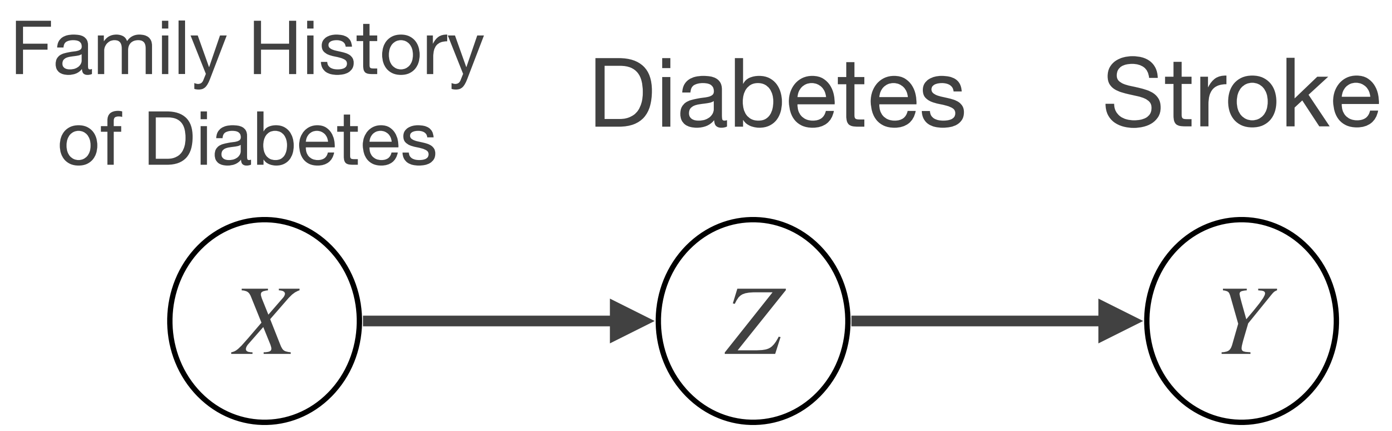
Z **as a common cause**



$$X \not\perp\!\!\!\perp Y$$
$$X \perp\!\!\!\perp Y \mid Z$$

Chain

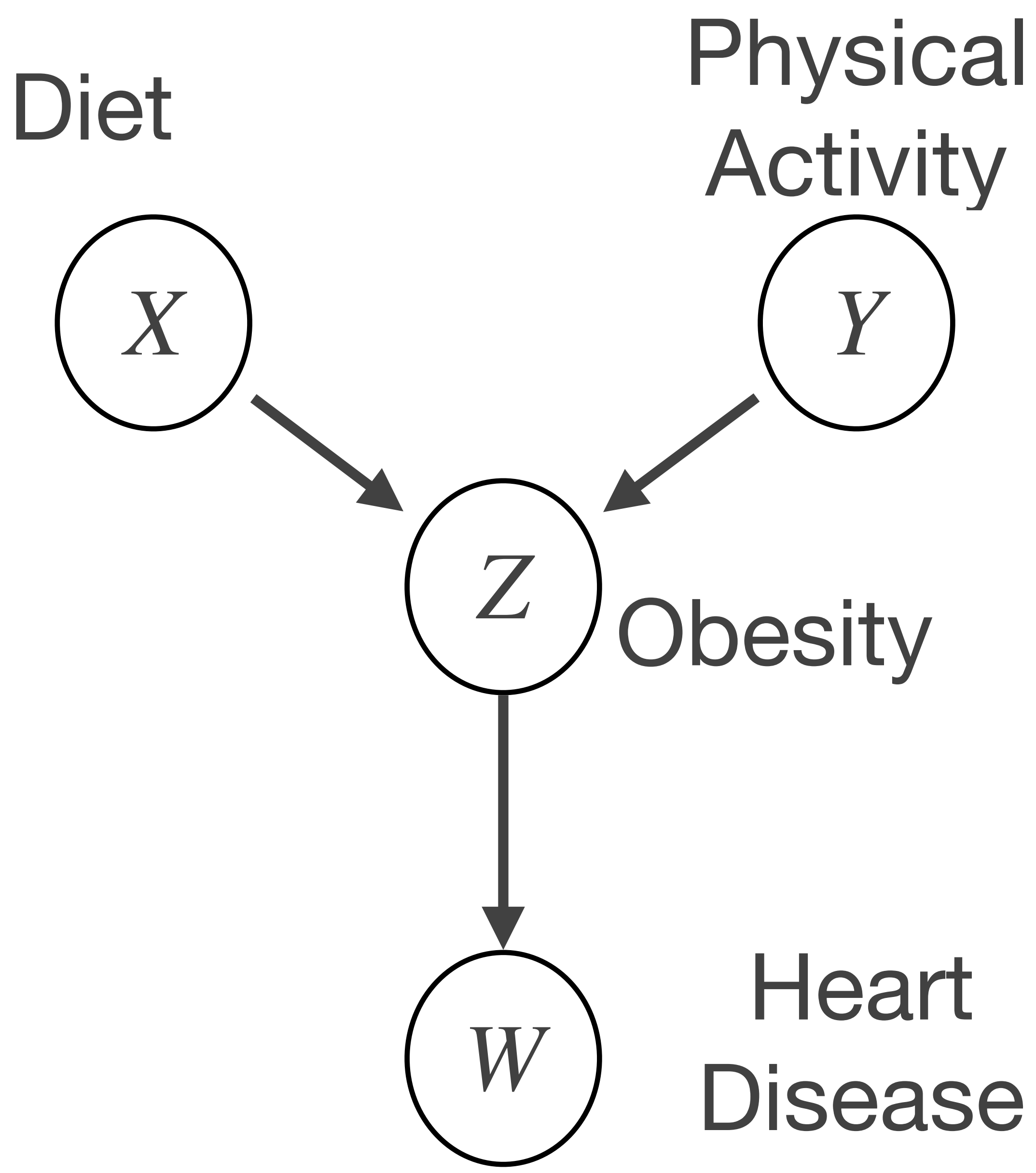
Z **as a mediator**



$$X \not\perp\!\!\!\perp Y$$
$$X \perp\!\!\!\perp Y \mid Z$$

V-Structure

Z **as a collider or common effect**



$$X \perp\!\!\!\perp Y$$
$$X \not\perp\!\!\!\perp Y \mid Z$$
$$X \not\perp\!\!\!\perp Y \mid W$$

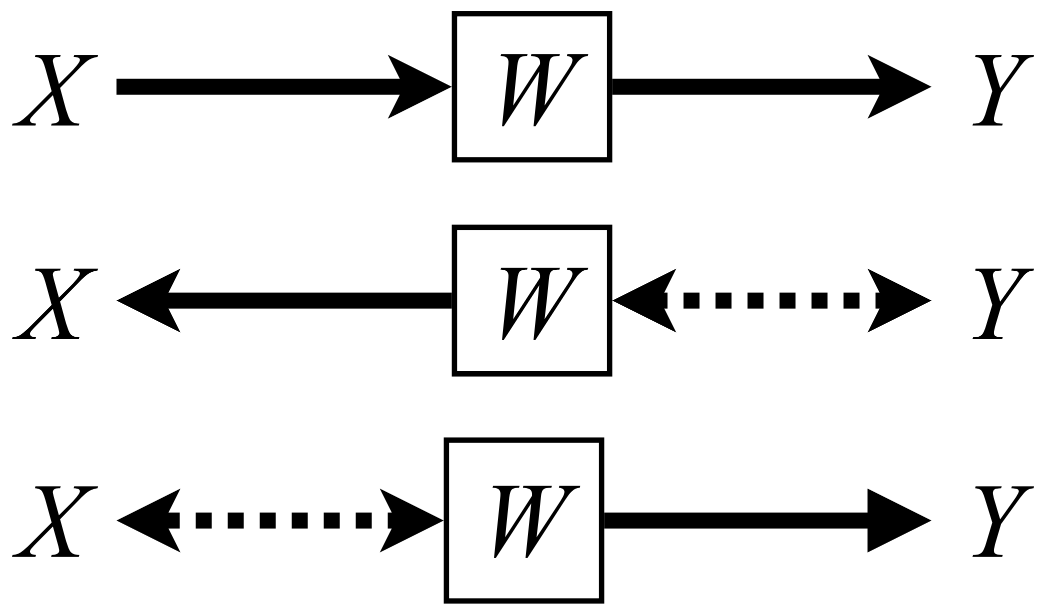
Two Markov-equivalent models.
Note that in both cases Z is a non-collider!

Active and Inactive Triplets

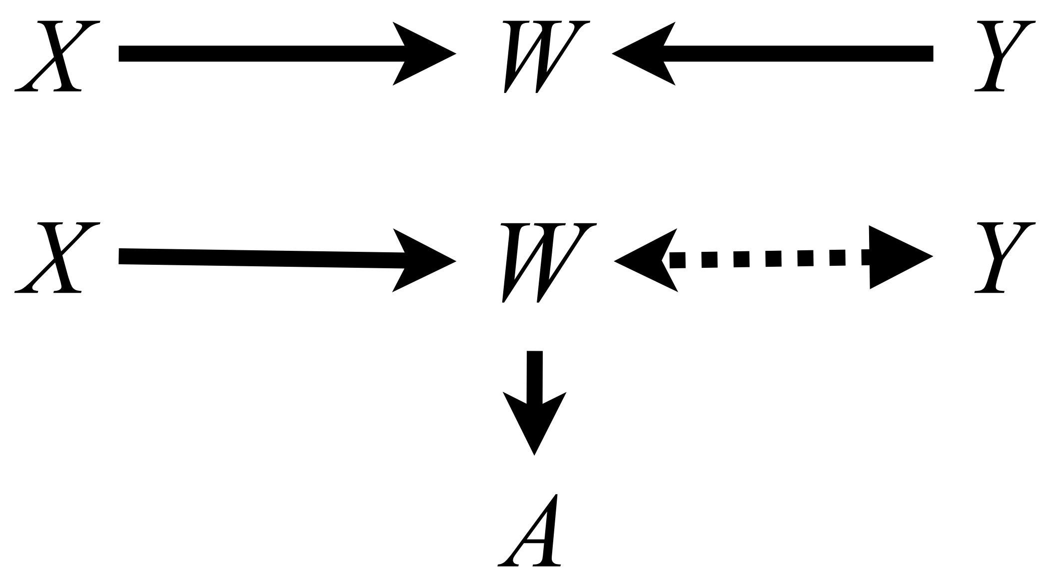
Definition (inactive): A triplet $\langle V_i, V_m, V_j \rangle$ is said to be ***inactive*** relative to a set $\mathbf{Z}$ if the middle node V_m:

1. Is a non-collider and is in $\mathbf{Z}$; or
2. Is a collider and neither it nor any of its descendants in $\mathbf{Z}$.

W is non-collider and $W \in \mathbf{Z}$



W is (descendant of) a collider and $W, A \notin \mathbf{Z}$

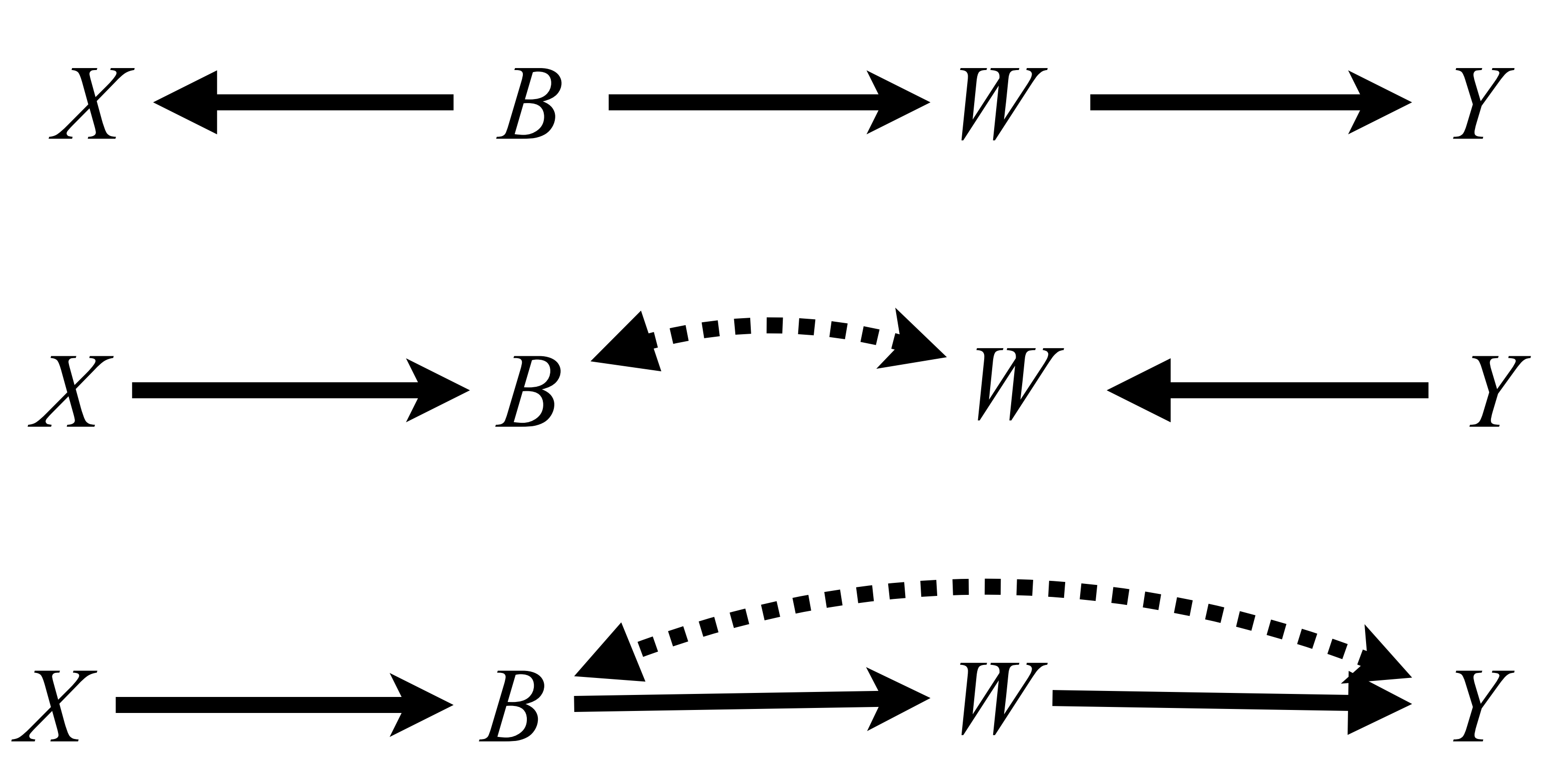


D-Separation

Definition (d-separation): A path p in an ADMG G is said to be ***d-separated*** (or blocked) by a set of variables $\mathbf{Z}$ if and only if p contains an inactive triplet in it.

A set $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ if and only if $\mathbf{Z}$ blocks every path between a node in $\mathbf{X}$ and a node in $\mathbf{Y}$. We denote that by $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_G$.

Does $\mathbf{Z}$ d-separate X and Y?



Graph	$\{\}$	$\{B\}$	$\{W\}$	$\{B, W\}$
1: $X \leftarrow B \rightarrow W \rightarrow Y$	✗	✓	✓	✓
2: $X \rightarrow B \leftrightarrow W \leftarrow Y$	✓	✓	✓	✗
3: $X \rightarrow B \rightarrow W \rightarrow Y$, $B \leftrightarrow Y$	✗	✗	✗	✗

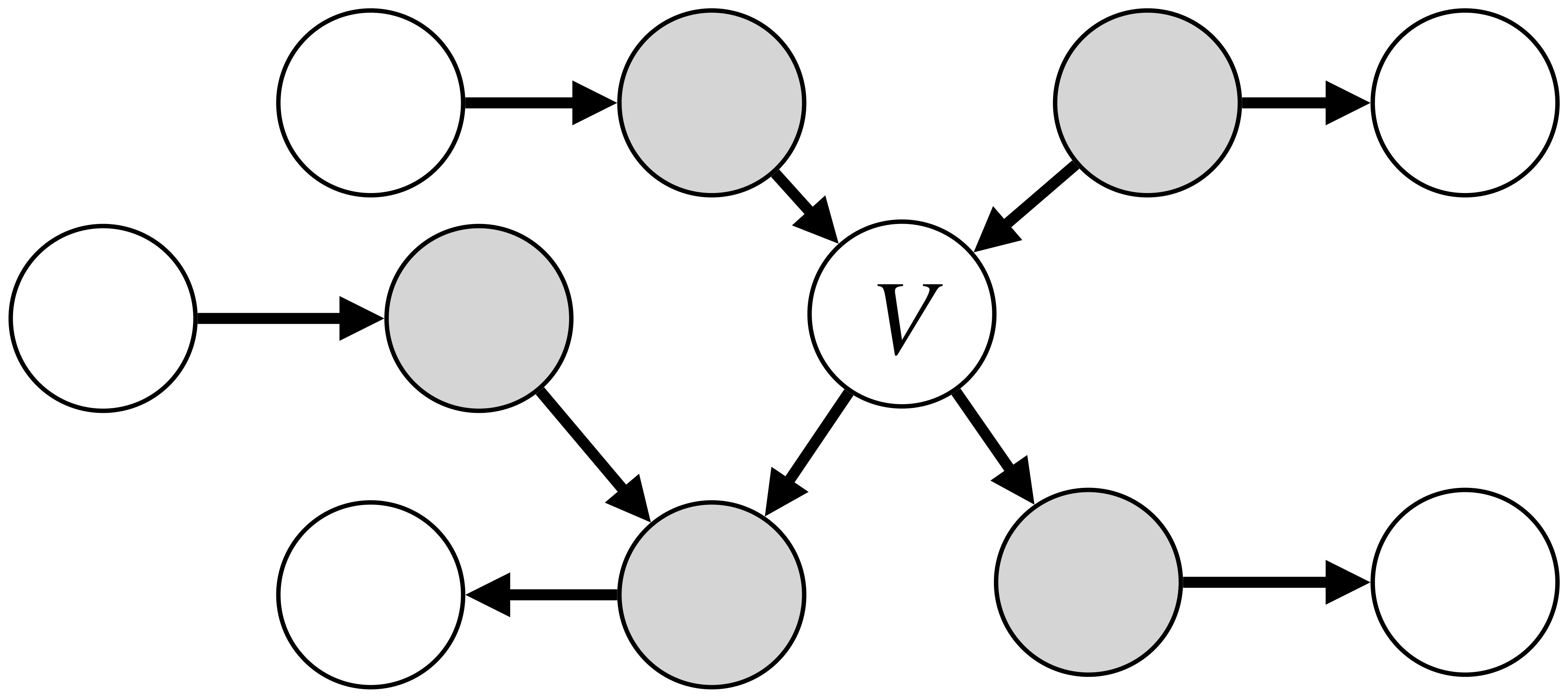
$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_G \Rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_P$$

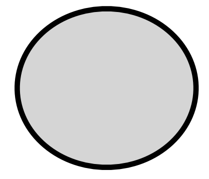
D-separations in G correspond to conditional independencies in P

Markov Blanket (Markovian)

Markov Blanket (MB) of a *Markovian* BN over V: the union of parents, children, and parents of the children V.

$$\text{mb}_G(V) = Pa(V) \cup Ch(V) \cup Pa(Ch(V))$$





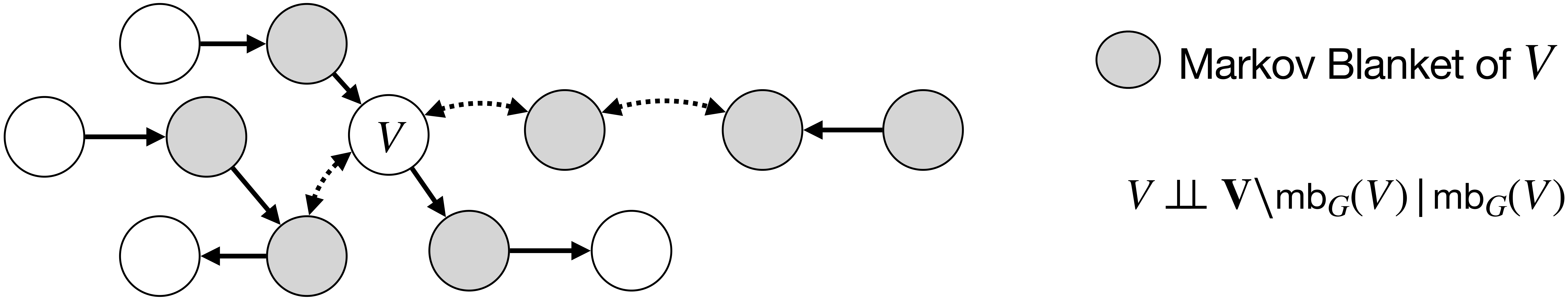
$$V \perp\!\!\!\perp \mathbf{V} \setminus \text{mb}_G(V) \mid \text{mb}_G(V)$$

Markov Blanket (Semi-Markovian)

Markov Blanket (MB) of a *Semi-Markovian* BN over V: is the district of V and the parents of the district of V (excluding V itself) i.e.:

$$\text{mb}_G(V) = \text{dis}_G(V) \cup \text{Pa}_G(\text{dis}_G(V)) \setminus \{V\}$$

District of V, $\text{dis}_G(V)$, is the set of variables connected with V through an edge or a bidirected path.



$$V \perp\!\!\!\perp \mathbf{V} \setminus \text{mb}_G(V) \,|\, \text{mb}_G(V)$$

Richardson, T. (2003). Markov Properties for Acyclic Directed Mixed Graphs. *Scandinavian Journal of Statistics*, *30*(1), 145–157

Causal Discovery

Learning the Markov equivalence class from observational data.

Super-Exponential Growth

The space of DAGs grows super-exponentially with the number *n* of variables, as shown by the following recurrence relation (Robinson, 1973):

$$|DAG(n)| = \sum_{i=1}^{n} \binom{n}{1} 2^{i(n-i)} |DAG(n-1)|$$

n	$\|DAG(n)\|$
2	3
3	27
4	729
5	59,049
6	1.4349×10^7
7	1.0460×10^{10}
8	2.2877×10^{13}

Super-Exponential Growth

The space of ADMGs also grows super-exponentially with the number *n* of variables, and it is much bigger than the space of DAGs:

$$|ADMG(n)| = |DAG(n)| \times 2^{n(n-1)/2}$$

$$|ADMG(n)| \gg |DAG(n)|$$

Causal Discovery is not feasible through naive enumeration!

n	$\|DAG(n)\|$	$\|ADMG(n)\|$
2	3	6
3	27	216
4	729	46,656
5	59,049	6.0457×10^7
6	1.4349×10^7	4.7019×10^{11}
7	1.0460×10^{10}	2.1936×10^{16}
8	2.2877×10^{13}	6.1410×10^{21}

Learning the Markov Equivalence Class

Identifiability: In non-parametric settings (i.e., without making parametric or distributional assumptions) and solely from observational data, causal discovery algorithms can only learn a graphical representation of a *Markov equivalence class*!

Algorithms: Score-Based vs Constraint-Based

Systems: Causal Sufficient vs Causal Insufficient

Causal Sufficiency: assumption that all confounding variables have been observed — although strong, it has been widely employed to simplify causal discovery and inference.

Score-Based Causal Discovery Algorithms

Strategy: search for the most probable causal structure by assessing goodness-of-fit scores of different possible structures.

Common Scores: Bayesian Information Criterion (BIC) for Gaussian variables and the BDeu score for multinomial variables.

Under causal sufficiency:

- **GES:** Greedy Equivalence Search, by Chickering, 2003.
- **FGES:** Fast GES, by Ramsey et al., 2017 — extension of the GES that improves the runtime of the algorithm by using parallelization.

Score-Based Causal Discovery Algorithms

Accounting latent confounding:

Use BIC, assuming linear Gaussian models

- **GSMAG:** a greedy search algorithm for learning MAGs, by Triantafillou, S. and Tsamardinos, I., 2016.
- **MAGSL:** search based on **dynamic programming and branch and bound**, by Rantanen et al., 2021 — it is guaranteed to find a globally optimal MAG.
- **Diff-discovery:** solves a **continuous optimization problem** with differentiable procedures to find the best fitting ADMG, by Bhattacharya et al., 2021.
- **N-ADMG:** Neural ADMG Learning, by Ashman et al., 2013 — extends Diff-discovery to the setting where the true causal diagram is bow-free and corresponds to a non-linear SCM with additive noise.

Constraint-Based Causal Discovery Algorithms

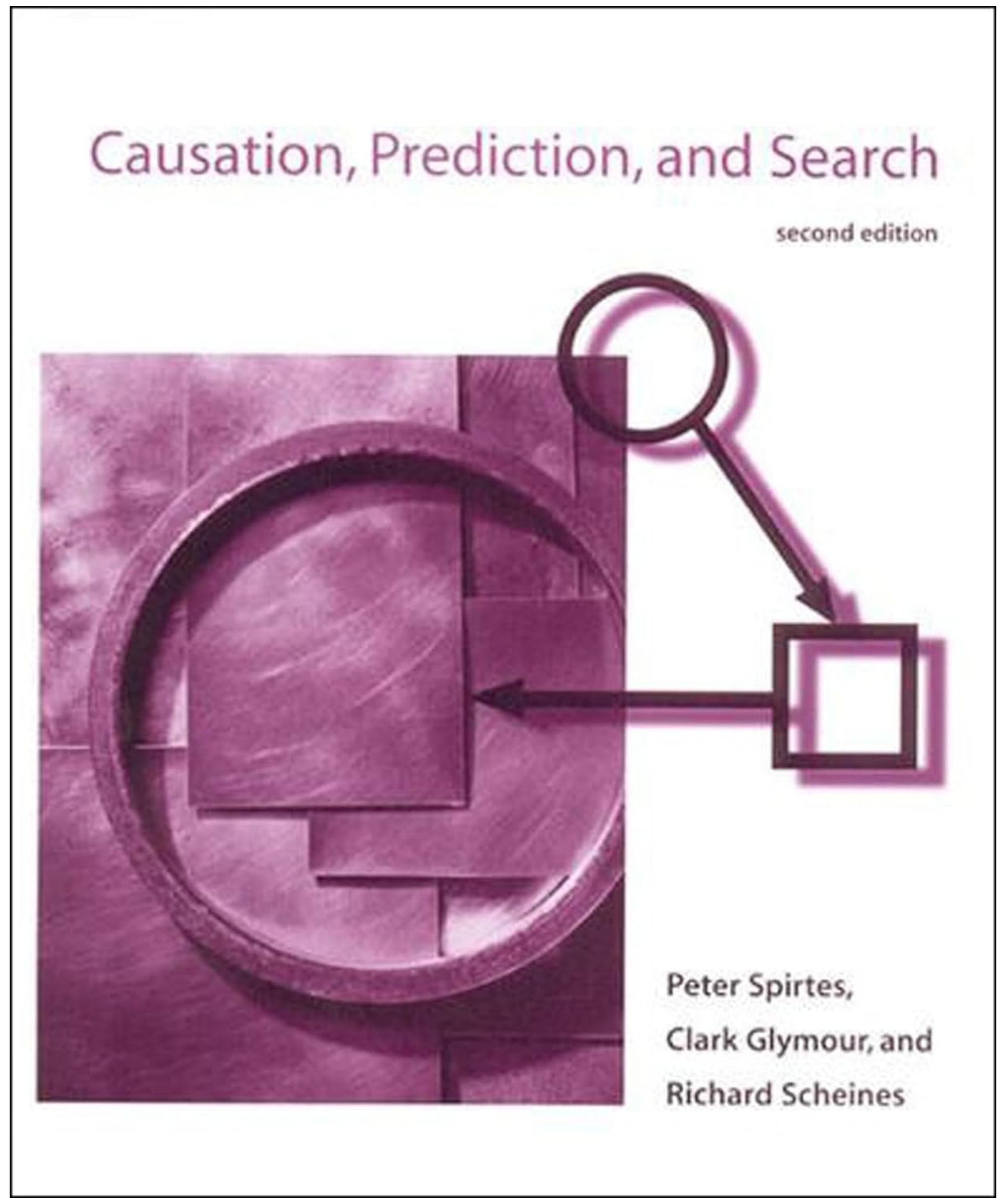
Strategy: construct a causal structure that aligns with all observed conditional independencies, identified using conditional independence tests.

Under causal sufficiency:

IC: Inductive Causation, by Verma and Pearl, 1990.

PC: Peter-Clark, by Spirtes and Glymour, 1991.

They start with an adjacency (skeleton) phase, based on conditional independence tests, followed by an orientation phase.



Spirtes, P., Glymour, C., and Scheines, R. (2001).
Causation, Prediction, and Search, 2nd edn. Cambridge, MA: MIT Press.

Constraint-Based Causal Discovery Algorithms

Accounting for latent confounding:

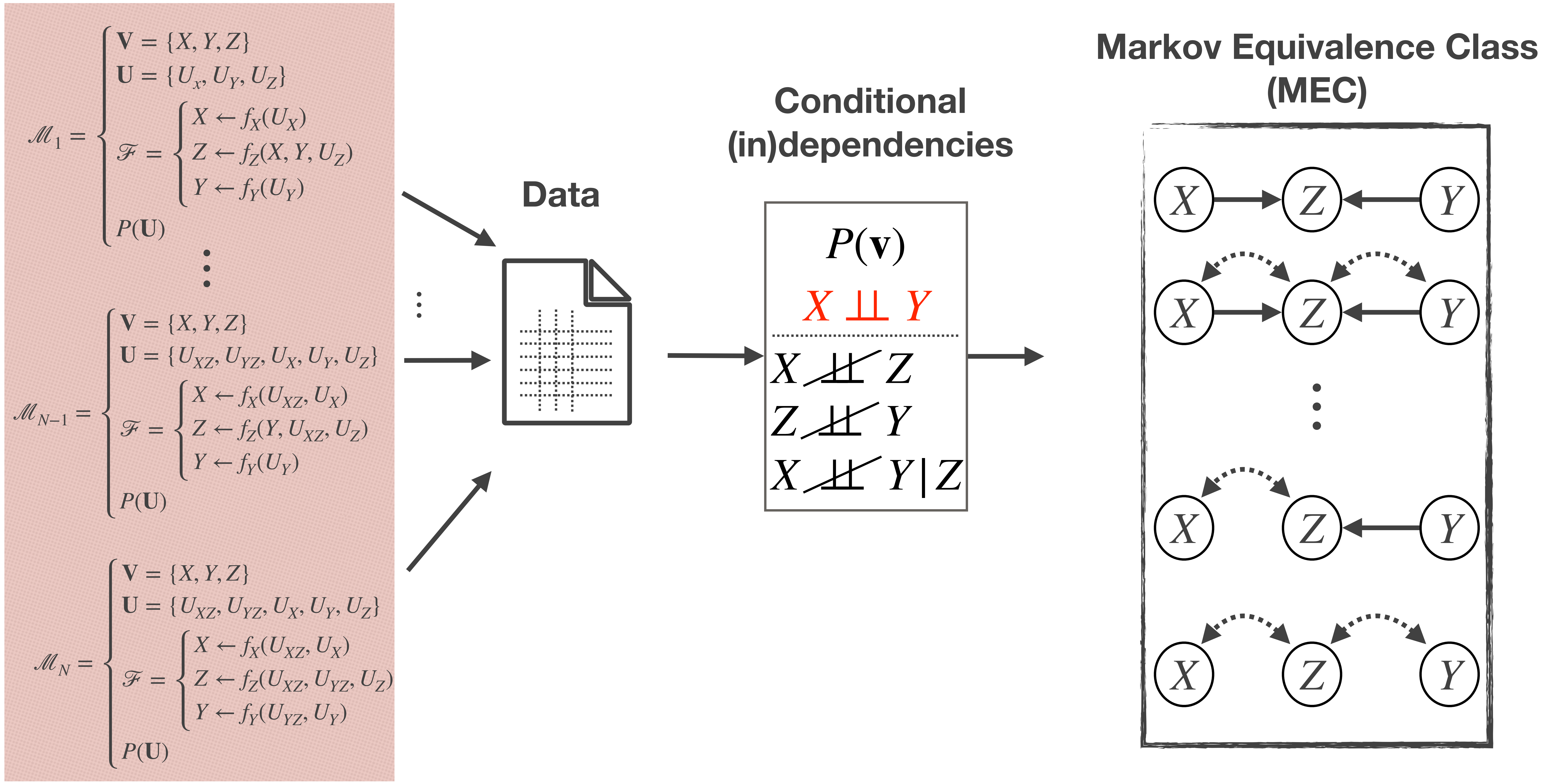
- **FCI:** Fast Causal Inference, by Spirtes et al., 1995 — most prominent extension of the PC and IC/IC* algorithms. Together with the additional rules by Zhang, J. (2008), is a complete algorithm accounting for both latent confounding and selection bias.

- **FCI variants:** Anytime FCI (**AFCI**), by Spirtes P., 2001, Conservative FCI (**CFCI)** and Really FCI (**RFCI**), by Colombo et al. 2012; and **FCI+**, by Claassen et al. 2013.

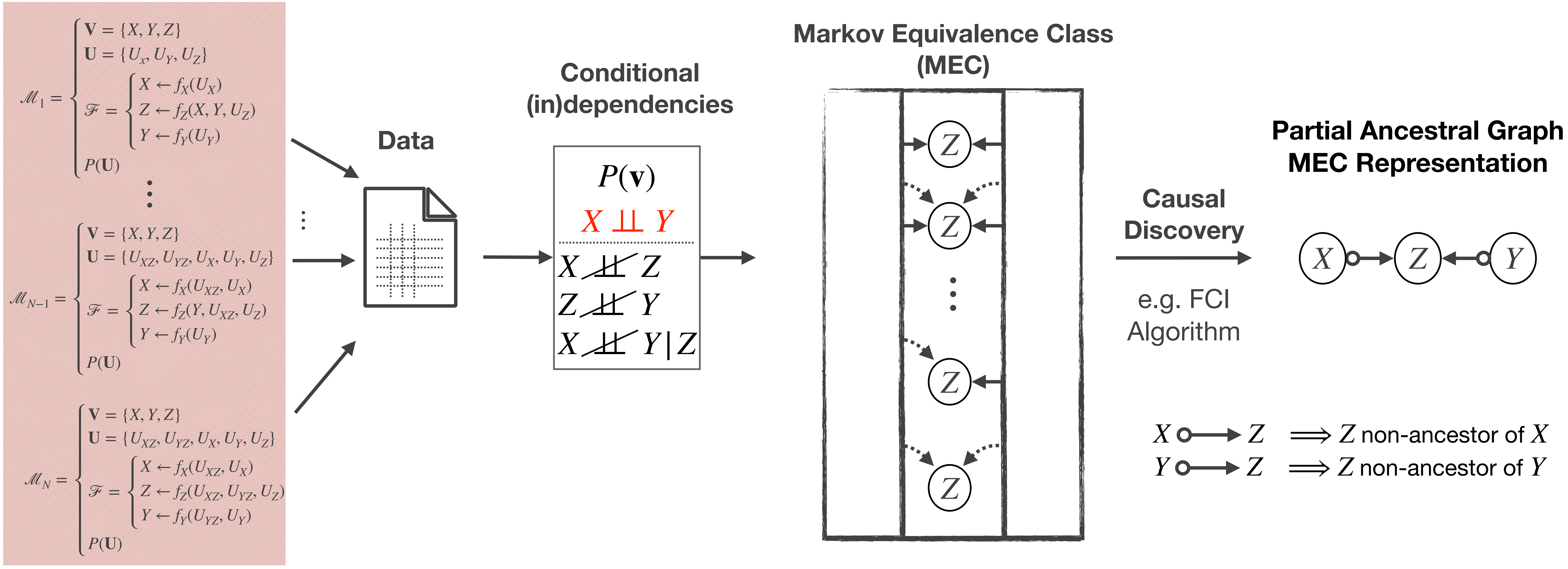
- **SAT-Based:** uses a Answer Set Programming (ASP) solver to find a causal structure that most satisfies the minimal observed conditional independencies, by Hyttinen et al., 2014.

- **ACI:** Ancestral Causal Inference — a logic-based algorithm by Magliacane et al., 2016.

Causal Discovery: Learning Structural Invariances



Causal Discovery: Learning Structural Invariances

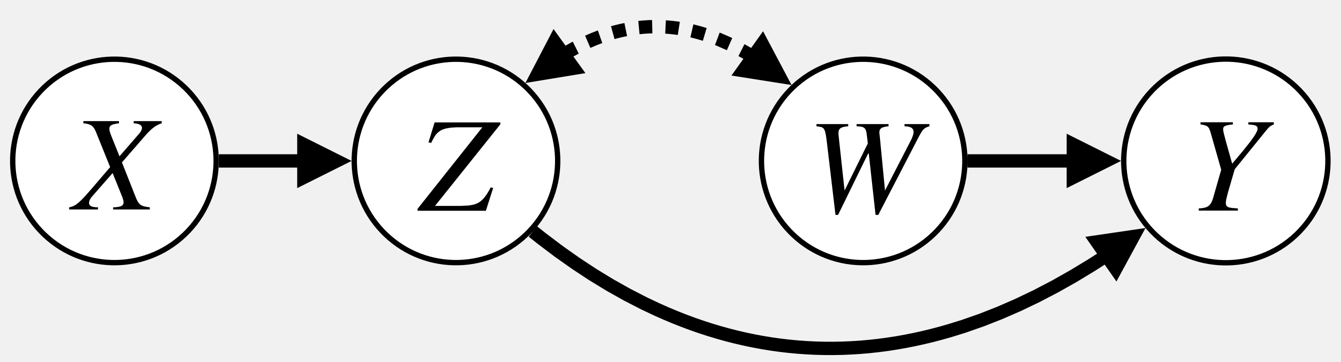


Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896. [Link]

FCI Algorithm - Pipeline

Unknown Reality

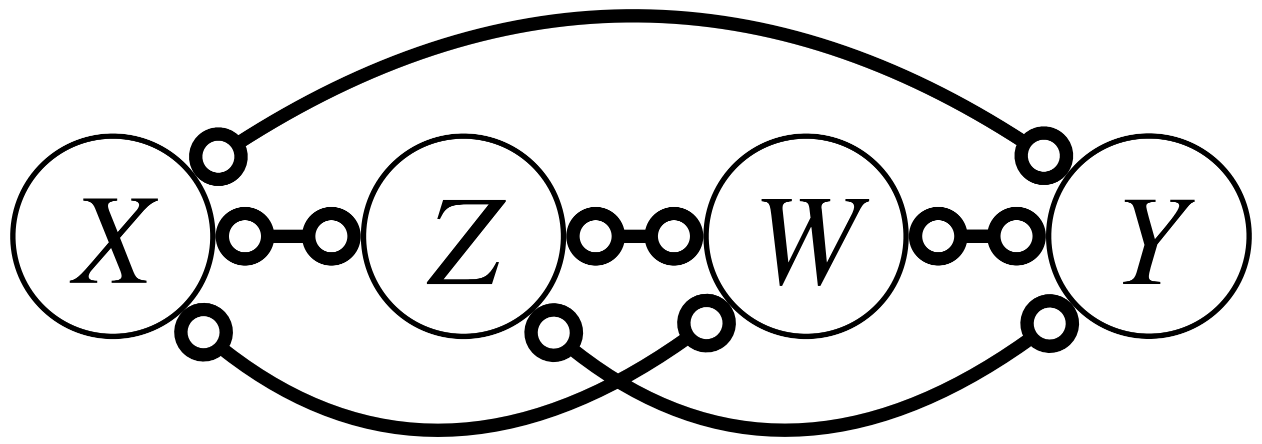
True causal diagram



$X \perp\!\!\!\perp W$

$X \perp\!\!\!\perp Y | Z, W$

Implied by the ADMG using d-separation



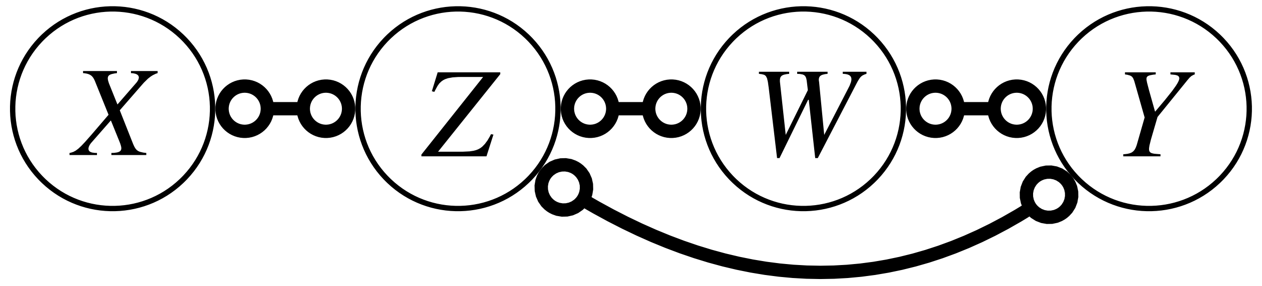
Complete Graph

Conditional Independence Tests

$X \perp\!\!\!\perp W$

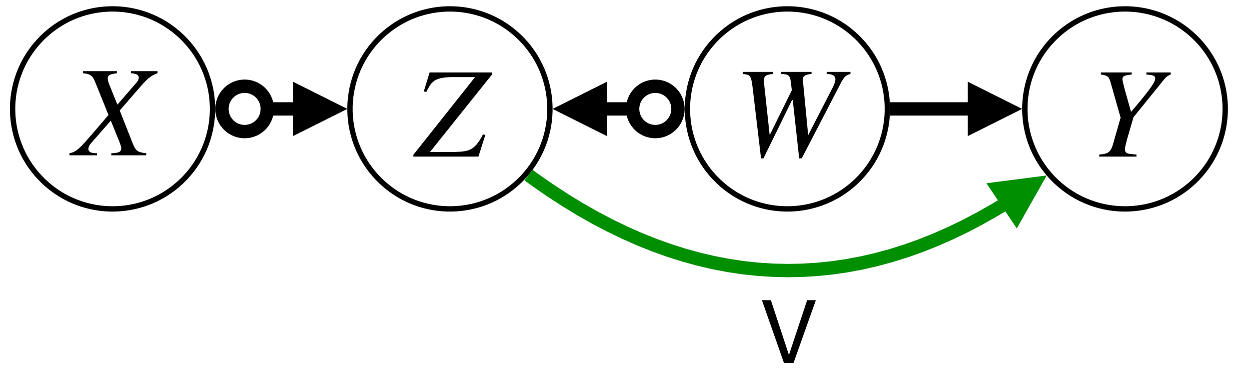
$X \perp\!\!\!\perp Y | Z, W$

By **faithfulness**, are correctly observed in the data



Skeleton

FCI Rules $(R1) - (R10)$



Partial Ancestral Graph (PAG)

Implied by the PAG using m-separation

$X \perp\!\!\!\perp W$

$X \perp\!\!\!\perp Y | Z, W$

$A \circ\!\!\rightarrow B \implies$ B non-ancestor of A

$A \rightarrow B \implies$ A ancestror of B

$A \leftrightarrow B \implies$ spurious association

$A - B \implies$ selection bias

Z is not an ancestor of X or W.

Z and W are ancestors of Y.

Z is not confounded with Y.

Conditional Independence Tests

Gaussian errors and independent observations: partial correlation test

Fisher, R.A. (1921). *On the Probable Error of a Coefficient of Correlation Deduced from a Small Sample.*
R package: https://cran.r-project.org/web/packages/pcalg/

Kernel-based non-parametric test:

Zhang, K., Peters, J., Janzing, D., & Schölkopf, B. (2012). *Kernel-based conditional independence test and application in causal discovery.* In: Uncertainty in artificial intelligence. AUAI Press; 2011. p.804–13
R package: https://cran.r-project.org/web/packages/CondIndTests

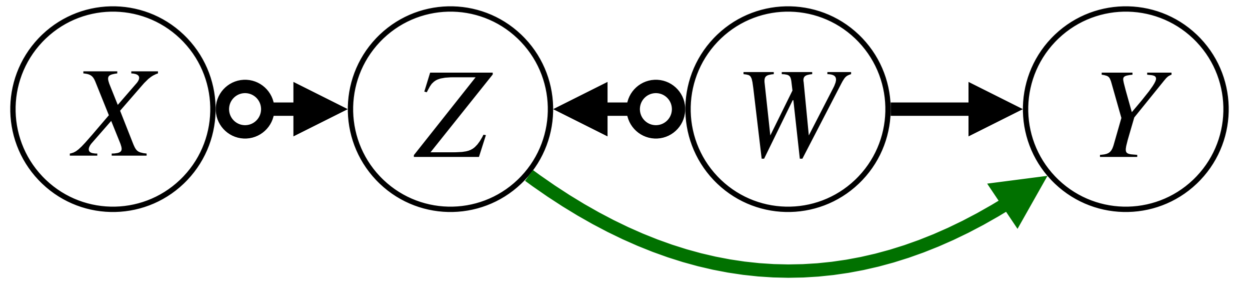
Continuous (conditional Gaussian) or Discrete (Binary, Ordinal, Multinomial) - Linear Regression

- **Tsagris, M., Borboudakis, G., Lagani, V. *et al.*** (2018) Constraint-based causal discovery with mixed data. *Int J Data Sci Anal* **6**, 19–30. (Link)
- R package: https://cran.r-project.org/web/packages/MXM/

Gaussian errors and correlated observations (family data) :

Ribeiro A.H., Soler J.M.P. (2020). *Learning Genetic and environmental graphical models from family data*, Statistics in Medicine.
R package: https://github.com/adele/FamilyBasedPGMs

PAG: Representation of the Markov Equivalence Class



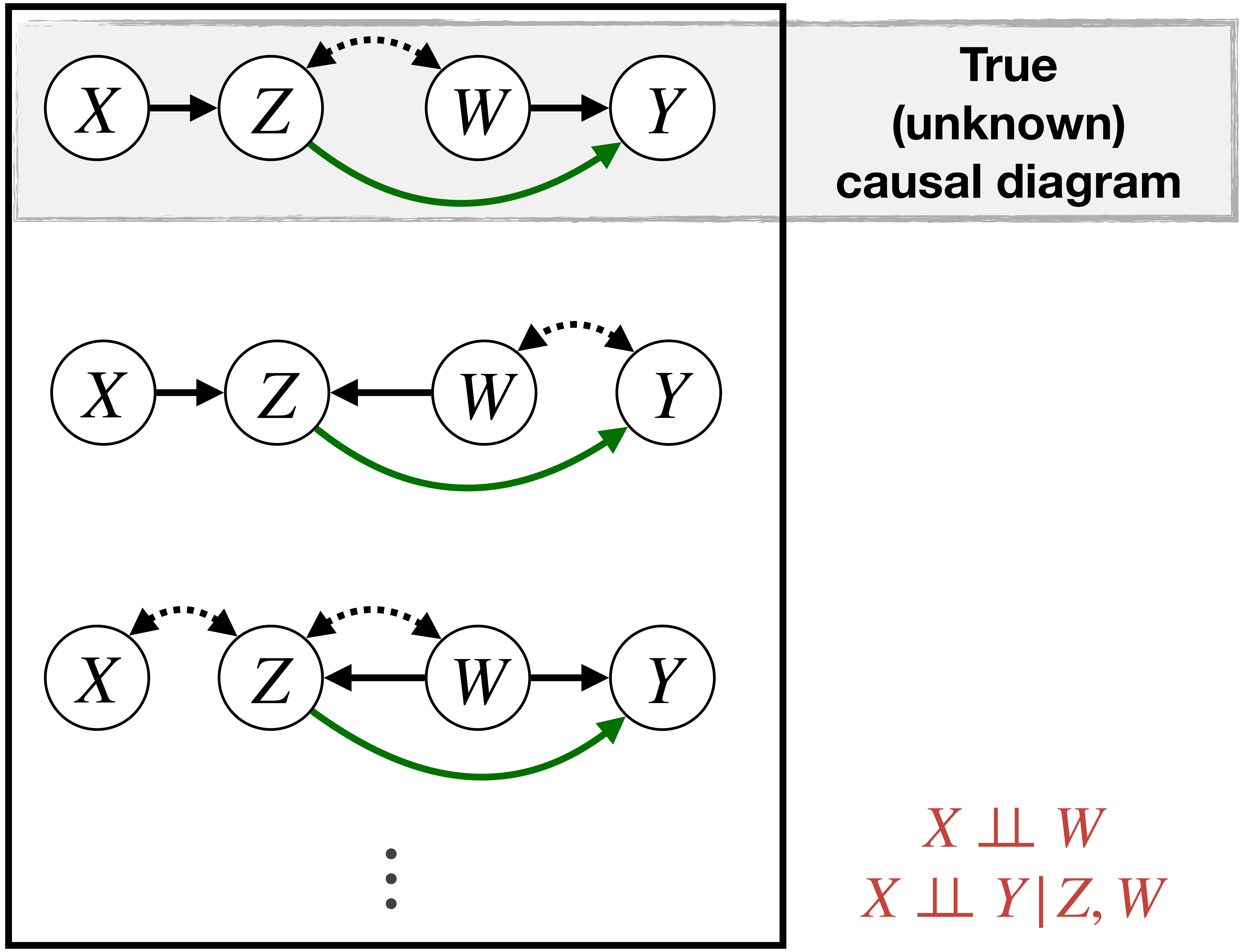
V

Partial Ancestral Graph
(PAG)

Z is not an ancestor of X or W.

Z and W are ancestors of Y.

Z is not confounded with Y.

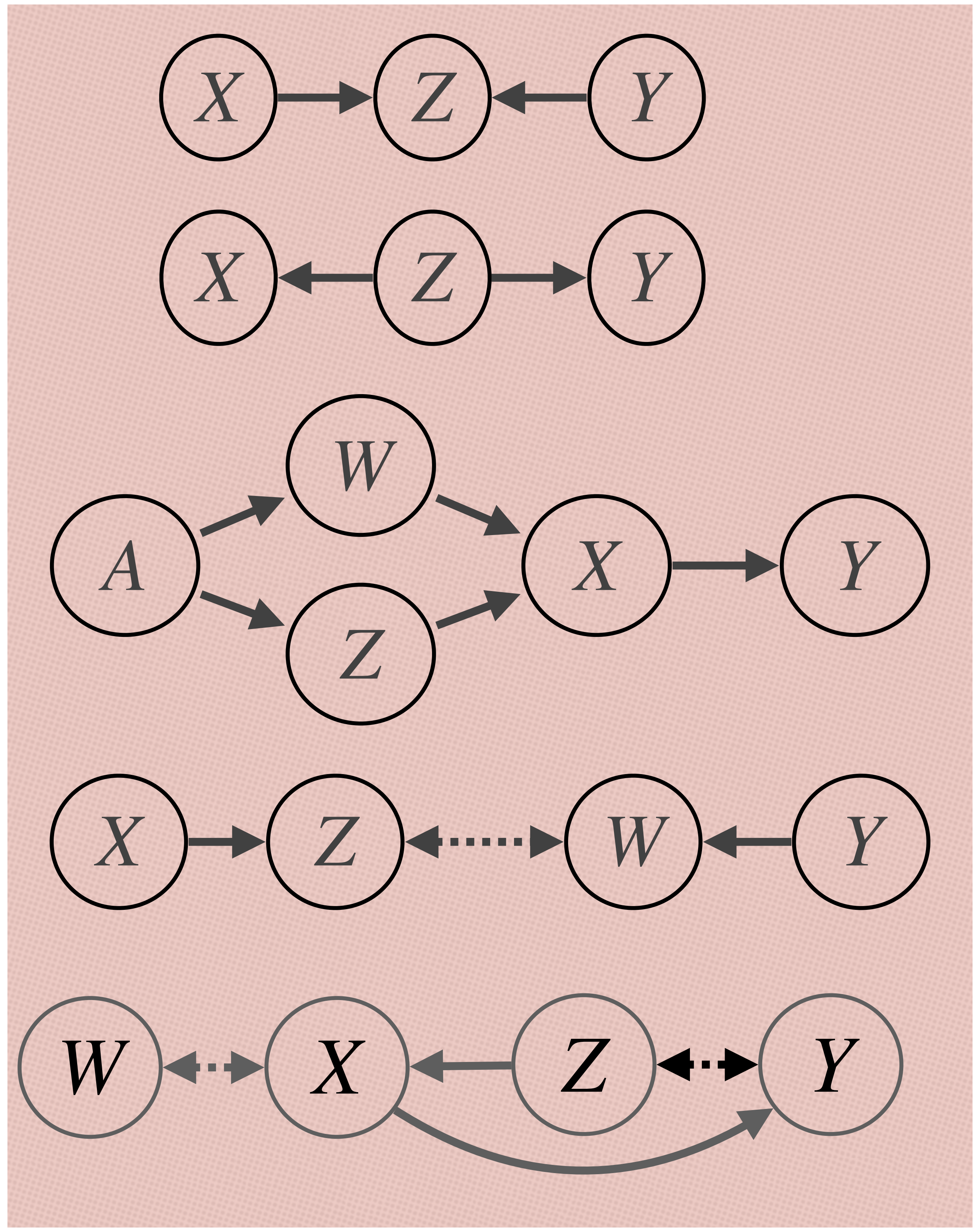


$X \perp\!\!\!\perp W$

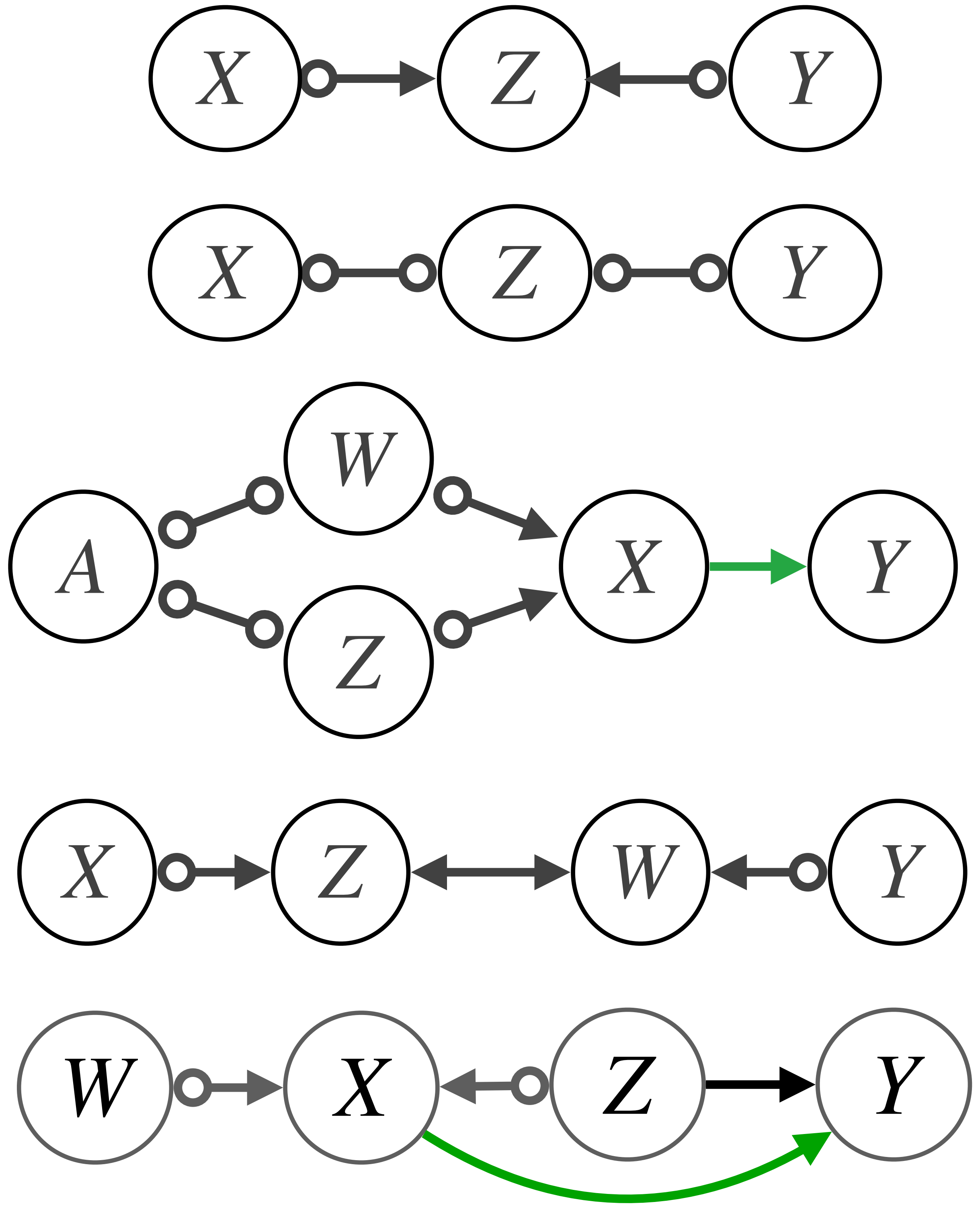
$X \perp\!\!\!\perp Y \mid Z, W$

Fast Causal Inference (FCI) Algorithm

Underlying Causal Diagram

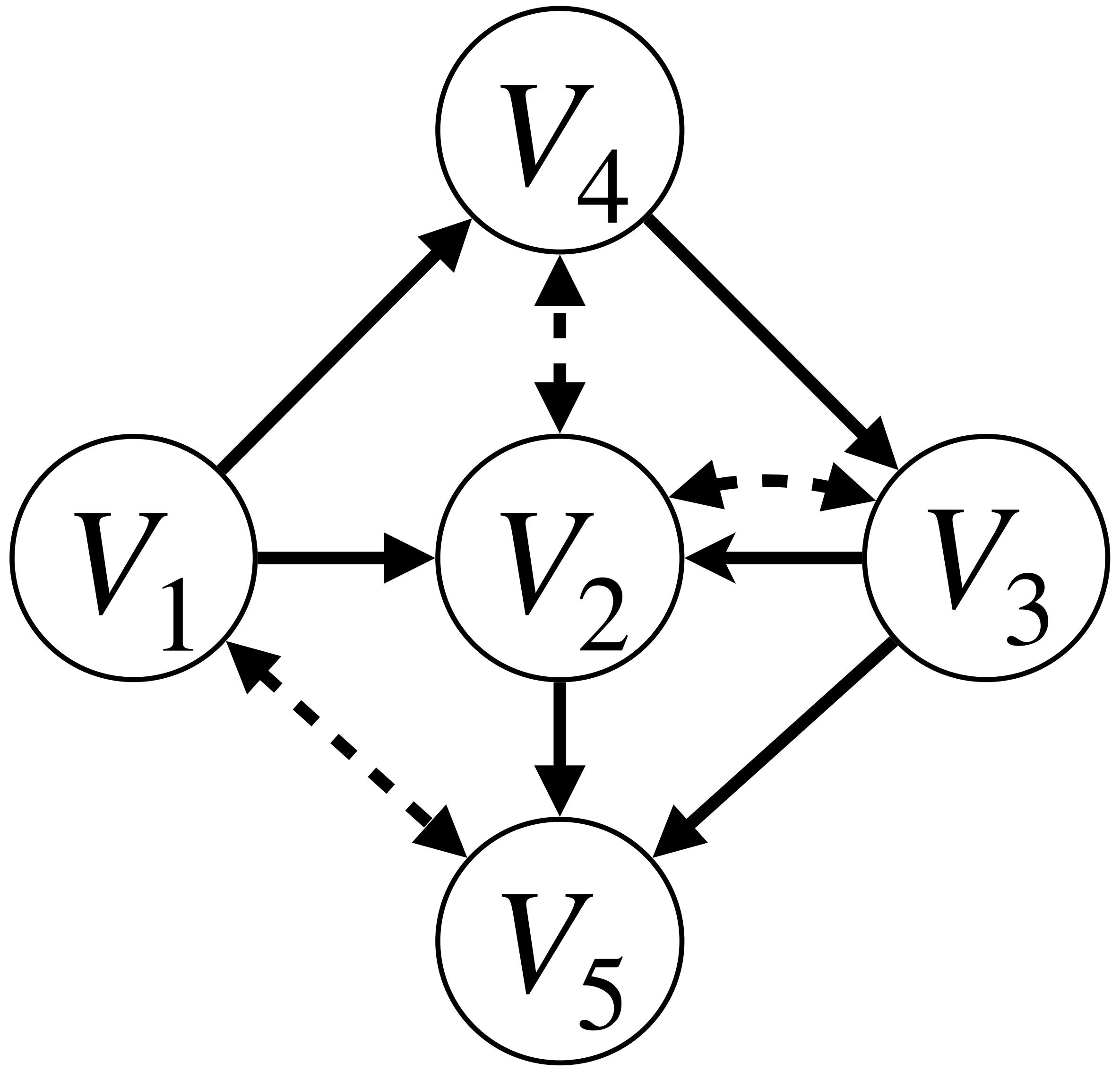


Partial Ancestral Graph

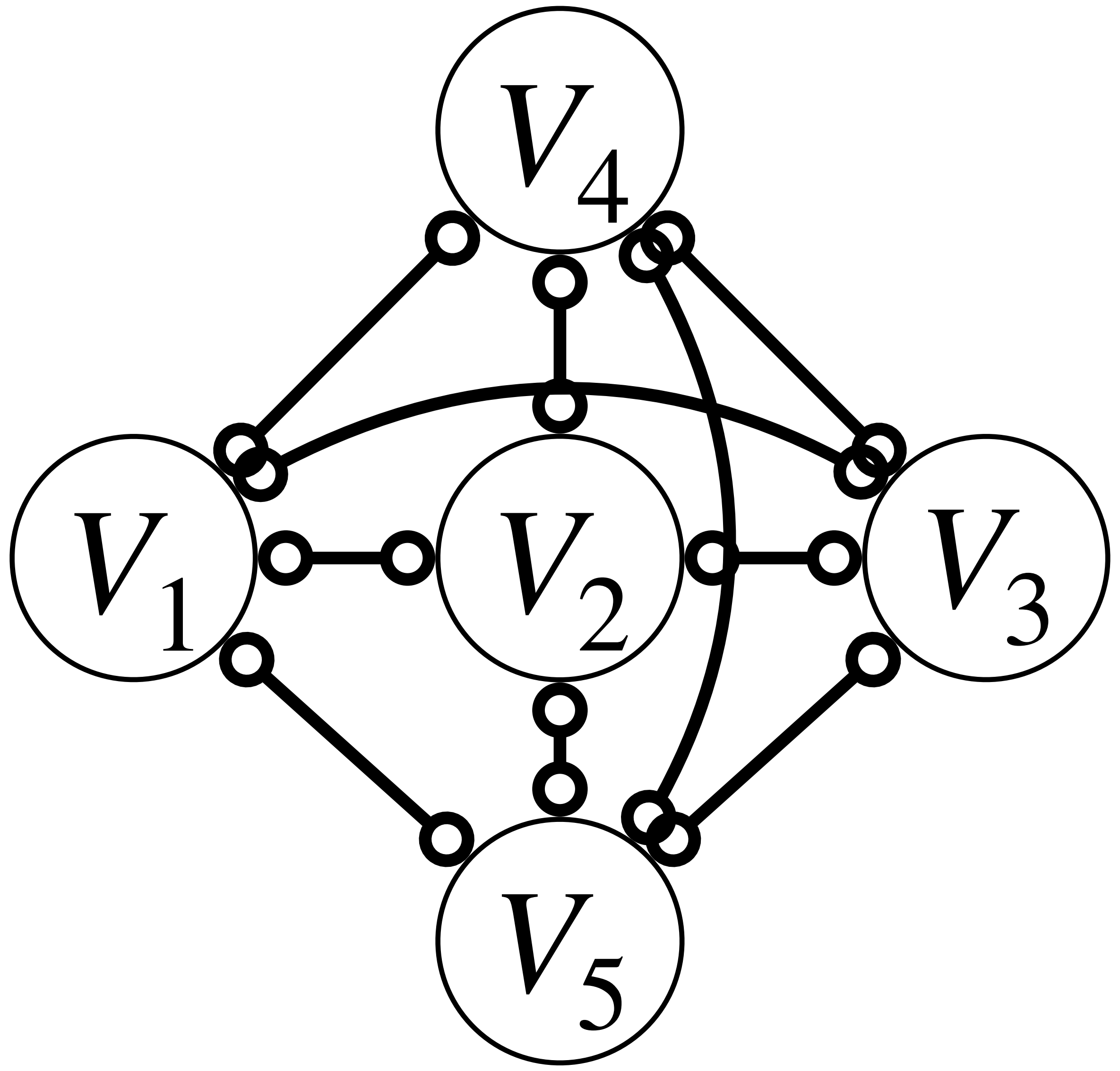


FCI - Skeleton

Form a complete graph on the set of variables, in which there is a circle-circle edge between every pair of variables;

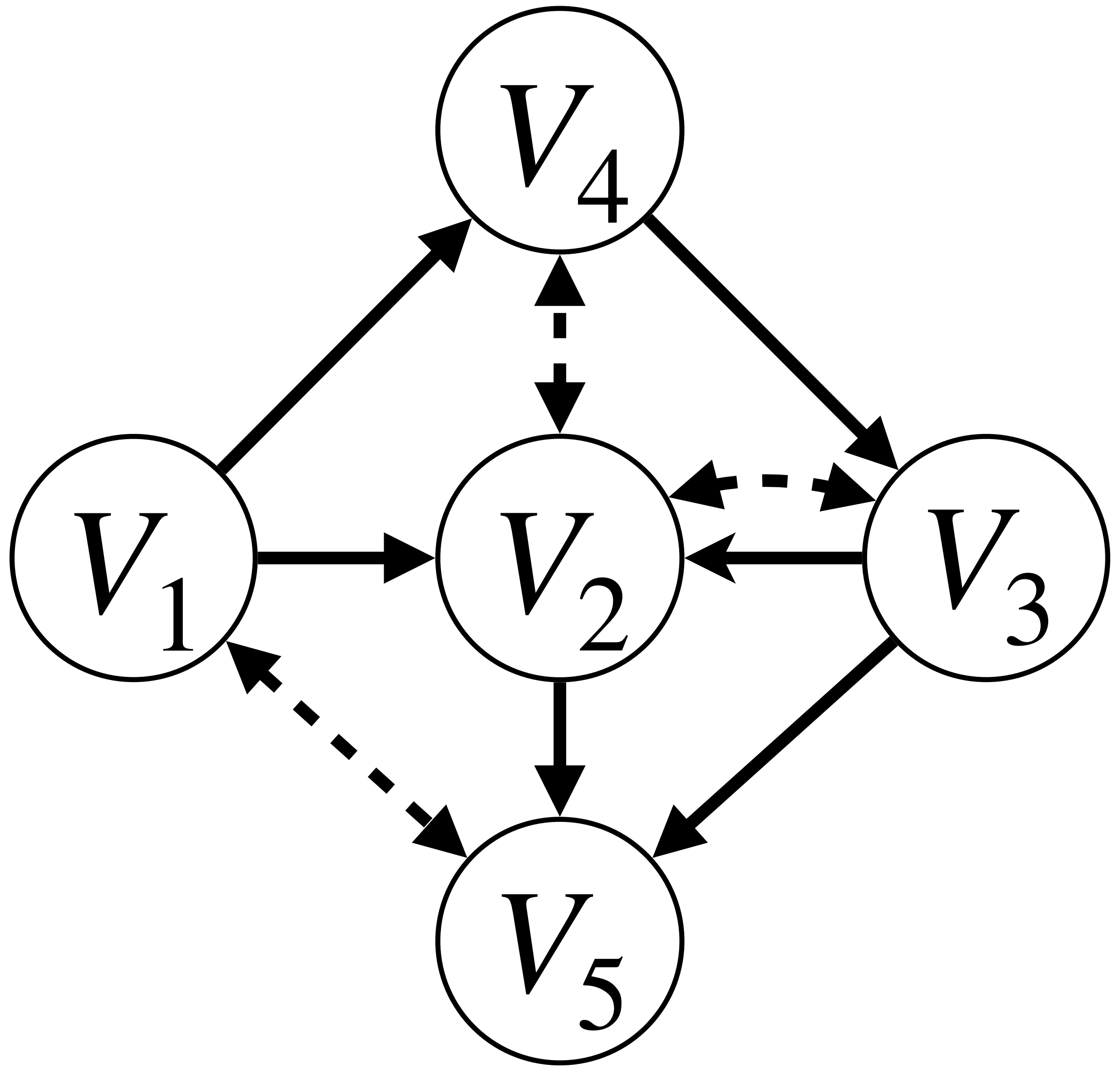


True, unknown ADMG

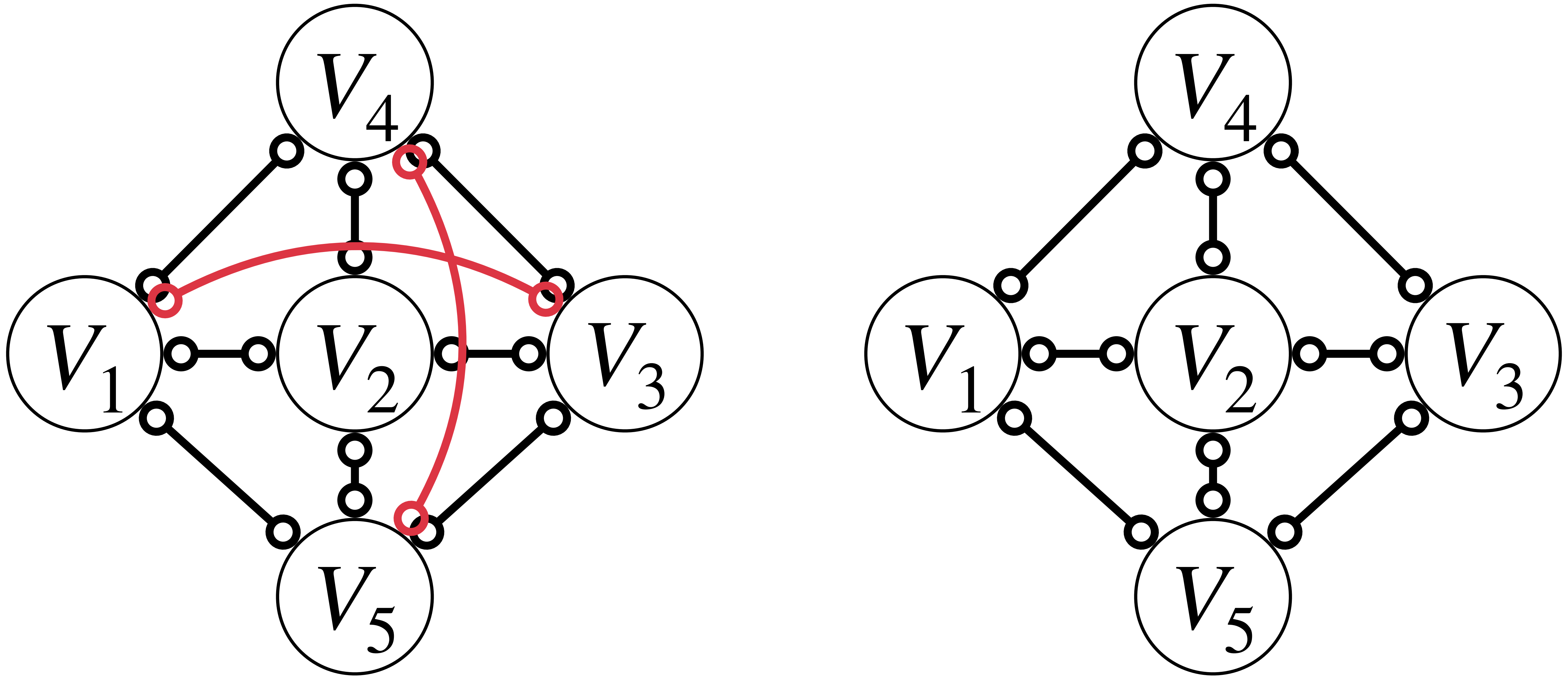


FCI - Skeleton

For every pair of variables V_1 and V_2, if exists a set $\mathbf{S}_{1,2}$ such that $V_1 \perp\!\!\!\perp V_2 \mid \mathbf{S}_{1,2}$, then remove the edge between V_1 and V_2 and add $\mathbf{S}_{1,2}$ in *Sepset*(V_1, V_2).



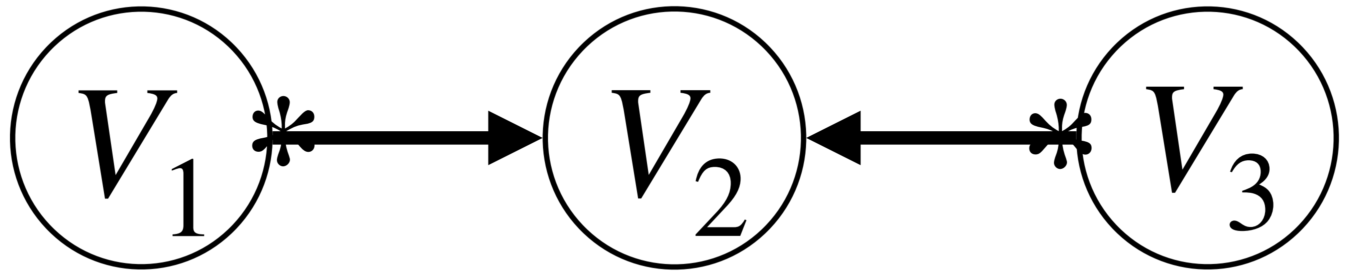
True, unknown ADMG



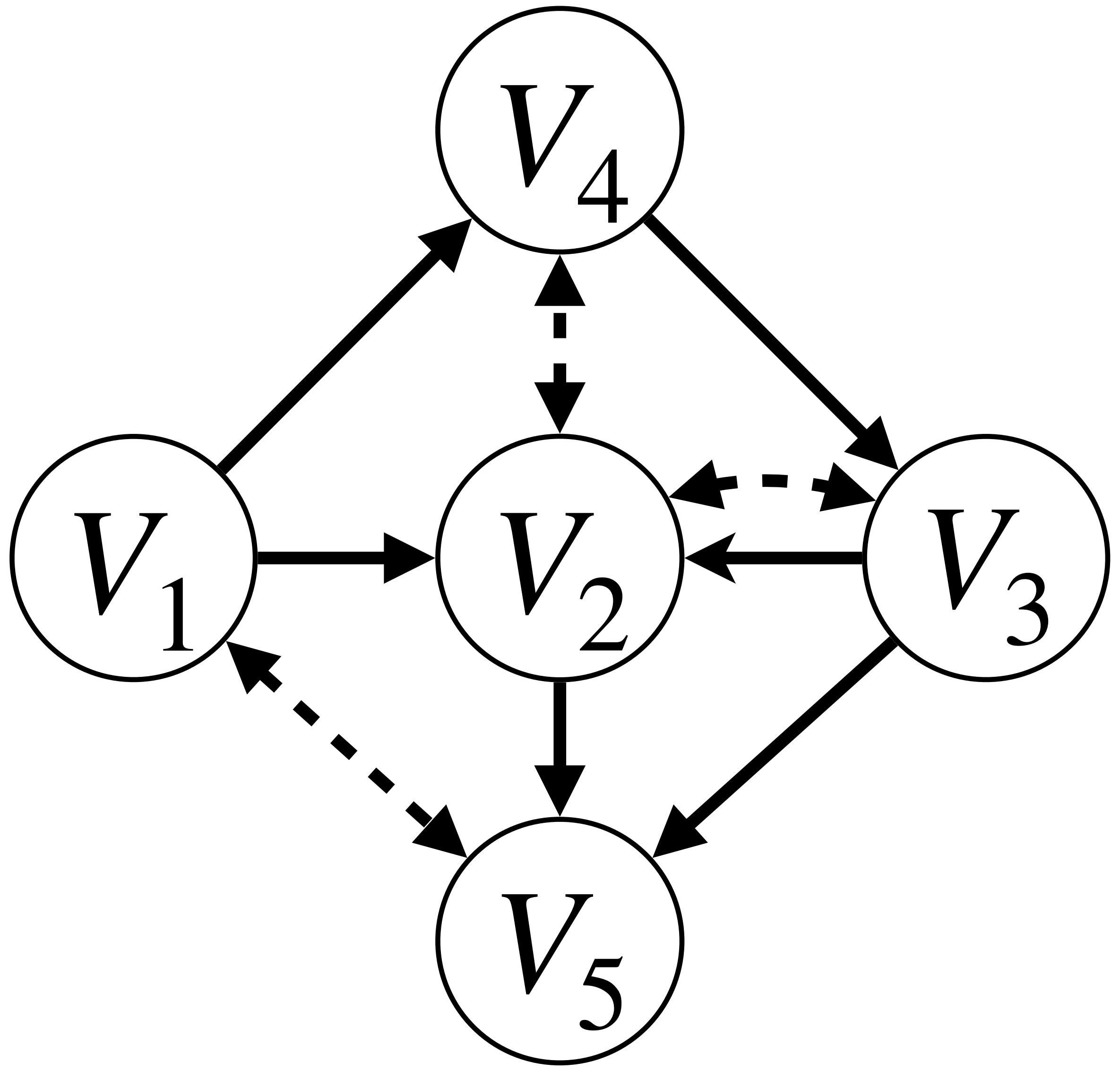
$V_1 \perp\!\!\!\perp V_3 \mid V_4$ and $V_4 \perp\!\!\!\perp V_5 \mid V_1, V_2, V_3$

FCI - Orienting the Colliders

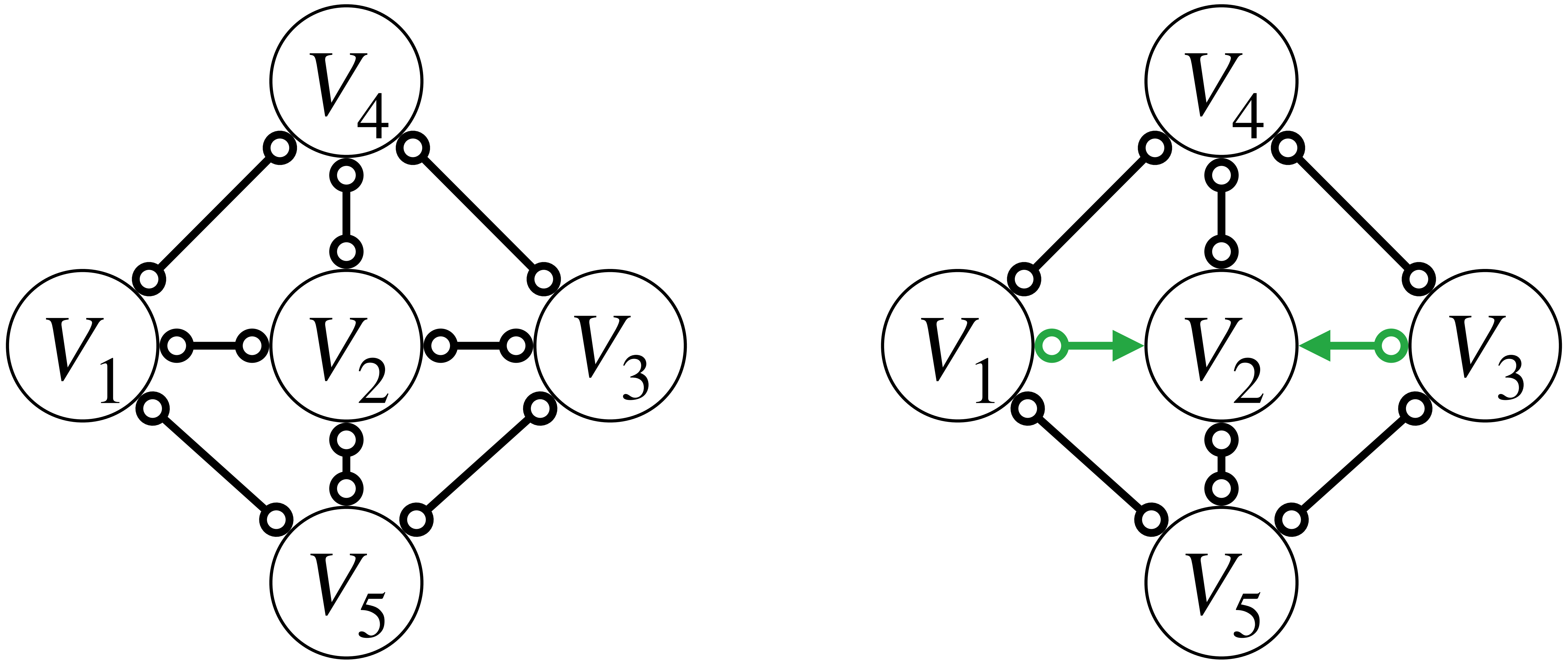
R0: If $\langle V_1, V_2, V_3 \rangle$ is unshielded and $V_2 \notin \textit{Sepset}(V_1, V_3)$, then



That is the only way for the path between V_1 and V_3 to be blocked when not conditioning on V2



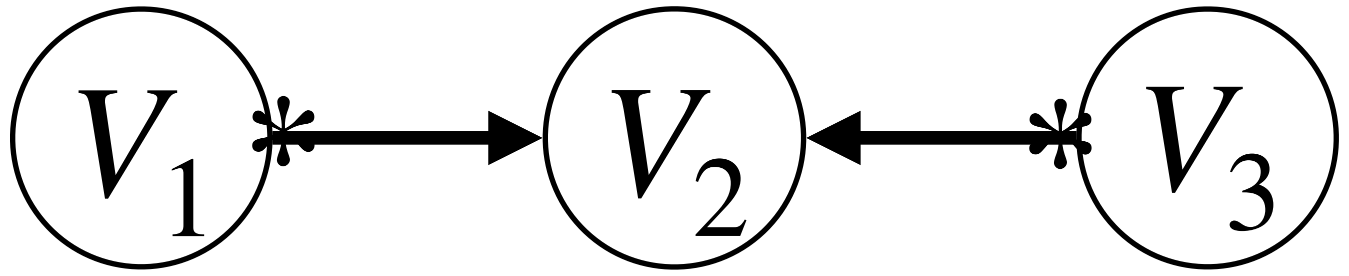
True, unknown ADMG



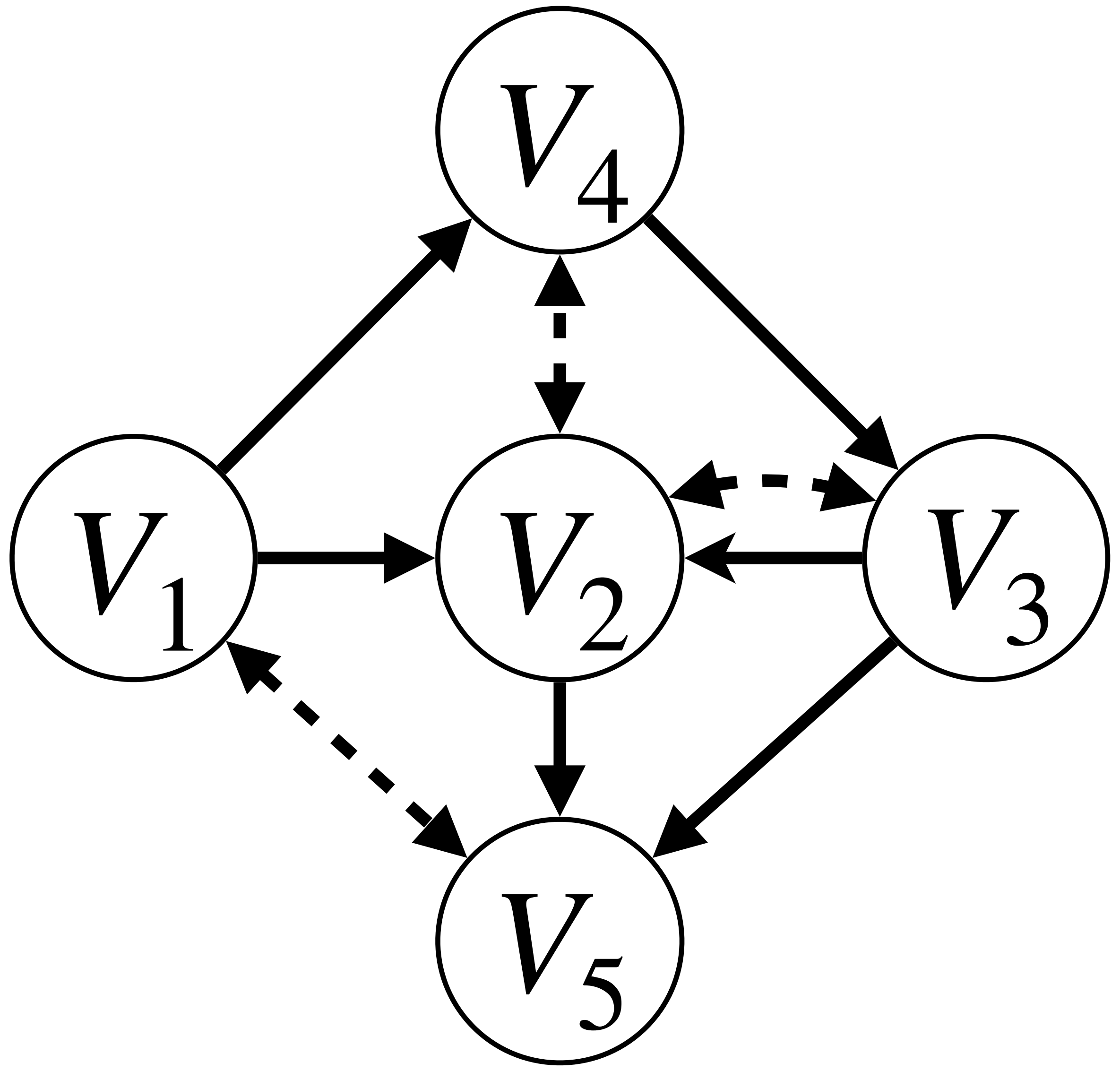
$V_1 \perp\!\!\!\perp V_3 \mid V_4$ and $V_1 \not\!\perp\!\!\!\perp V_3 \mid V_4, V_2$

FCI - Orienting the Colliders

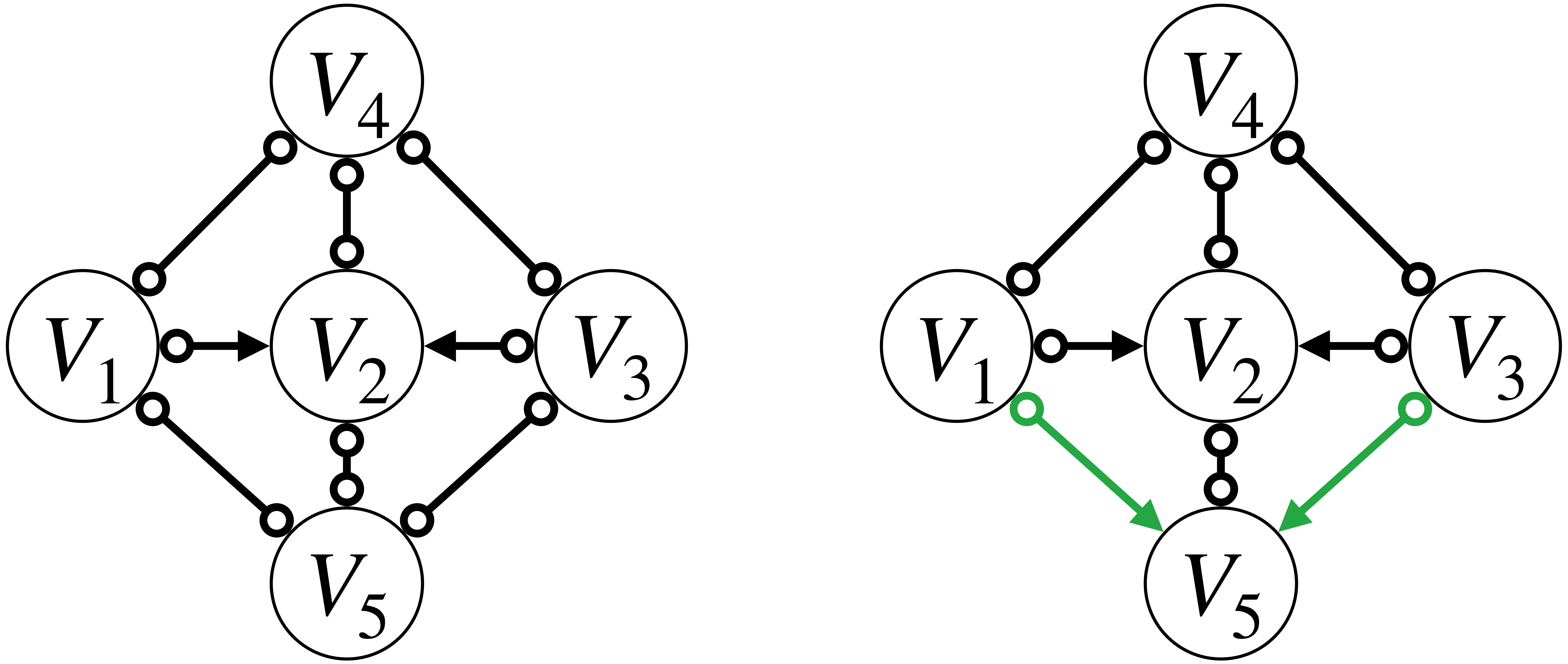
R0: If $\langle V_1, V_2, V_3 \rangle$ is unshielded and $V_2 \notin Sepset(V_1, V_3)$, then



We apply R0 until no more collider can be oriented!



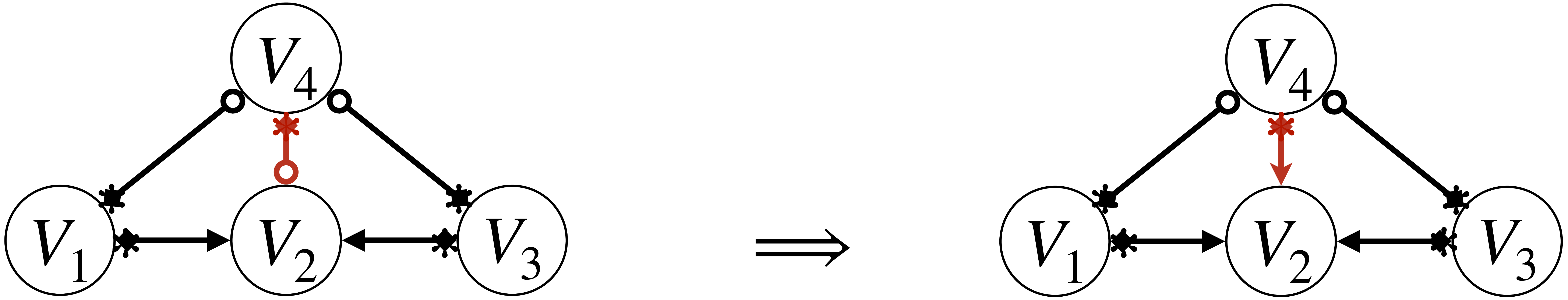
True, unknown ADMG



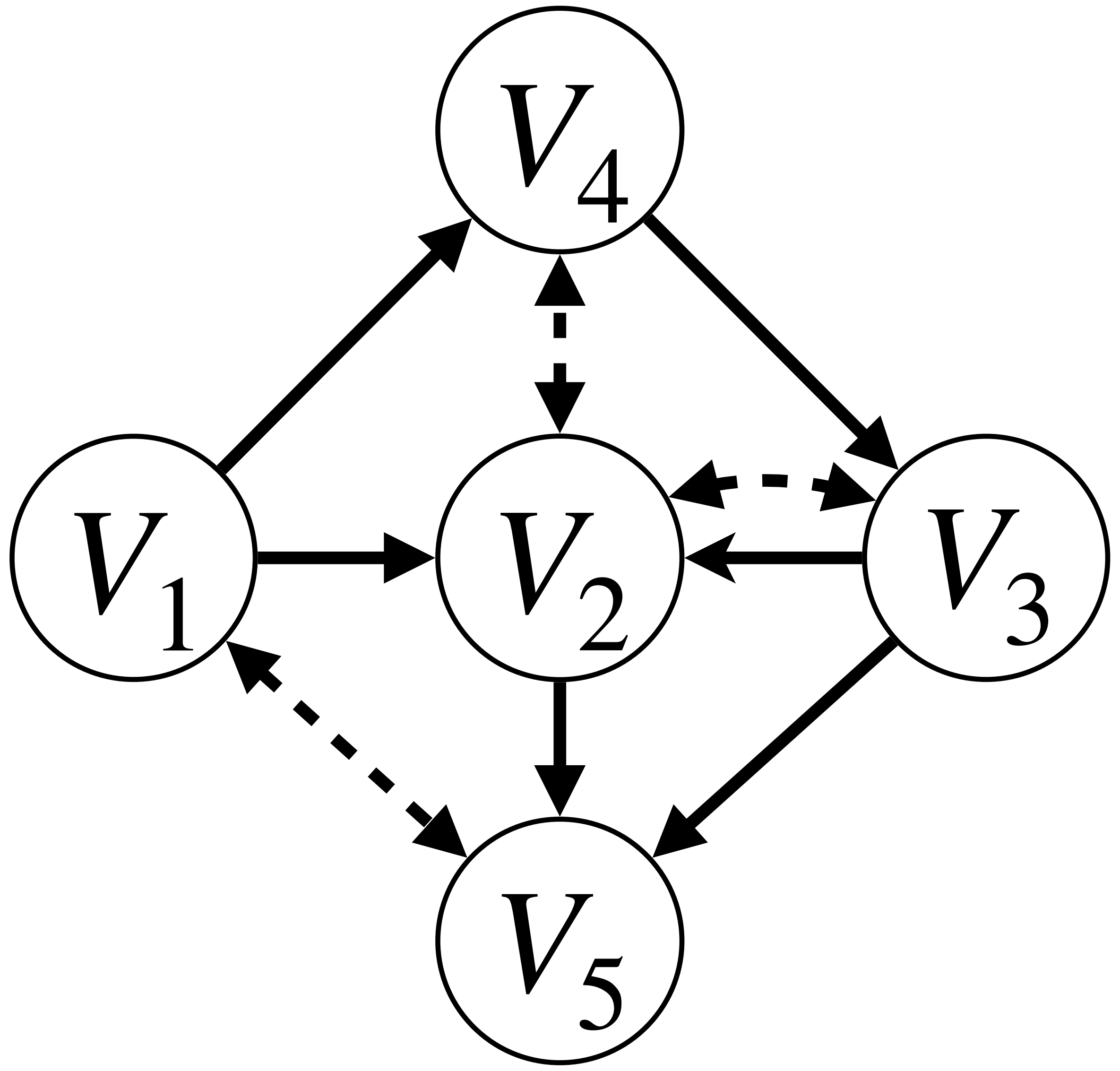
$V_1 \perp\!\!\!\perp V_3 \mid V_4$ and $V_1 \not\perp\!\!\!\perp V_3 \mid V_4, V_5$

Applying Mark Inference Rules

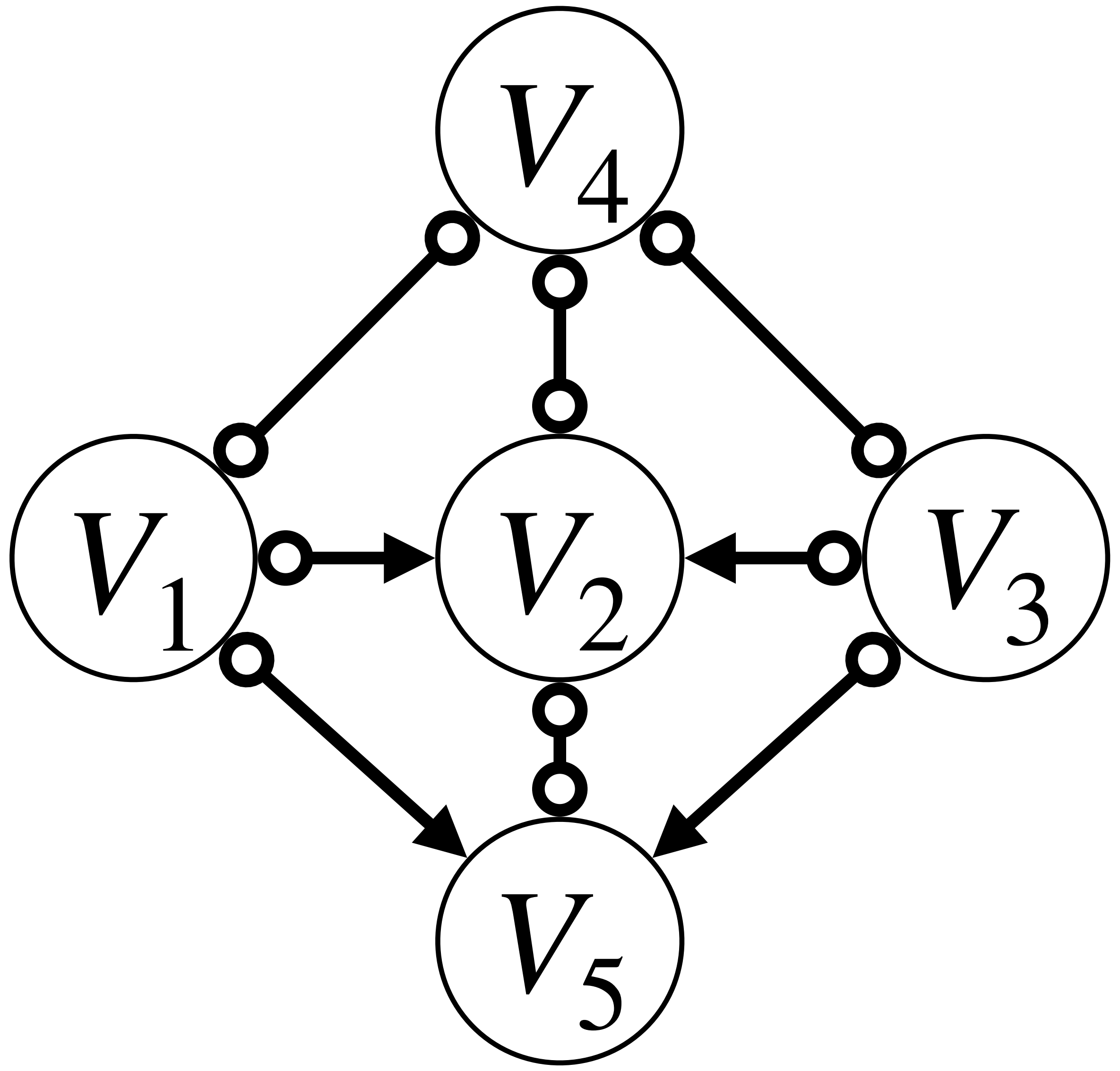
R3:



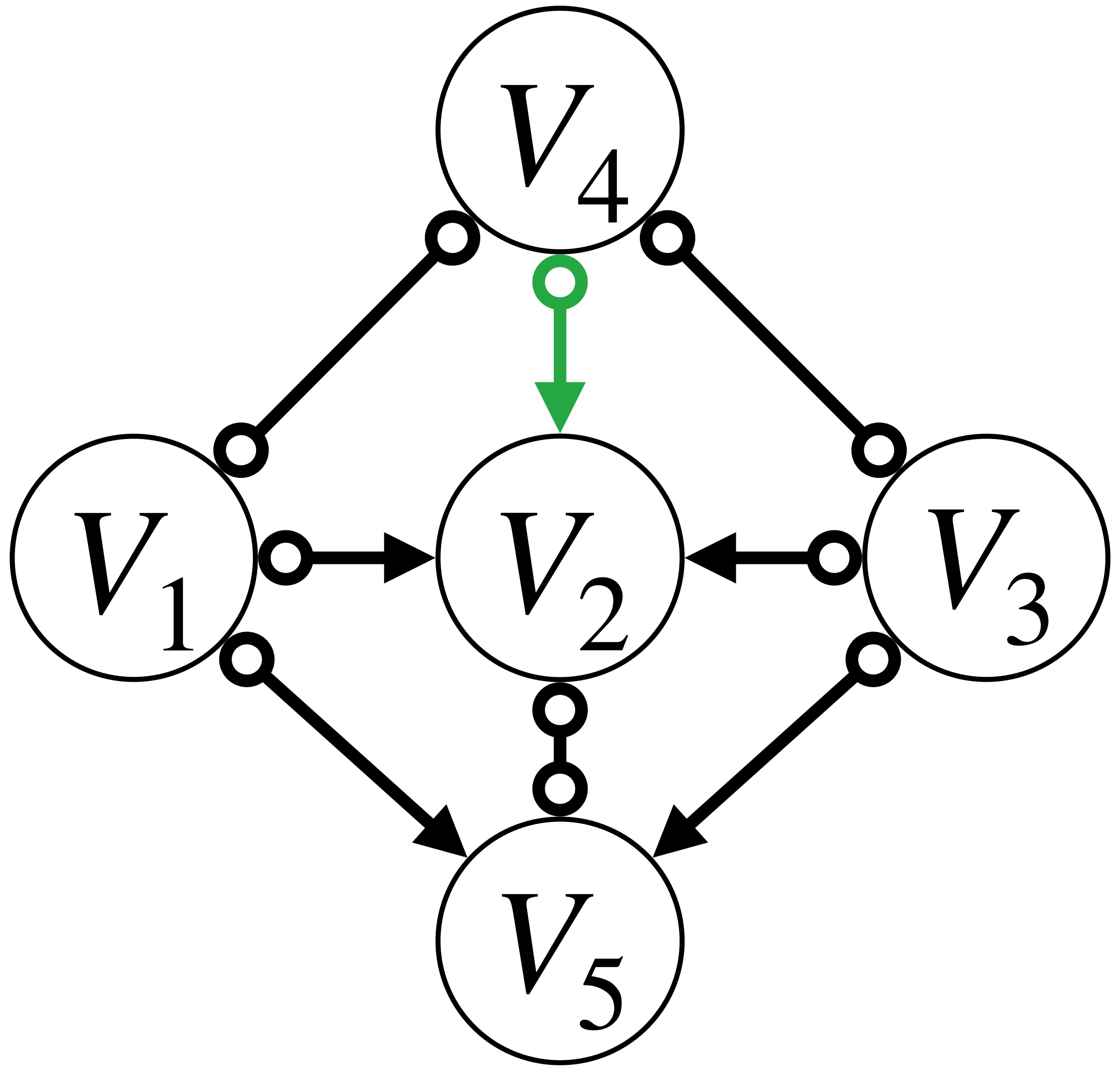
where V_1 and V_3 are not adjacent



True, unknown ADMG



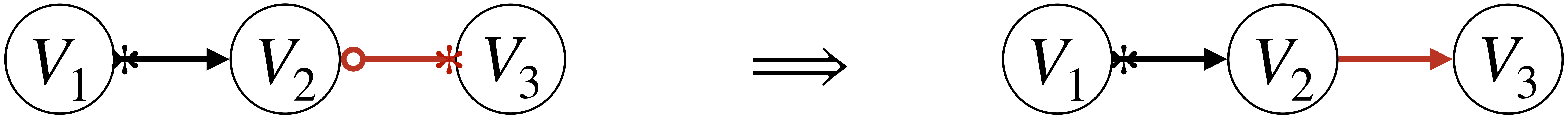
After Skeleton + R0



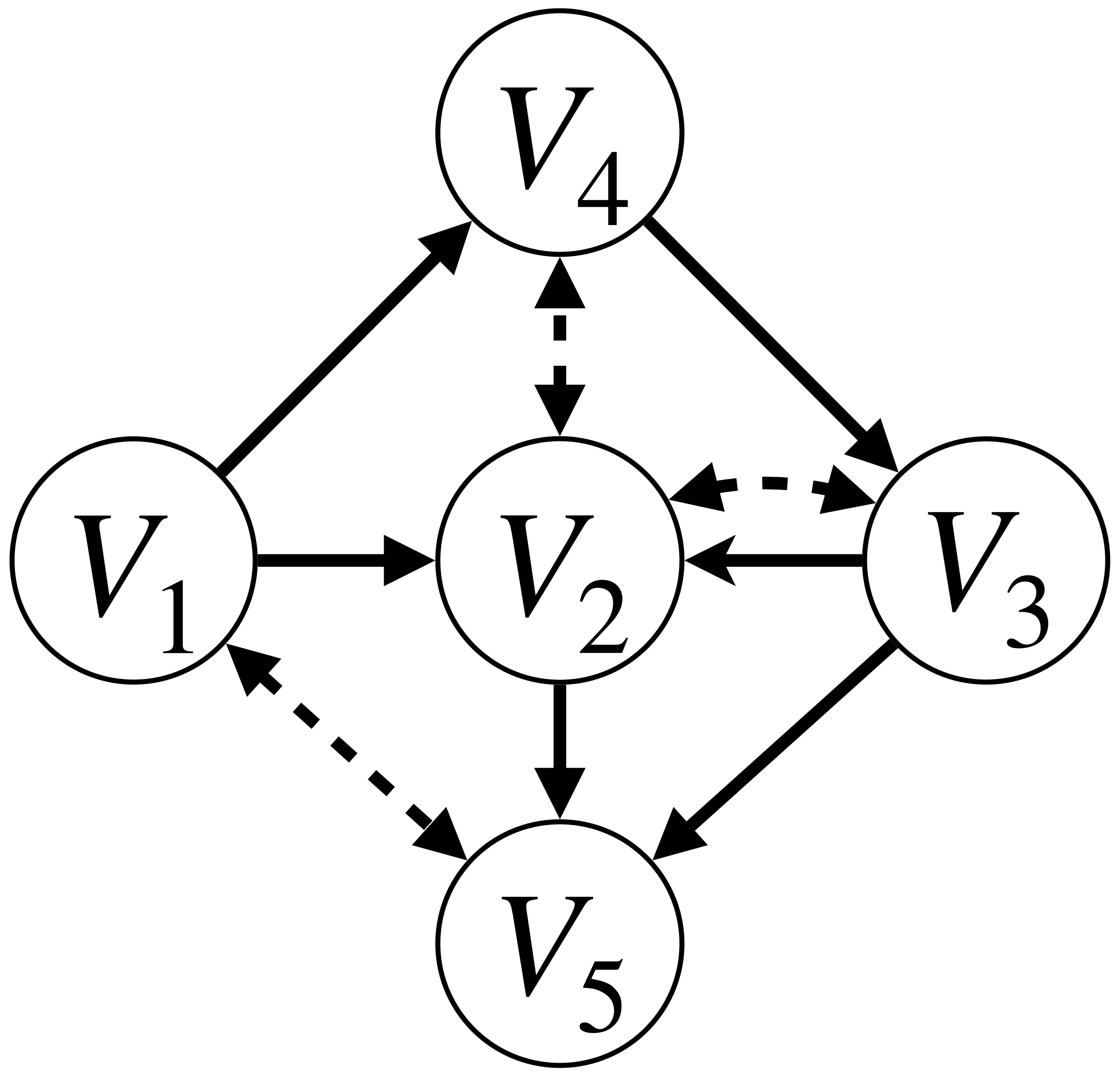
Applying R3

Applying Mark Inference Rules

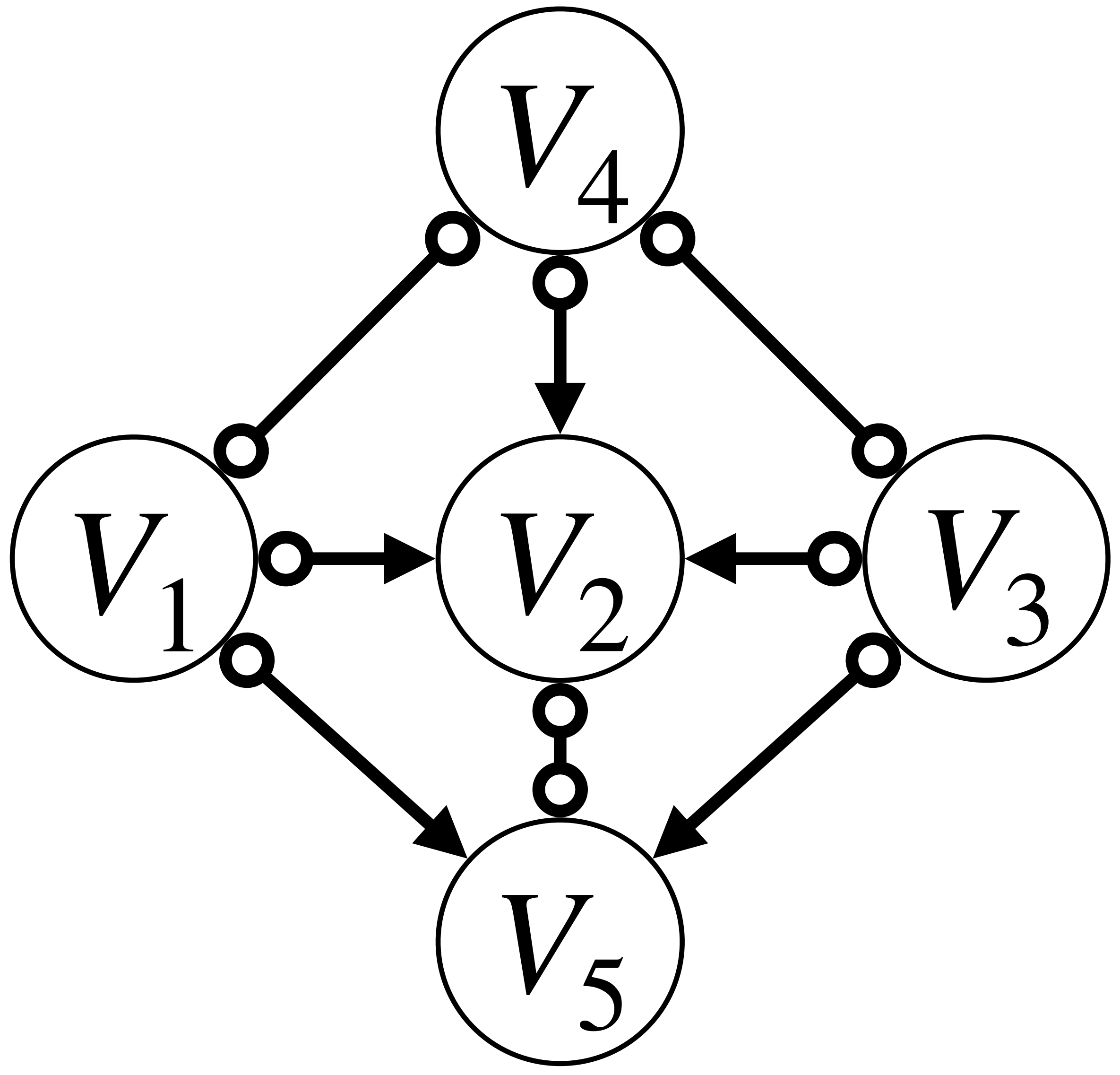
R1:



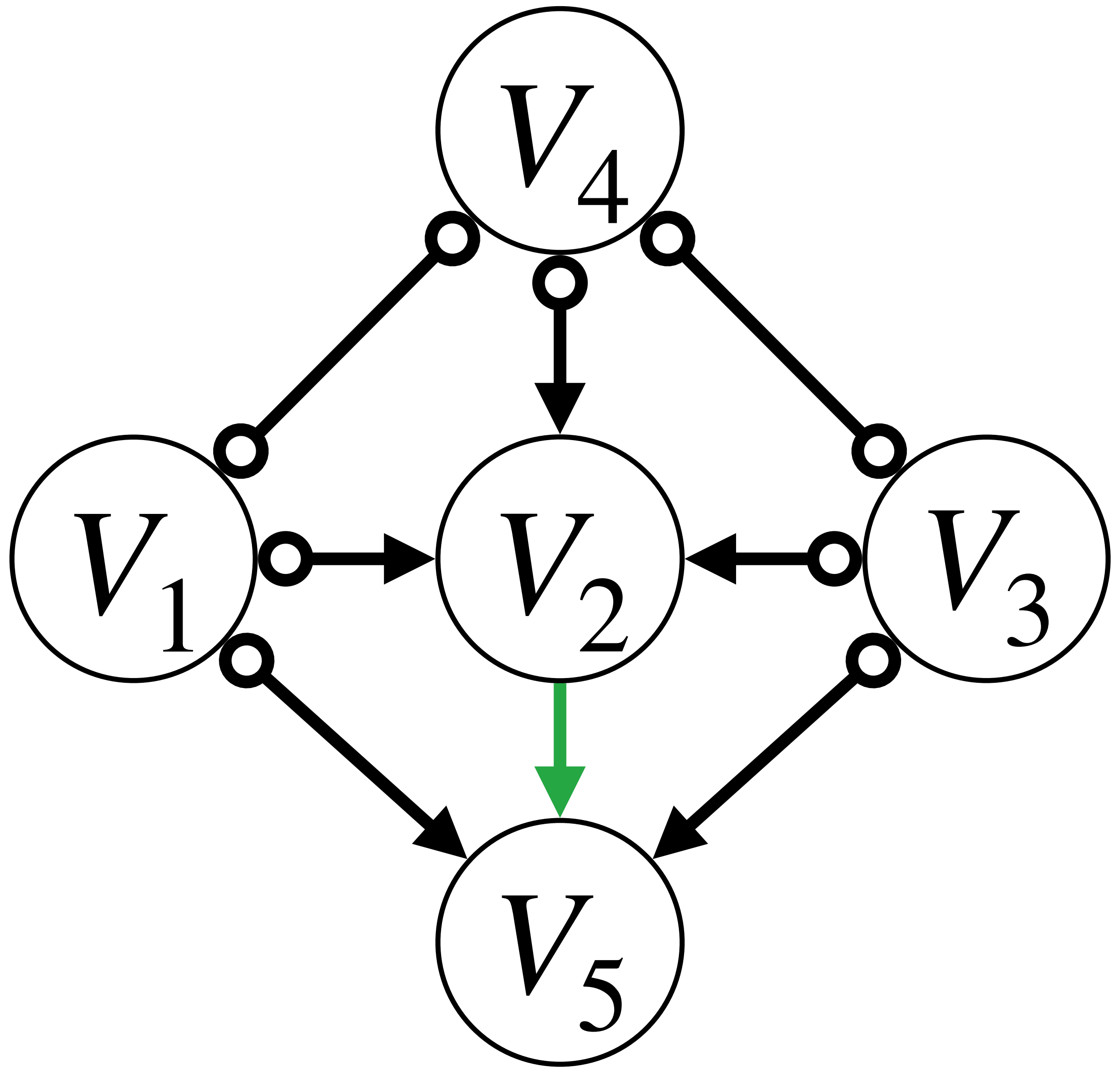
where V_1 and V_3 are not adjacent



True, unknown ADMG



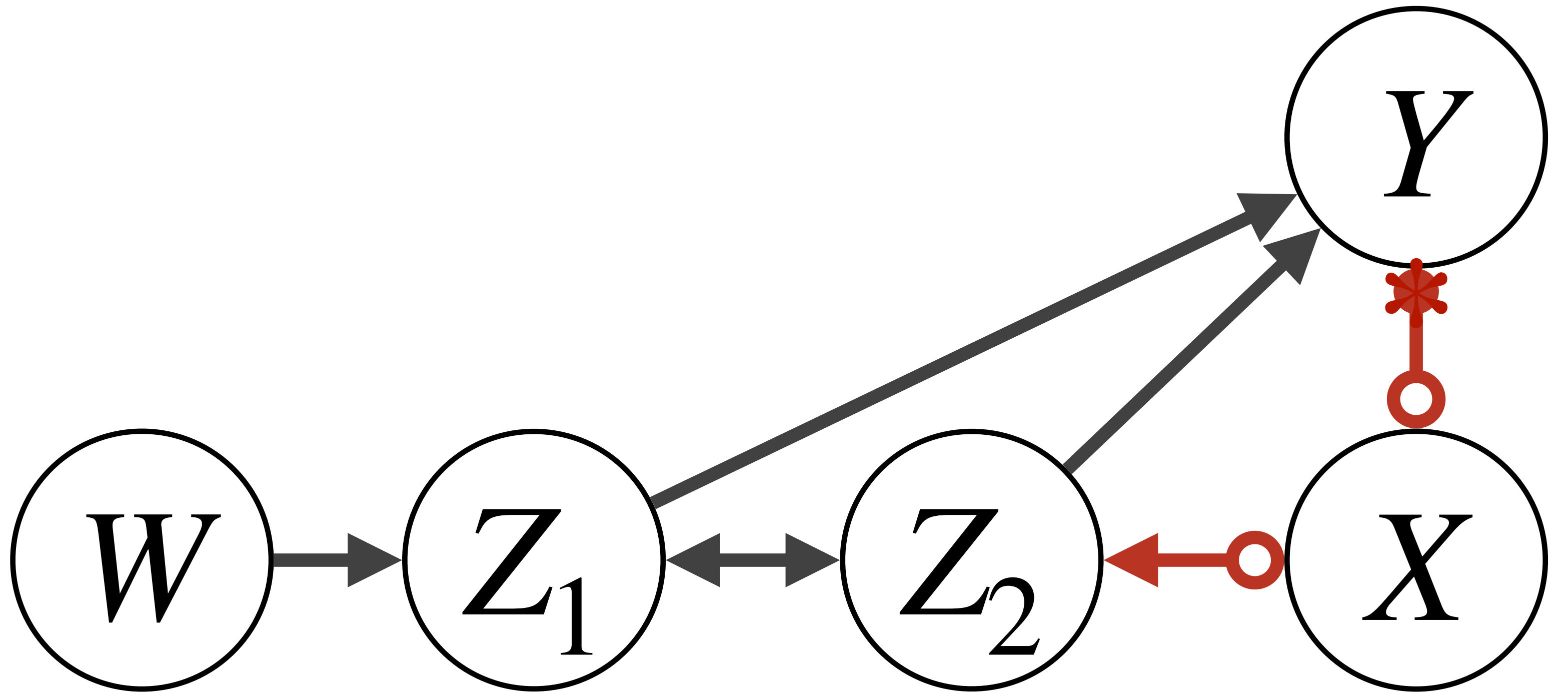
After Skel + R0 + R3



Applying R1

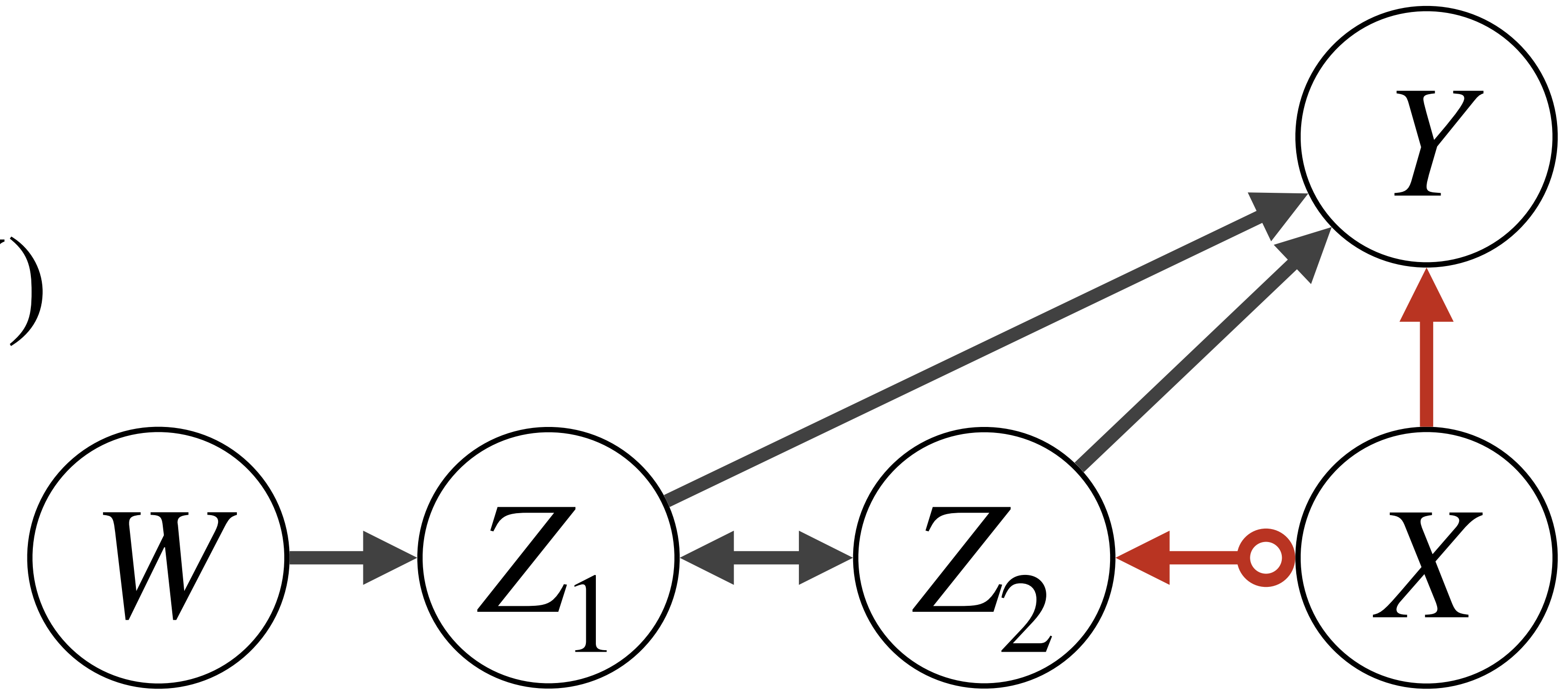
Applying Mark Inference Rules

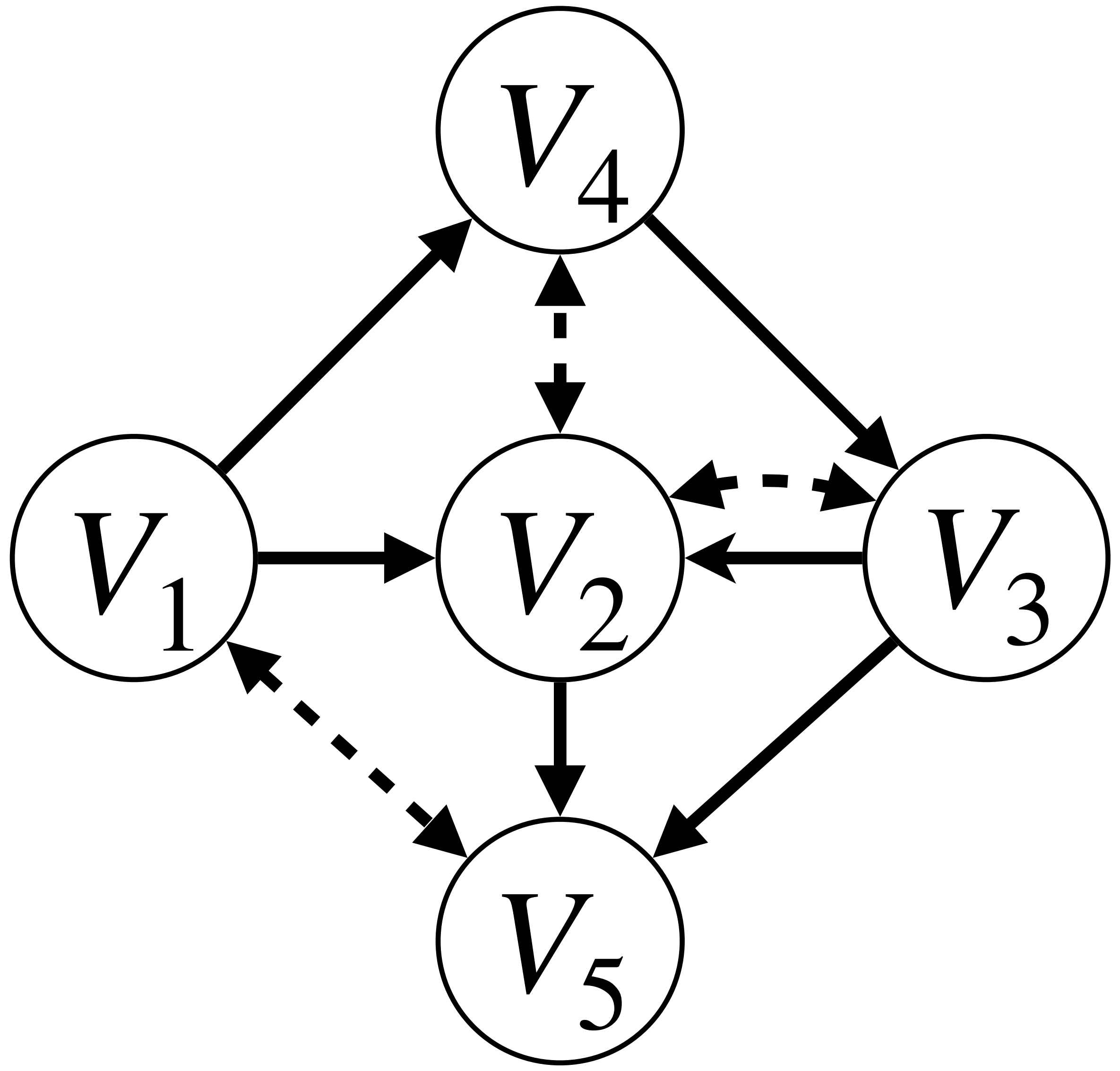
R4:



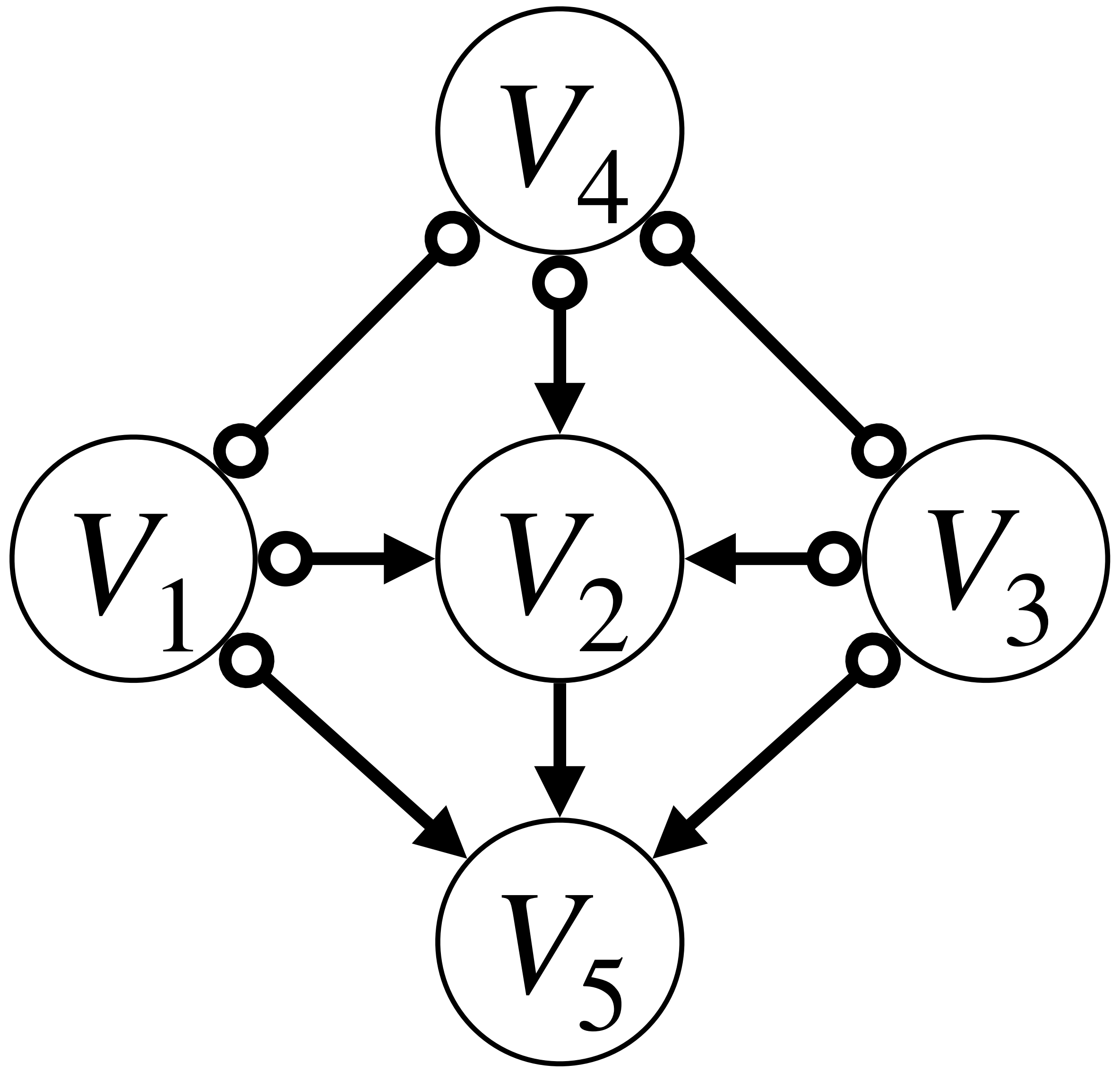
$\langle W, Z_1, Z_2, X, Y \rangle$ is a discriminating path for X

$X \in Sepset(W, Y)$ $\Longrightarrow$





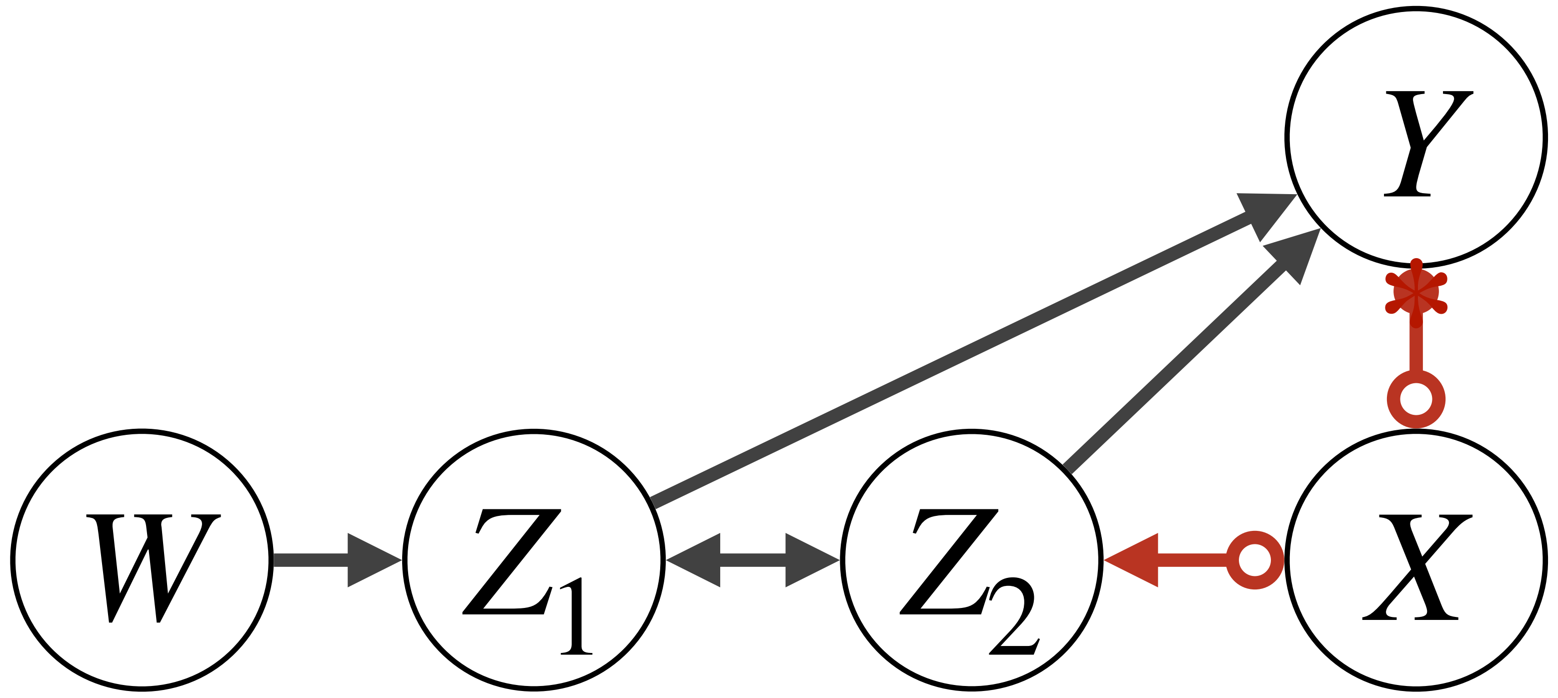
True, unknown ADMG



After Skel + R0 + R3 + R1

Applying Mark Inference Rules

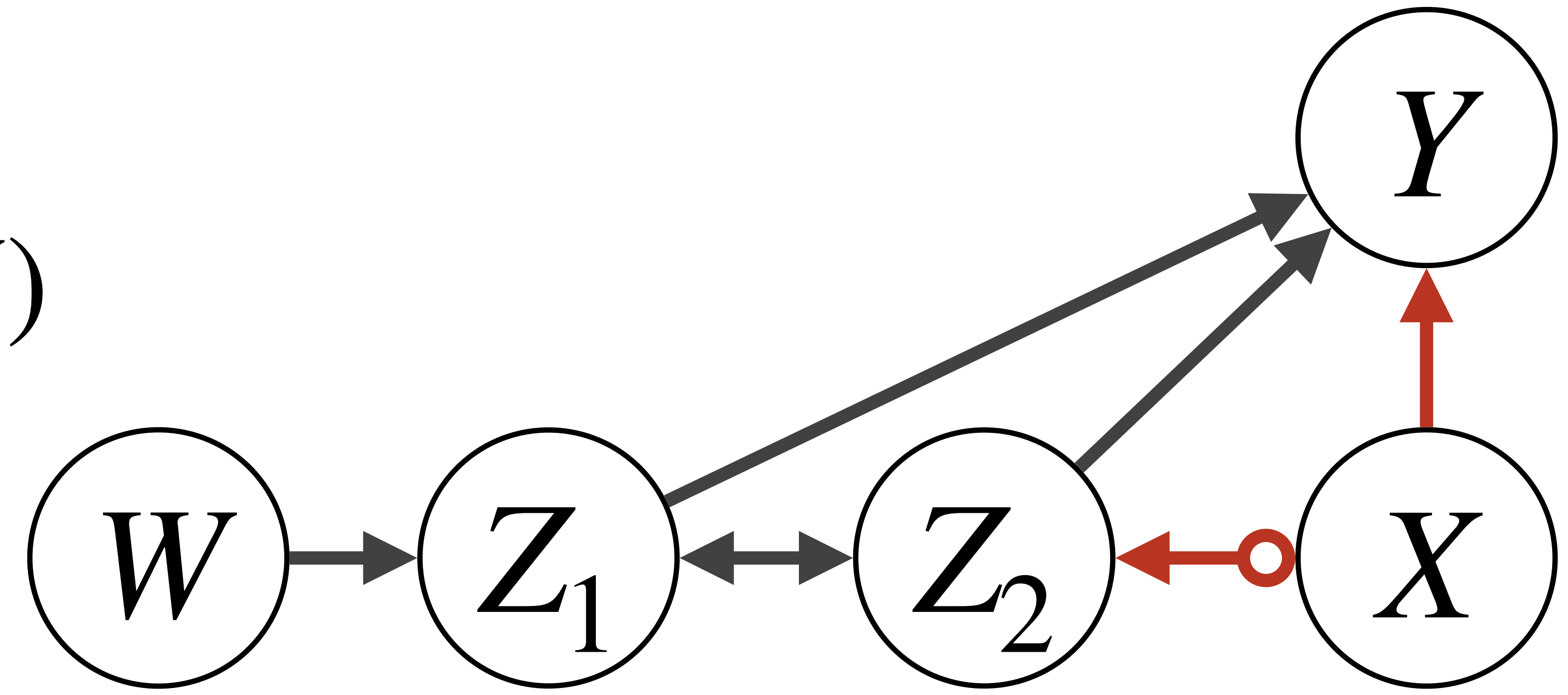
R4:

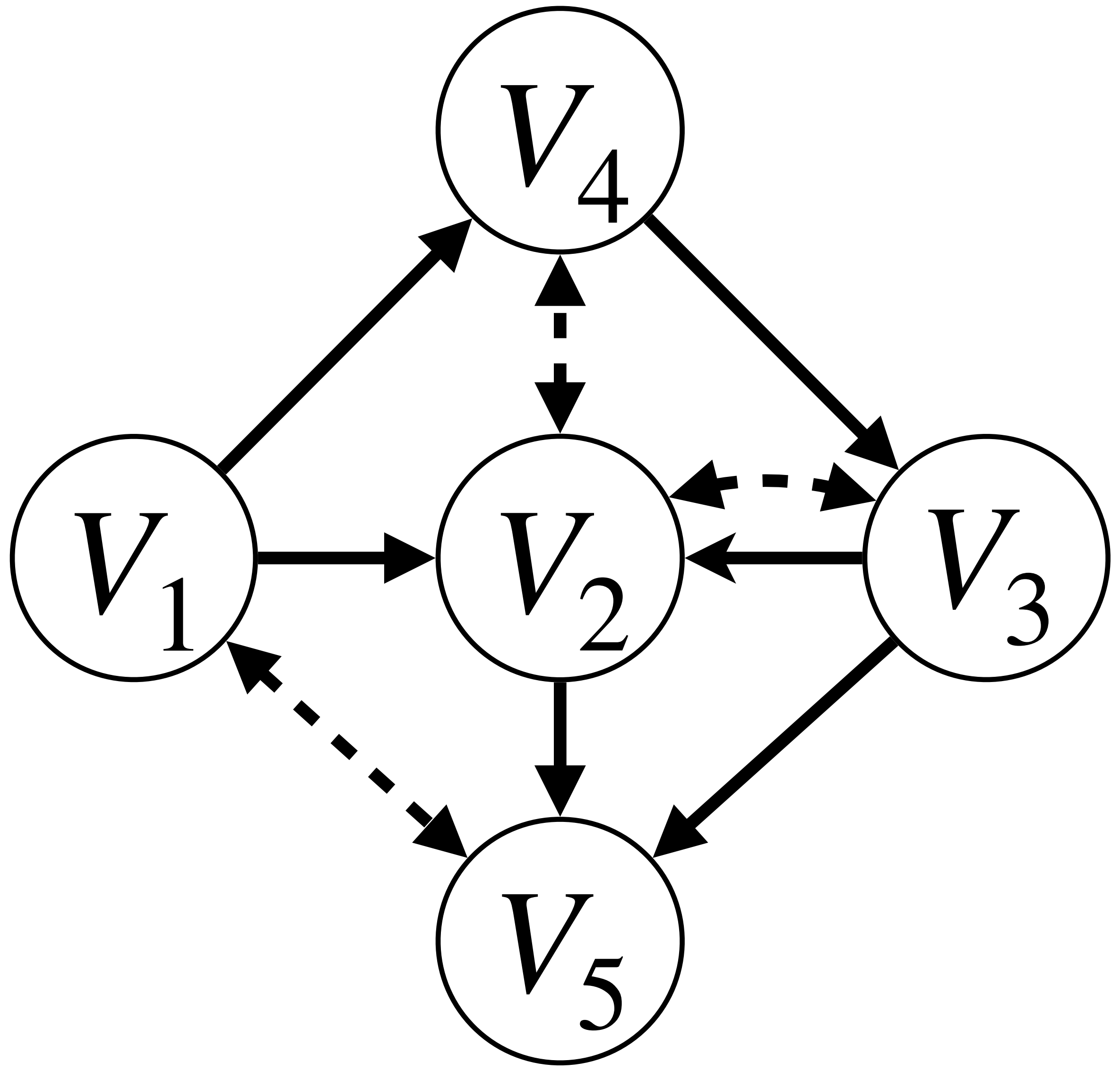


$\langle W, Z_1, Z_2, X, Y \rangle$ is a discriminating path for X

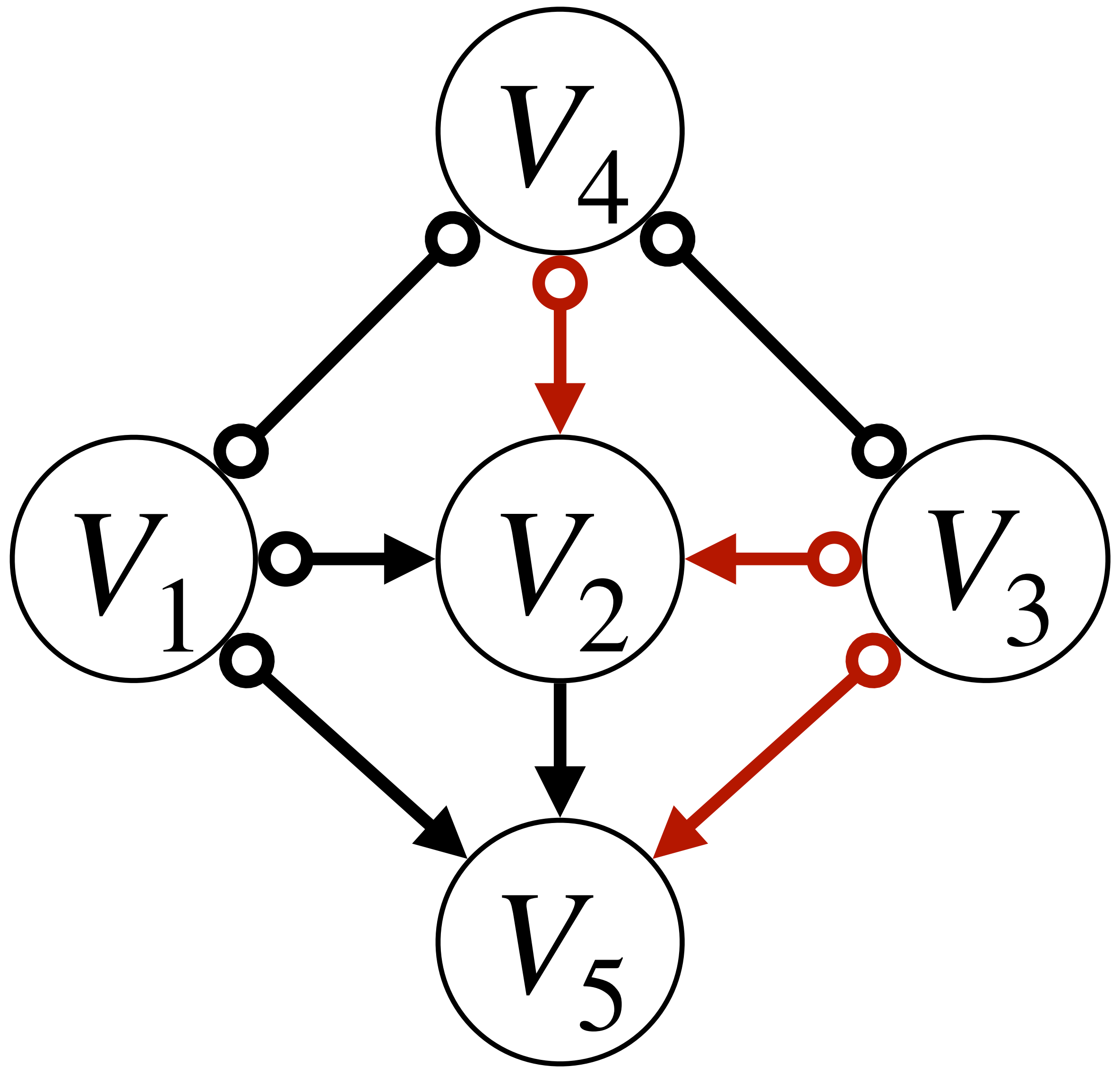
$X \in Sepset(W, Y)$

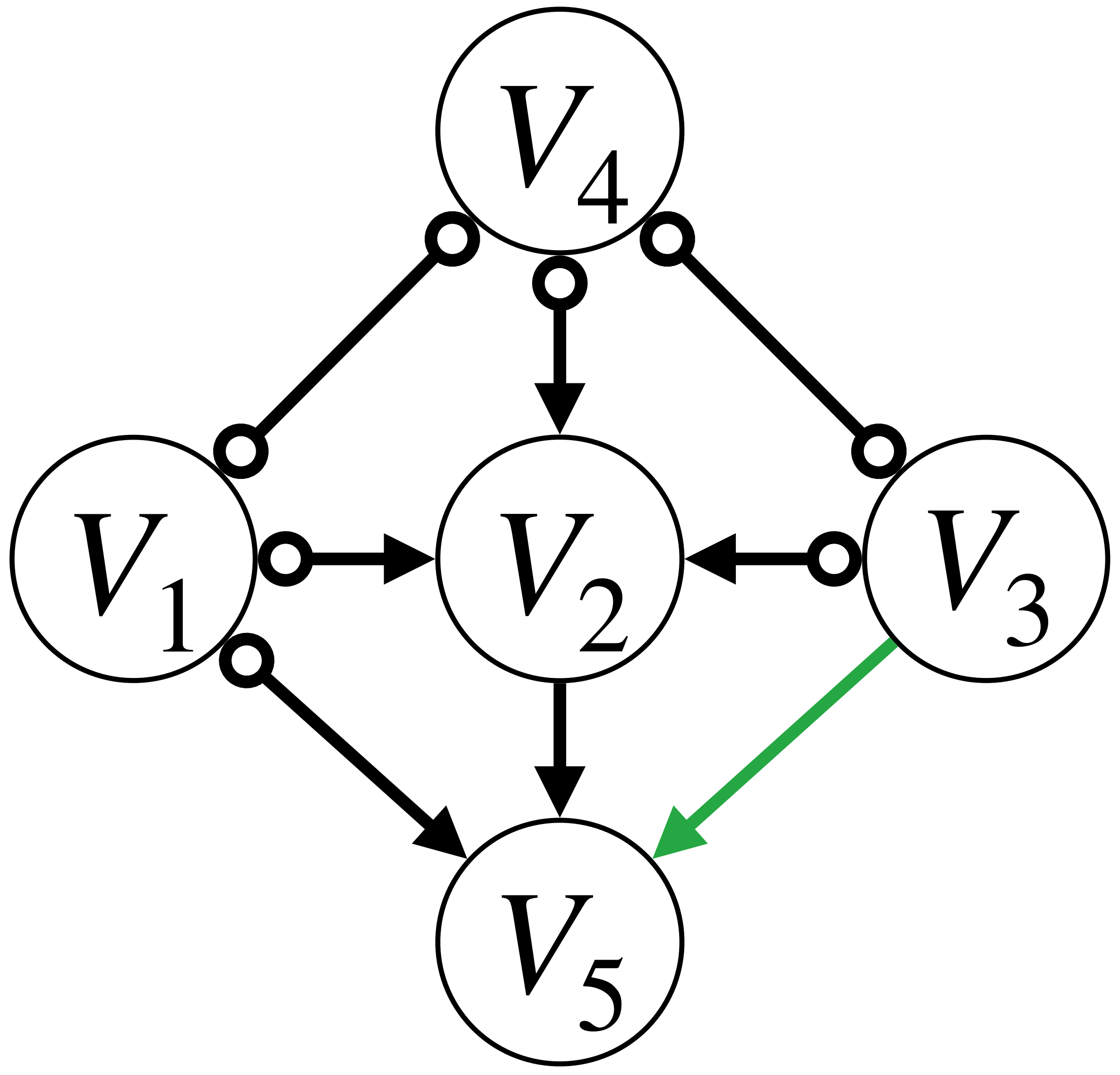
$\Longrightarrow$





True, unknown ADMG



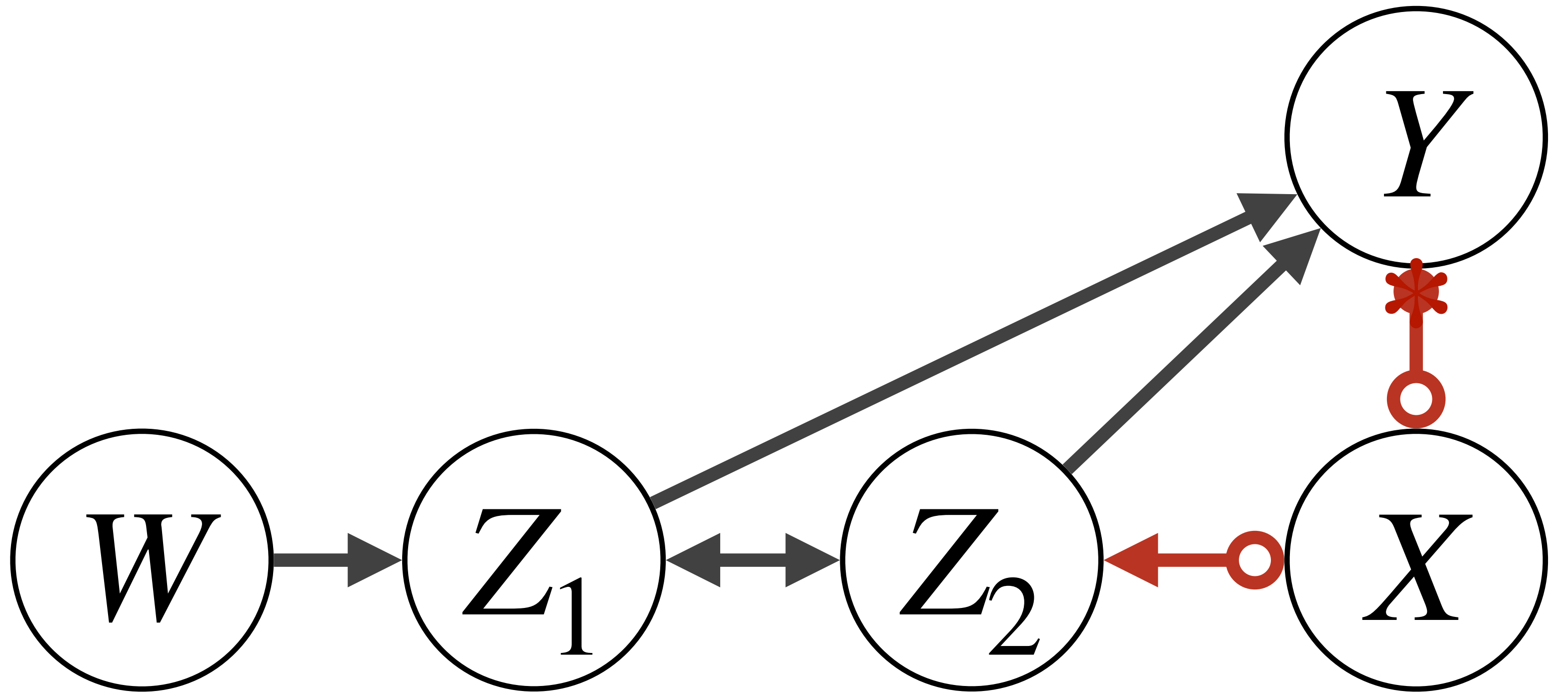


$\langle V_4, V_2, V_3, V_5 \rangle$ is a discriminating path for V_3 and

$V_3 \in Sepset(V_4, V_5)$ $-$ $V_4 \perp\!\!\!\perp V_5 \mid V_1, V_2, V_3$

Applying Mark Inference Rules

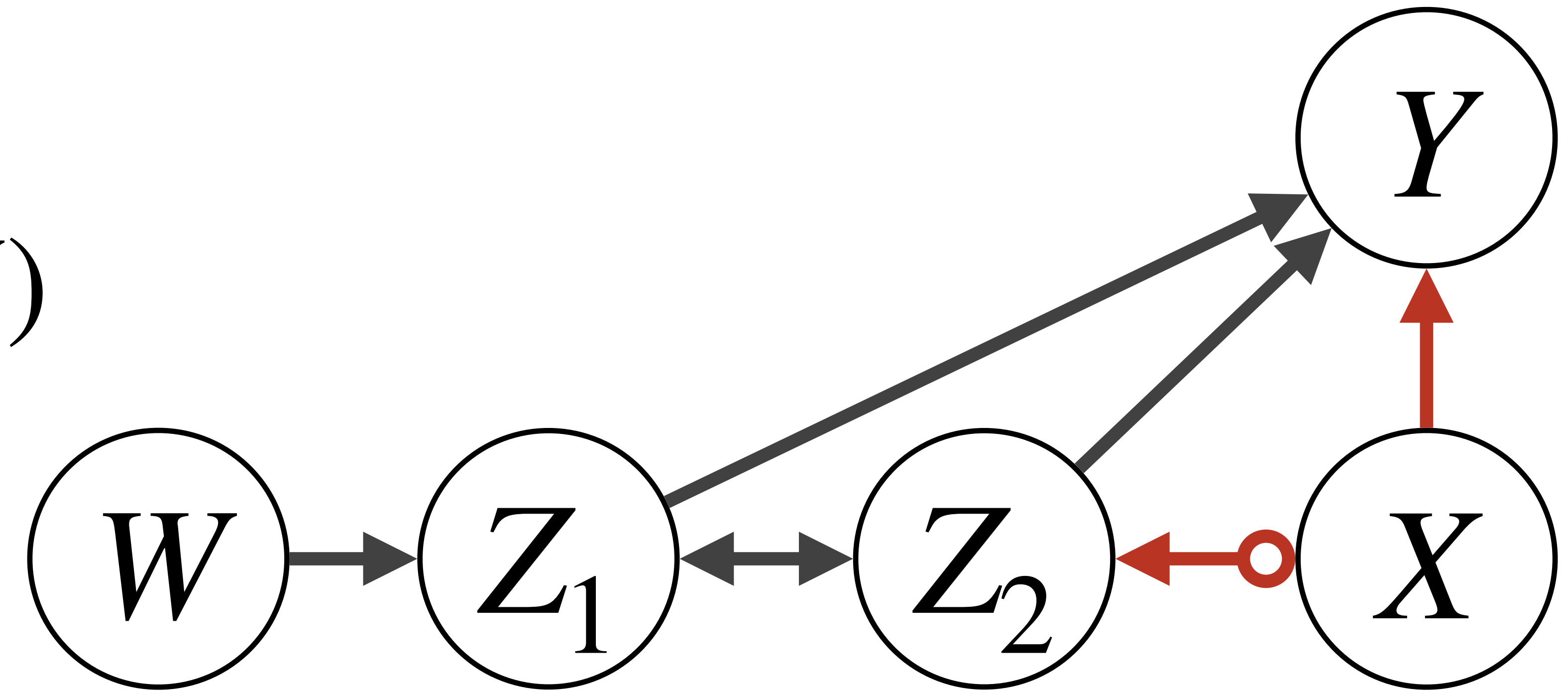
R4:

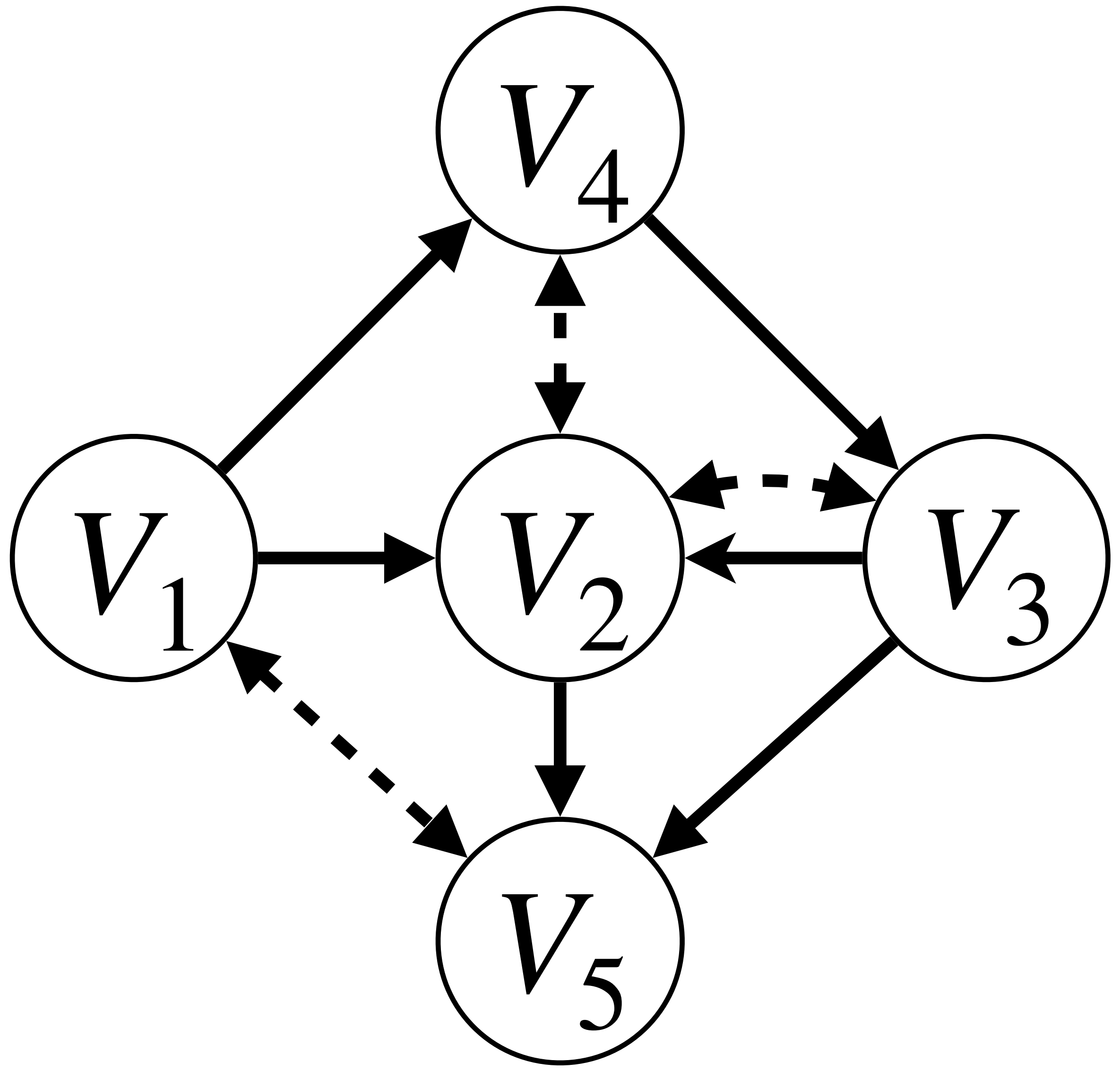


$\langle W, Z_1, Z_2, X, Y \rangle$ is a discriminating path for X

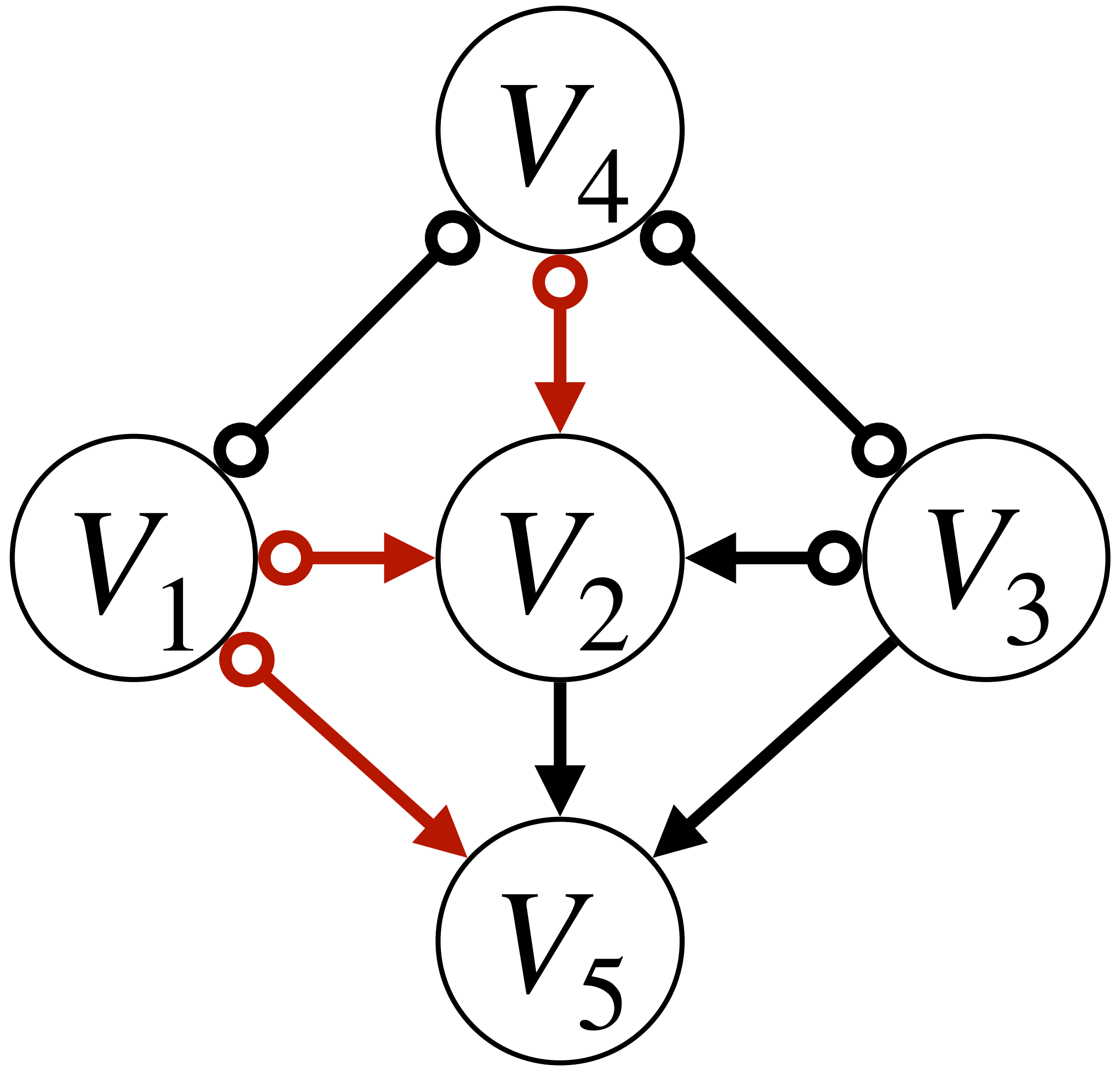
$X \in Sepset(W, Y)$

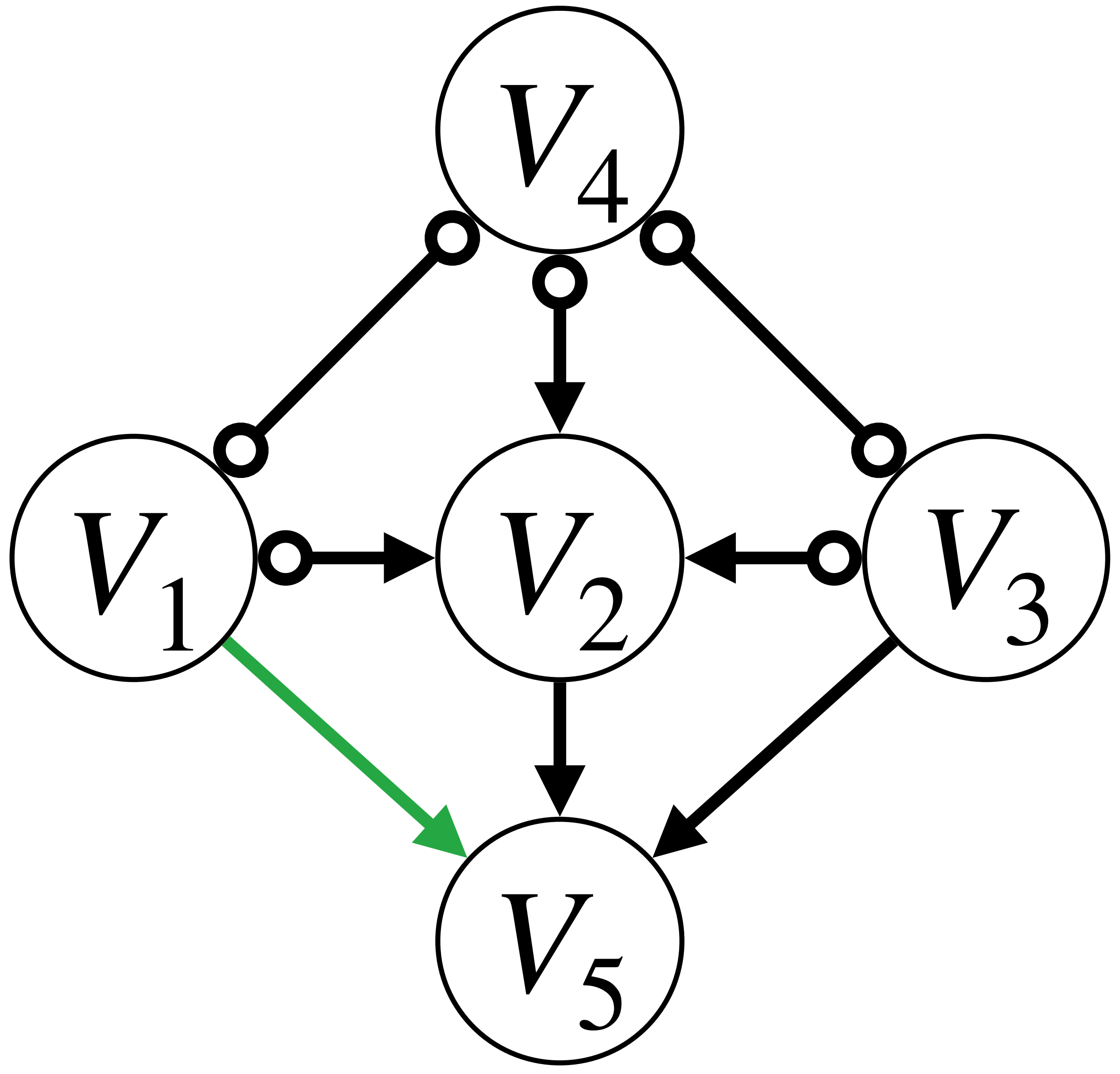
$\Longrightarrow$





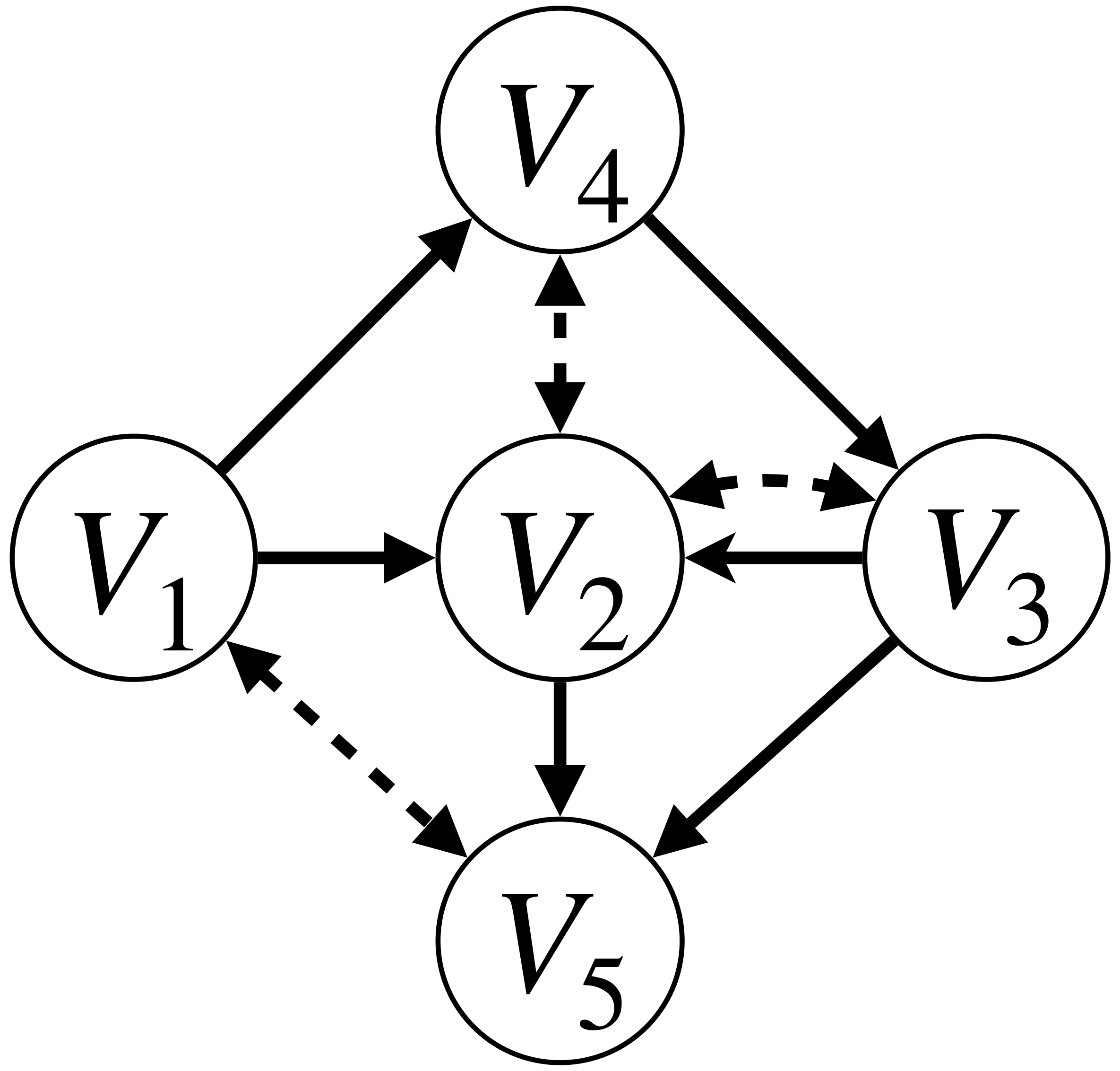
True, unknown ADMG





$\langle V_4, V_2, V_1, V_5 \rangle$ is a discriminating path for V_1 and $V_1 \in Sepset(V_4, V_5)$ – $V_4 \perp\!\!\!\perp V_5 \mid V_1, V_2, V_3$

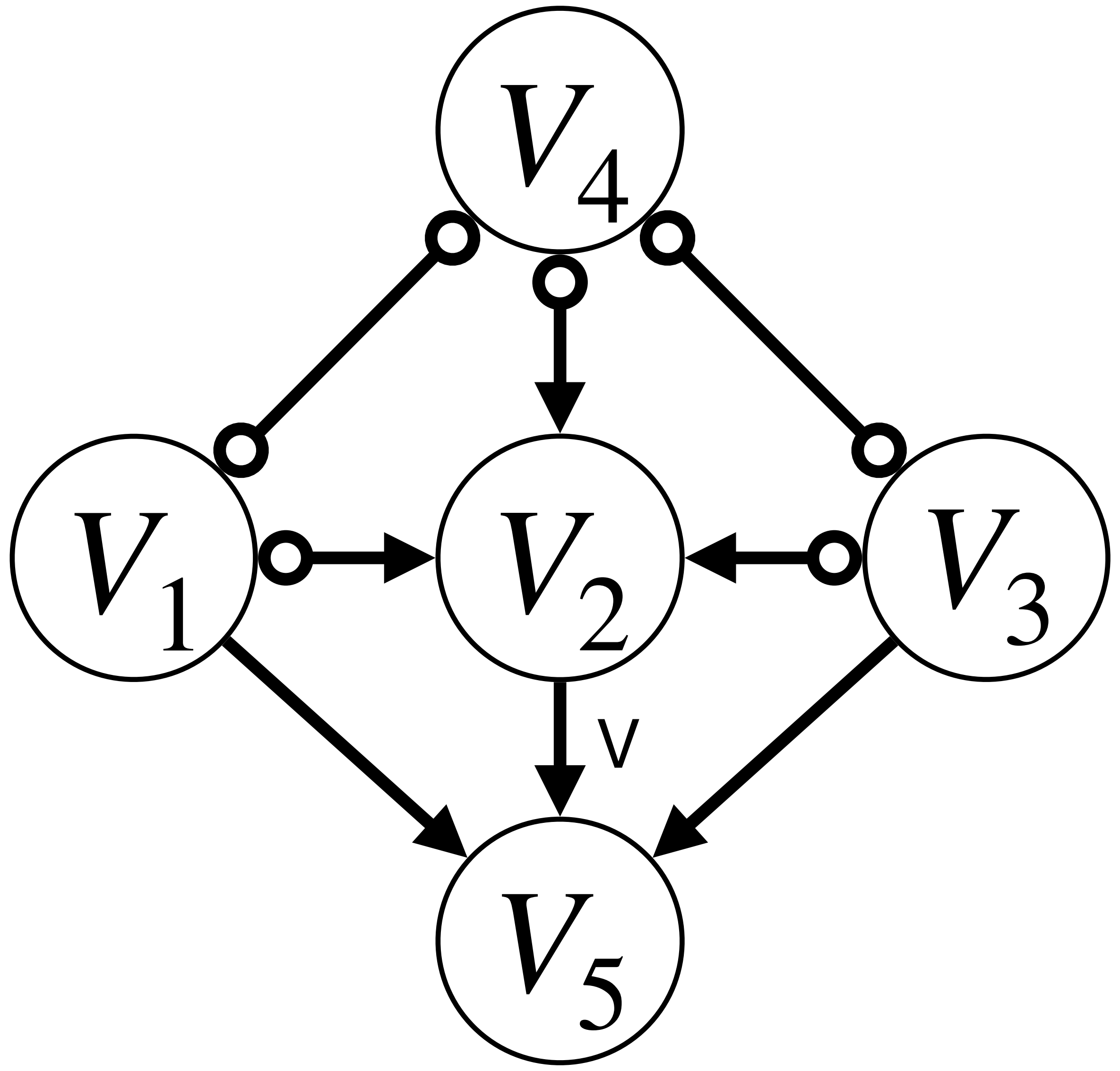
Final PAG



True, unknown ADMG

$V_1 \perp\!\!\!\perp V_3 \mid V_4$

$V_4 \perp\!\!\!\perp V_5 \mid V_1, V_2, V_3$



Final PAG

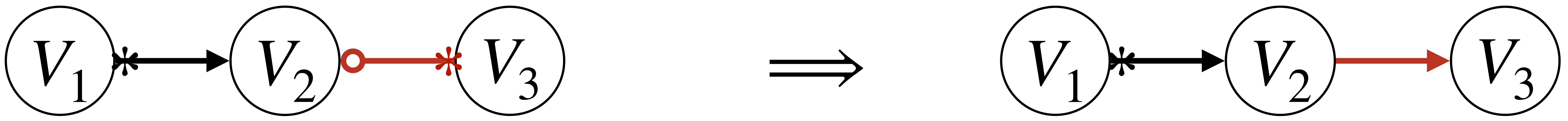
After Skel + R0 + R3 + R1 + R4 + R4

$V_1 \perp\!\!\!\perp V_3 \mid V_4$

$V_4 \perp\!\!\!\perp V_5 \mid V_1, V_2, V_3$

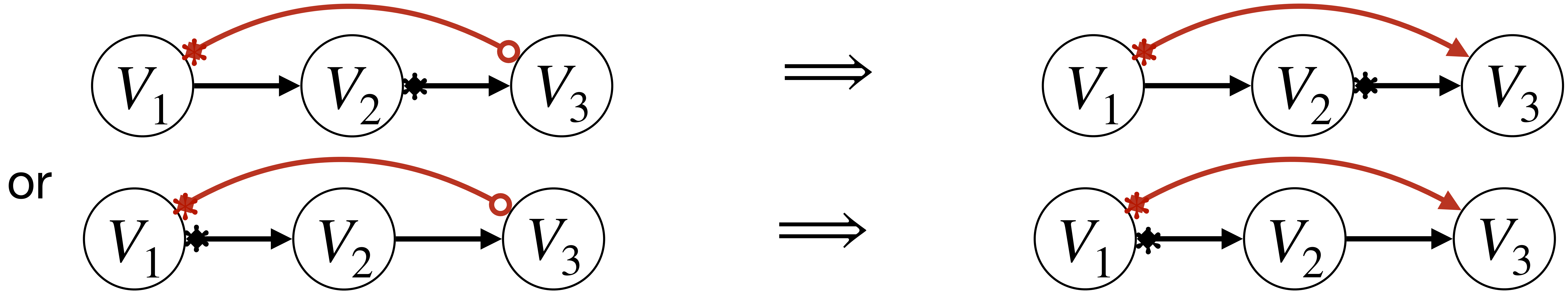
FCI - Complete Set of Mark Inference Rules

R1:

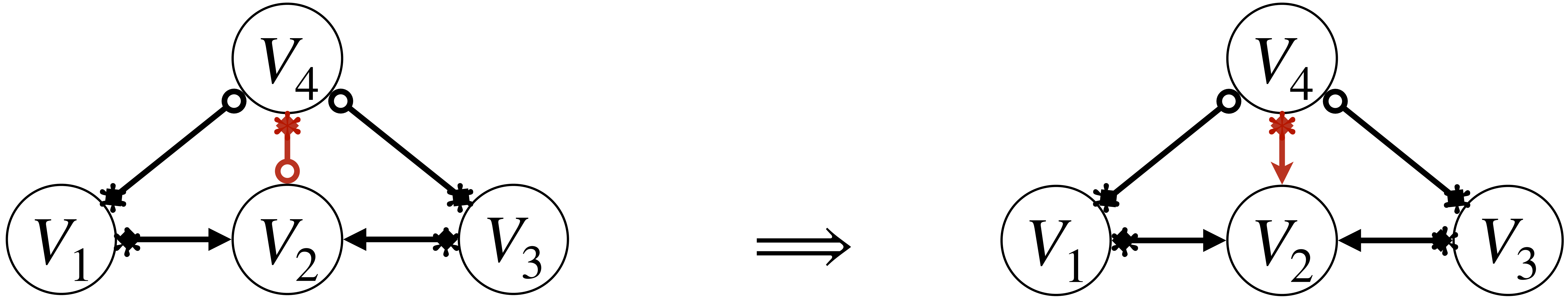


where V_1 and V_3 are not adjacent

R2:



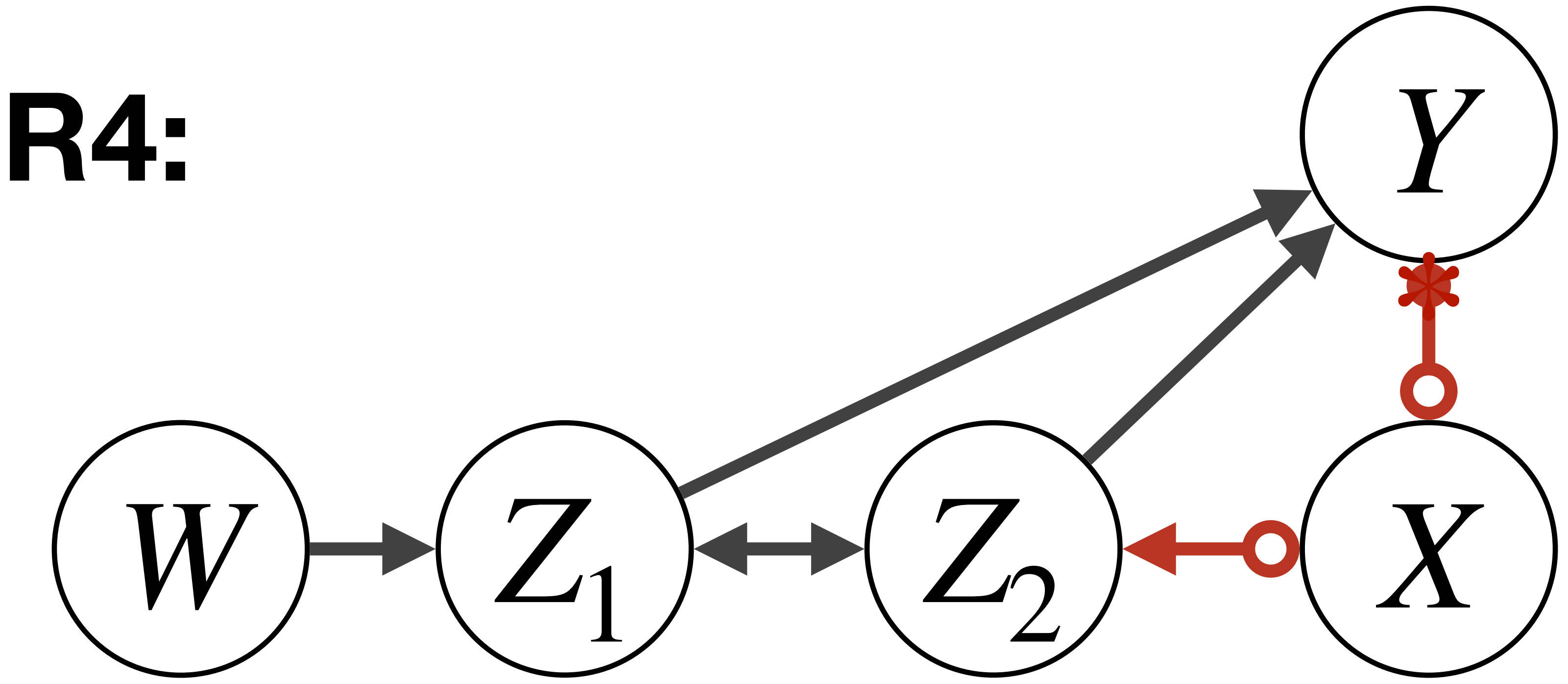
R3:



where V_1 and V_3 are not adjacent

FCI - Complete Set of Mark Inference Rules

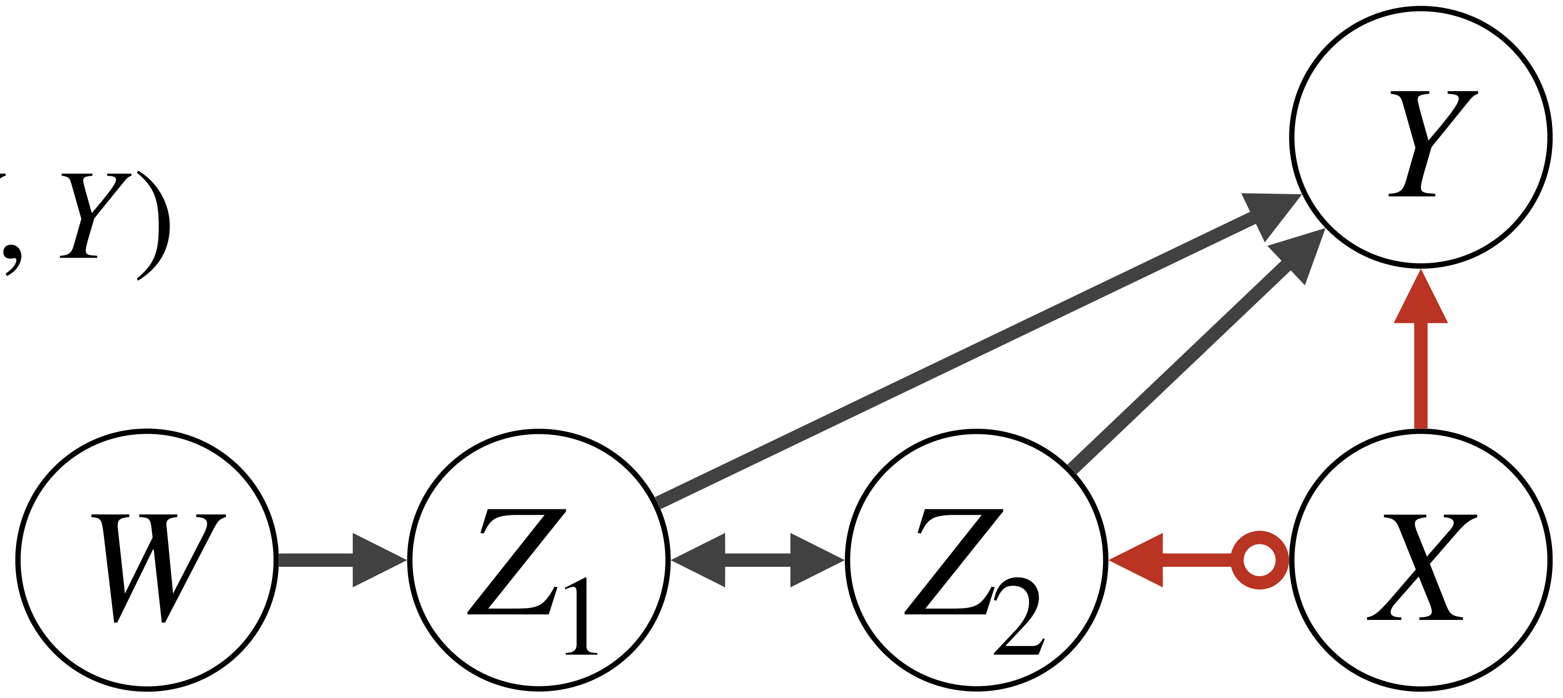
R4:



$\langle W, Z_1, Z_2, X, Y \rangle$ is a discriminating path for X

$X \in Sepset(W, Y)$

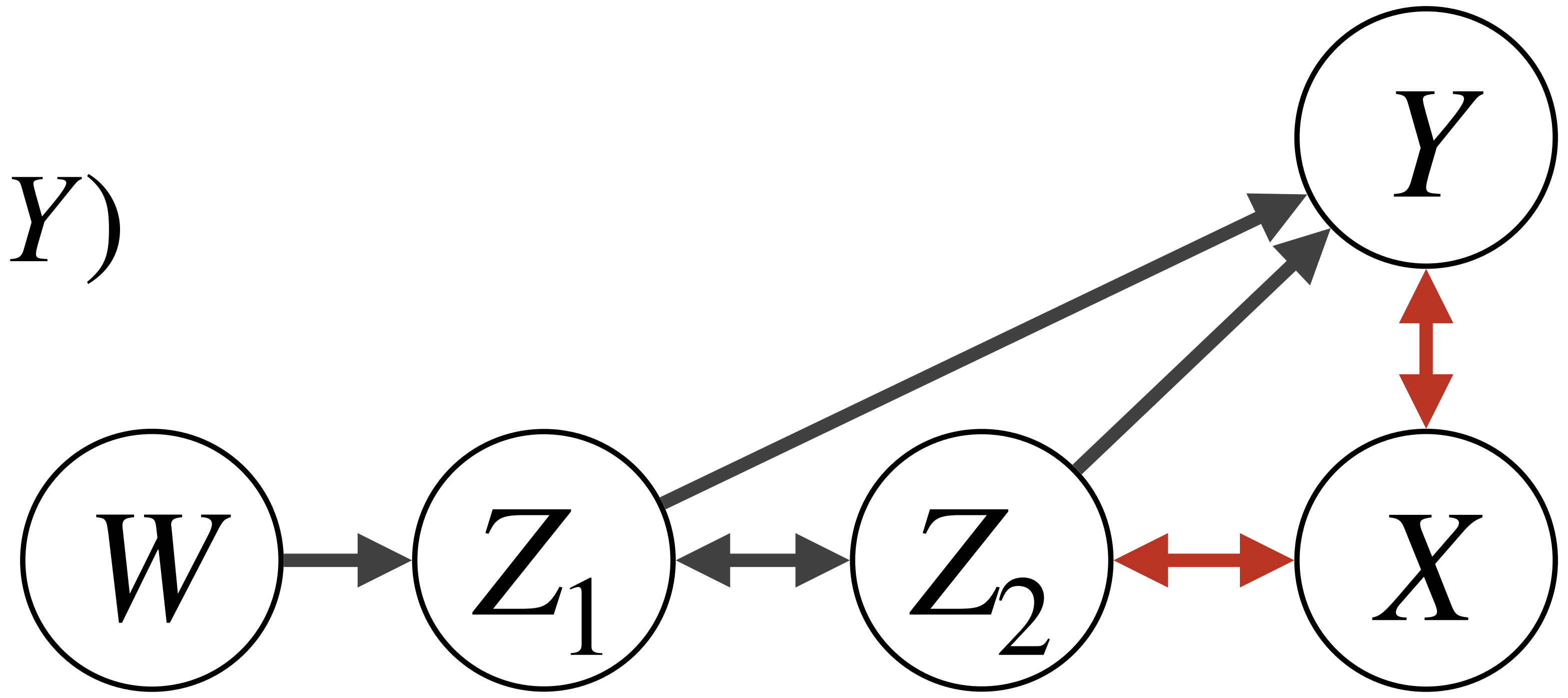
$\Longrightarrow$



X is a non-collider in $\langle Z_2, X, Y \rangle$

$X \notin Sepset(W, Y)$

$\Longrightarrow$



X is a collider in $\langle Z_2, X, Y \rangle$

Definition (discriminating path): A path $p = \langle X, \ldots, W, V, Y \rangle$ in a MAG is a discriminating path for V if

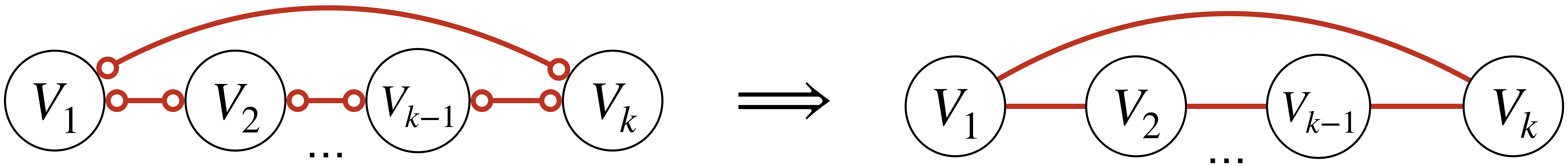
(i) p includes at least three edges;

(ii) V is a non-endpoint vertex on p, and is adjacent to Y on p; and

(iii) X is not adjacent to Y , and every vertex between X and V is a collider on p and is a parent of Y.

FCI - Complete Set of Mark Inference Rules

R5:

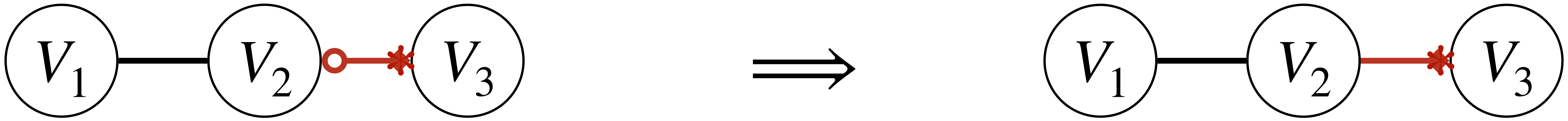


$\langle V_1, V_2, \ldots, V_{k-1}, V_k \rangle$ is an uncovered circle path

V_1 and V_{k-1} are not adjacent

V_2 and V_k are not adjacent

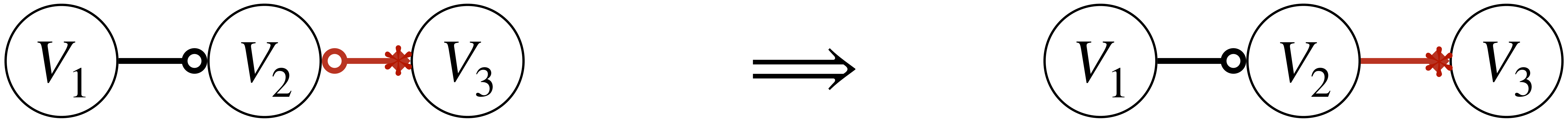
R6:



V_1 and V_3 may or may not be adjacent

FCI - Complete Set of Mark Inference Rules

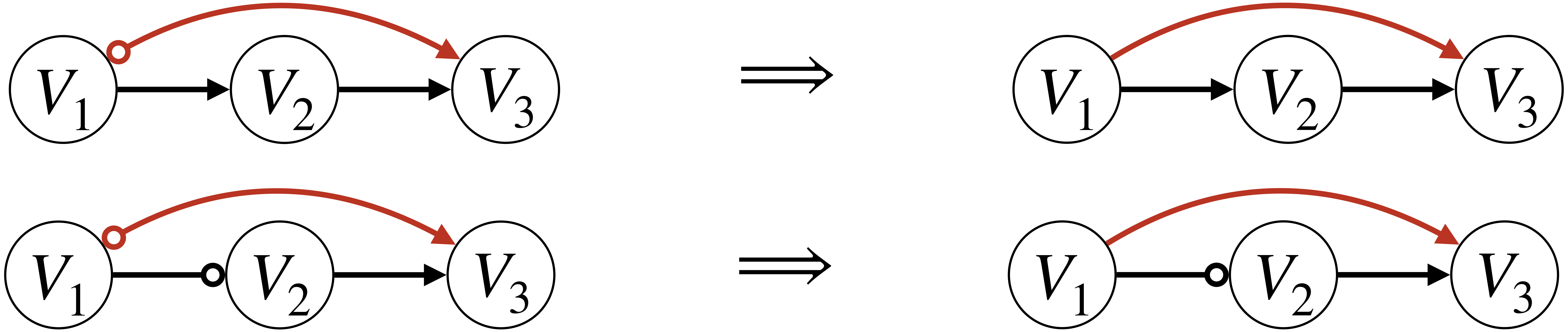
R7:



V_1 and V_3 are not adjacent

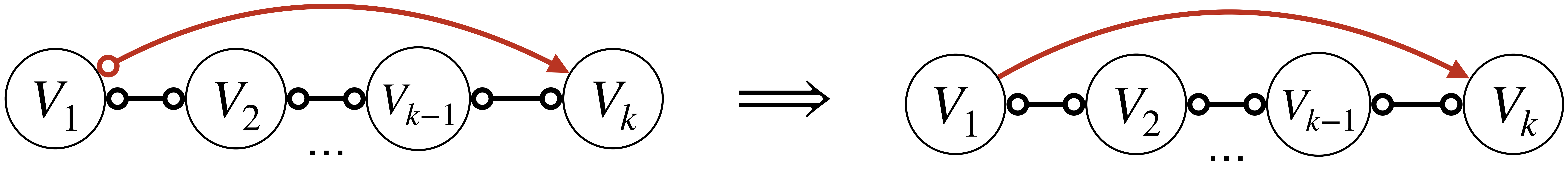
R8:

or



FCI - Complete Set of Mark Inference Rules

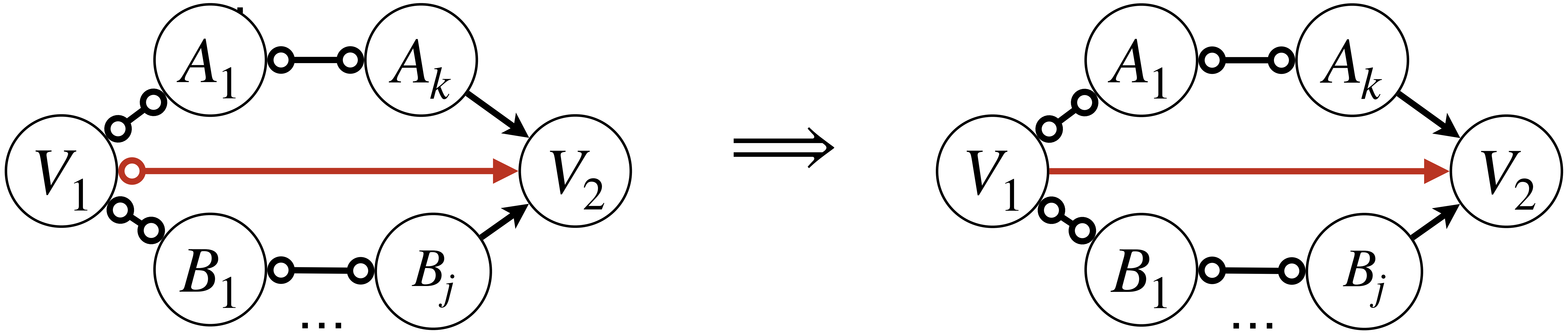
R9:



$\langle V_1, V_2, \ldots, V_{k-1}, V_k \rangle$ is an uncovered potentially directed path from V_1 to V_k

V_2 and V_k are not adjacent

R10:

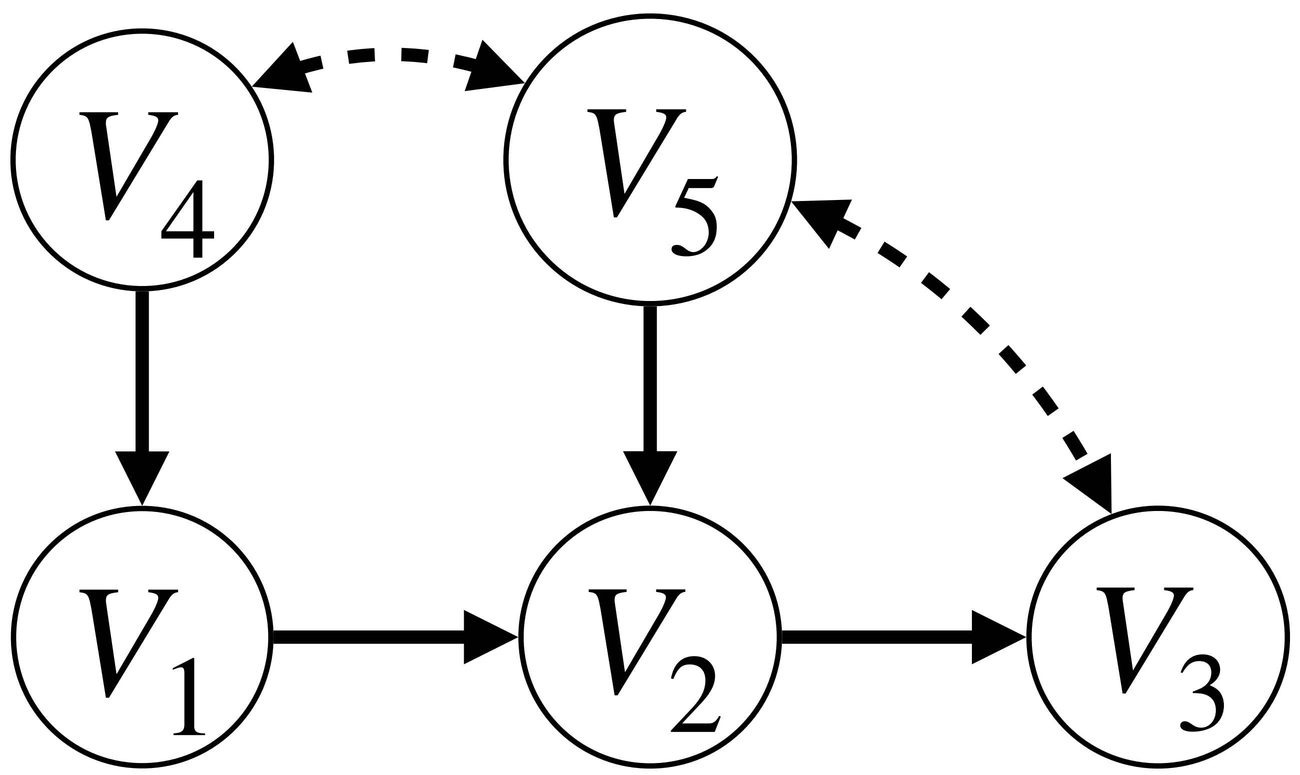


$\langle V_1, A_1, \ldots, A_k \rangle$ is an uncovered potentially directed path from V_1 to A_k (A_1 may be A_k)

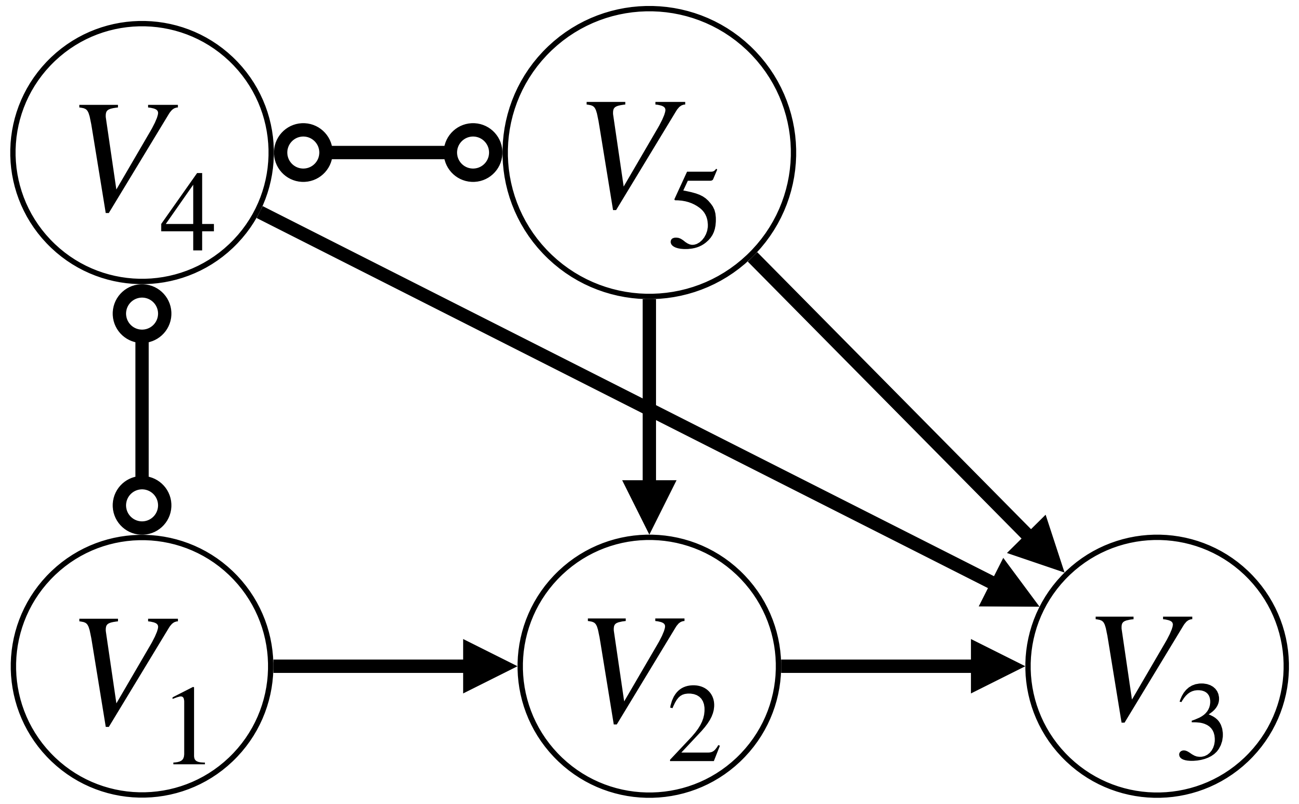
$\langle V_1, B_1, \ldots, B_k \rangle$ is an uncovered potentially directed path from V_1 to B_k (B_1 may be B_k)

$A_1 \neq B_1$ and A_1 and B_1 are not adjacent

Another Example



True, unknown
Causal Diagram



Corresponding PAG

$V_1 \perp\!\!\!\perp V_3 \mid V_2, V_4, V_5$
$V_1 \perp\!\!\!\perp V_5 \mid V_4$
$V_2 \perp\!\!\!\perp V_4 \mid V_1, V_5$

$V_1 \perp\!\!\!\perp V_3 \mid V_2, V_4, V_5$
$V_1 \perp\!\!\!\perp V_5 \mid V_4$
$V_2 \perp\!\!\!\perp V_4 \mid V_1, V_5$

Hint: apply Rules 0, 1, 2, 4 and then Rule 9 three times.

Available Implementations of the FCI

R Packages:

- pcalg R package:
  - https://cran.r-project.org/web/packages/pcalg/
  - https://github.com/cran/pcalg/
- RPy-Tetrad (Wrapper in R): https://github.com/cmu-phil/py-tetrad/tree/main/pytetrad/R

Python Packages:

- Do-discover in PyWhy: https://github.com/py-why/dodiscover
- Causal-Learn: https://causal-learn.readthedocs.io/en/latest/index.html
- Py-Tetrad (Wrapper in Python): https://github.com/bd2kccd/py-causal

Other Causal Discovery Algorithms

Hybrid approaches accounting for latent confounding:

- **M3HC:** Max-Min Hill Climbing, by Tsirlis et al., 2018 — extends GSMAG by introducing a constraint-based first phase that greatly reduces the space of structures to investigate.
- **GFCI:** Greedy FCI, by Ogarrio et al., 2016 — combines FGES and FCI. The skeleton and orientation phases are firstly performed using FGES and then refined by using the FCI.

Parametric approaches under causal sufficiency — Identifiable Structure

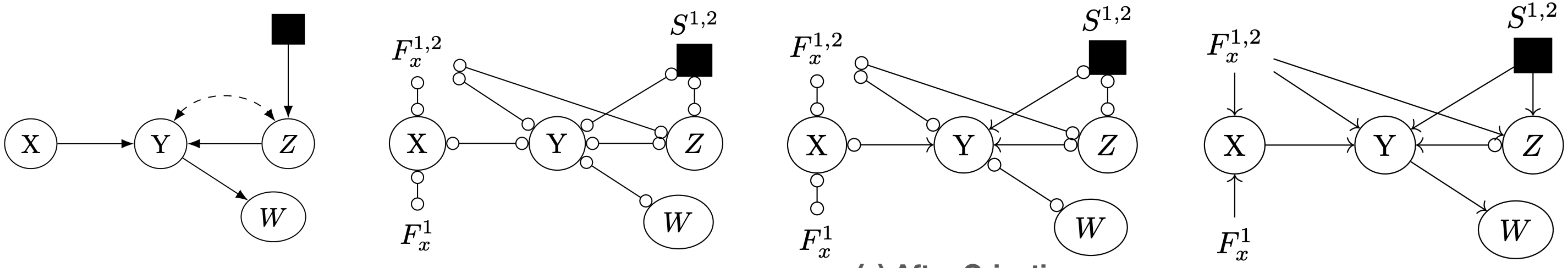
- **LiNGAM:** Linear, non-Gaussian, and Acyclic Model, by Shimizu et al., 2006 — leverage distributional asymmetries with linear causal mechanisms are non-Gaussian error terms.
- **ANM:** Non-linear additive noise model (Hoyer et al., 2009; Zhang and Hyvärinen, 2009a) — leverage distributional asymmetries with non-linear mechanisms and additive noise.

Advances in Causal Discovery with Unobserved Confounding

Going *Beyond* the Markov Equivalence Class:

1. Causal Discovery with Interventional Data

    - **Jaber, A., Kocaoglu, M., Shanmugam, K. and Bareinboim, E.,** (2020). Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems*, *33*, pp.9551-9561.
    - **A. Li, A. Jaber, E. Bareinboim.** Causal discovery from observational and interventional data across multiple environments. (2023) In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems* — NeurIPS-23.



(a) G_S (b) Skeleton (c) After Orienting Unshielded Colliders (d) Final S-PAG

Advances in Causal Discovery with Unobserved Confounding

Going *Beyond* the Markov Equivalence Class:

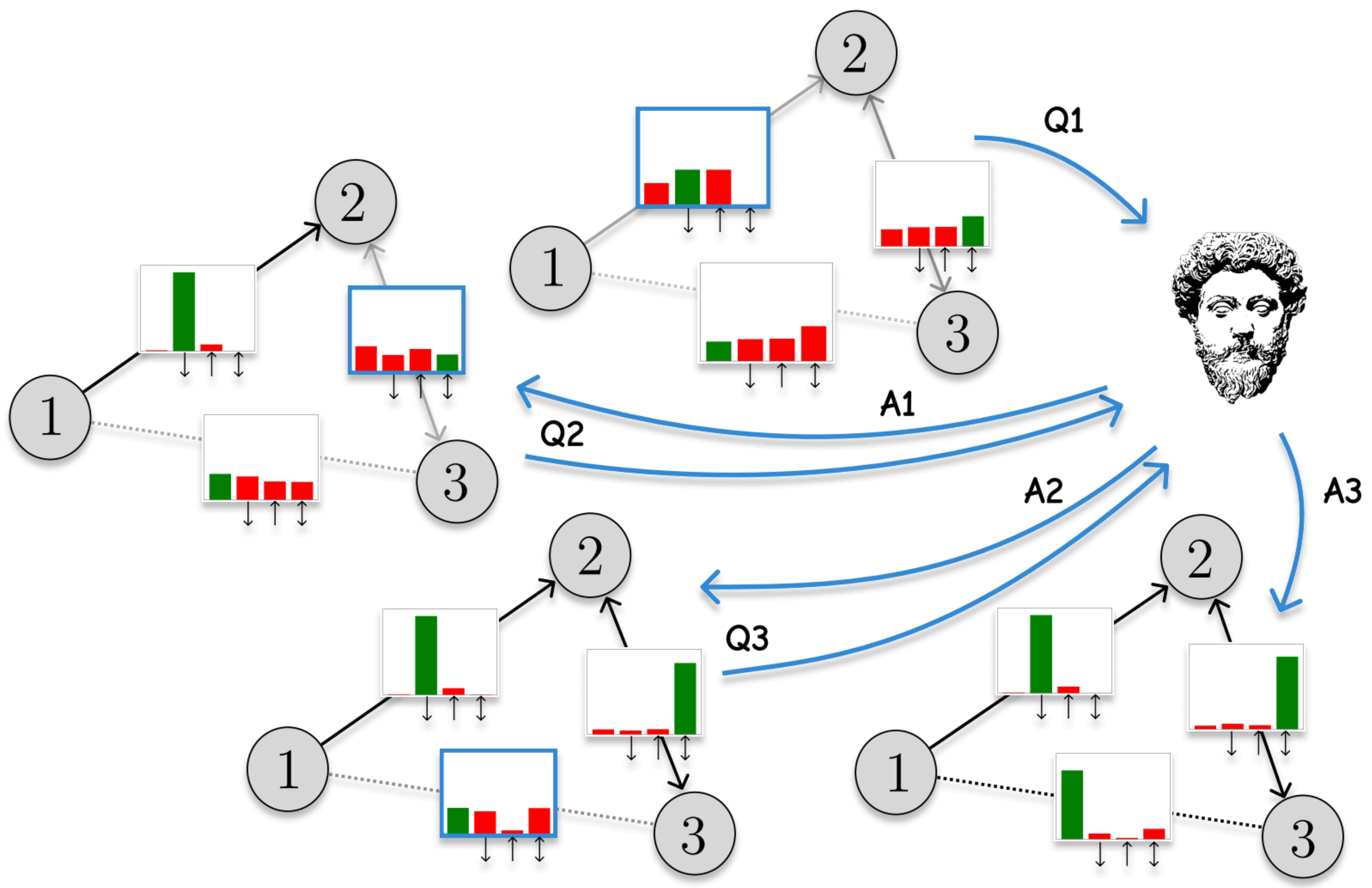
2. Causal Discovery with Prior Knowledge

   - **Wang, T. Z., Qin, T. and Zhou, Z.H.**, (2022). Sound and complete causal identification with latent variables given local background knowledge. *Advances in Neural Information Processing Systems*, *35*, pp.10325-10338.

   - **Venkateswaran, A., & Perkovic, E.** (2024). Towards Complete Causal Explanation with Expert Knowledge. arXiv preprint arXiv:2407.07338.

3. Human-in-the-Loop Probabilistic Causal Discovery

   - **da Silva, T., Silva, E., Ribeiro, A., Góis, A., Heider, D., Kaski, S., & Mesquita, D.** (2023). Human-in-the-Loop Causal Discovery under Latent Confounding using Ancestral GFlowNets. arXiv preprint arXiv:2309.12032.



Advances in Causal Discovery with Unobserved Confounding

Going *Beyond* the Markov Equivalence Class:

4. Causal Discovery in Linear Models

    - **Tashiro, T., Shimizu, S., Hyvärinen, A., & Washio, T.** (2014). ParceLiNGAM: A causal ordering method robust against latent confounders. *Neural computation*, *26*(1), 57-83.
    - **Wang, Y. S., & Drton, M.** (2023). Causal discovery with unobserved confounding and non-Gaussian data. *Journal of Machine Learning Research*, *24*(271), 1-61.

Relax the causal sufficiency assumption of LinGAN by Shimizu et al., 2006: order / ancestral identifiability under linear systems with non-gaussian error terms

5. Causal Discovery for Additive Noise Models

    - **Van Diepen, M. M., Bucur, I. G., Heskes, T., & Claassen, T.** (2023). Beyond the Markov Equivalence Class: Extending Causal Discovery under Latent Confounding. In *Conference on Causal Learning and Reasoning* (pp. 707-725). PMLR.

FCI-CDC: causal direction criterion (CDC) allows pairwise orientation in (weakly) additive noise models with independent causal mechanisms.

Advances in Causal Discovery with Unobserved Confounding

Learning Dynamic Systems:

1. Causal Discovery with Cycles

    - **Bongers, S., Forré, P., Peters, J., & Mooij, J. M.** (2021). Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, *49*(5), 2885-2915.

    - **Claassen, T. &; Mooij, J.M.. (2023).** Establishing Markov equivalence in cyclic directed graphs. Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence, PMLR 216:433-442, 2023.

2. Causal Discovery from Time-Series Data

    - **Gerhardus, A., & Runge, J.** (2020). High-recall causal discovery for autocorrelated time series with latent confounders. *Advances in Neural Information Processing Systems (NeurIPS 2020)*, *33*, 12615-12625.

Current Challenges & Open Problems

- Robustness in real-world scenarios, with small (unfaithful) datasets.
- Scalability in insufficient systems — development of adaptive approaches.
- Uncertainty modeling, lack of ground-truth in real-world applications.
- Integration of expert / human knowledge — completeness results.
- Causal experimental design — what if a causal relation is not identified?
- Learning from multi-modal datasets — connection with causal abstraction and causal representation learning.
- Continual causal discovery

Additional Resources

- Causality Tutorial: https://github.com/adele/Causality-Tutorial/

  → Causal Discovery — Google Colab Notebook: (Link)

- Tutorials, talks, and complete lectures on YouTube: (Link)

Feel free to reach out to me if you have any questions or are interested in collaborations.

adele.ribeiro@uni-marburg.de

Thank you! :)