Representation of Causal Knowledge and Causal Discovery

Adèle Helena Ribeiro

TECHNISCHE UNIVERSITÄT DARMSTADT

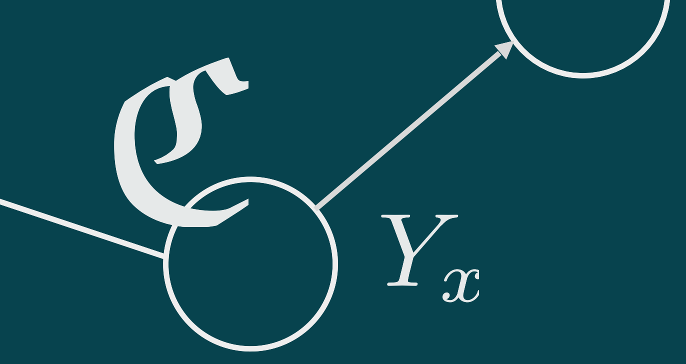
PHILIPPS-UNIVERSITÄT MARBURG

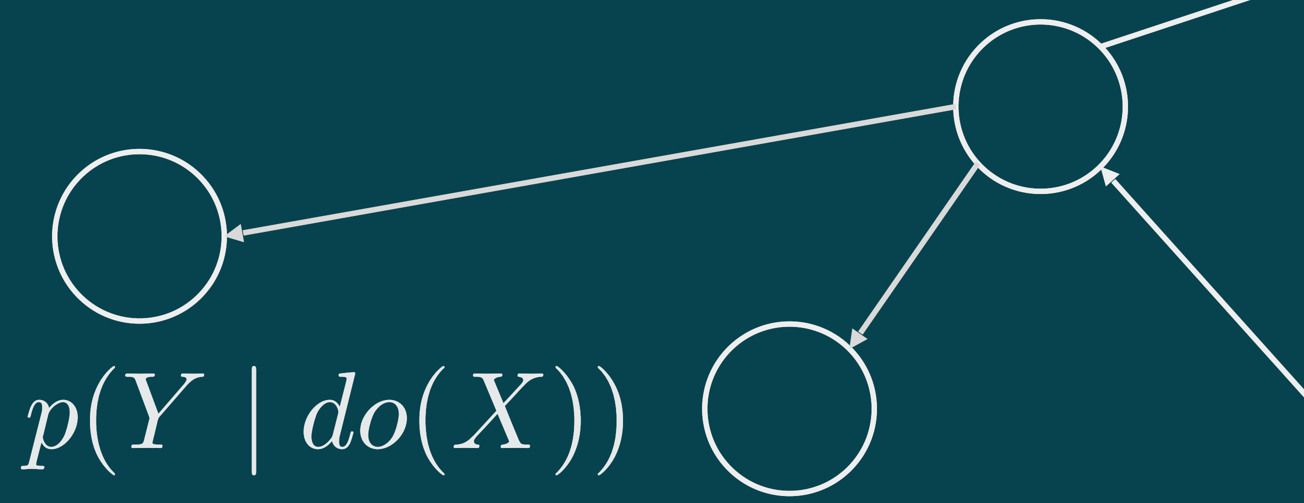
25th July 2023

ESSAI & ACAI 2023

LJUBLJANA, SLOVENIA

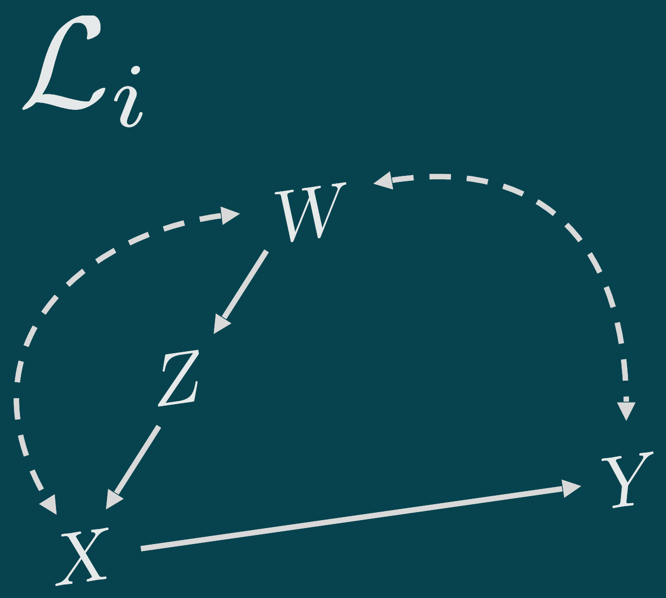
Machines Climbing Pearl's Ladder of Causation





1 | Encoding Causal Structural Knowledge

Causal Diagrams



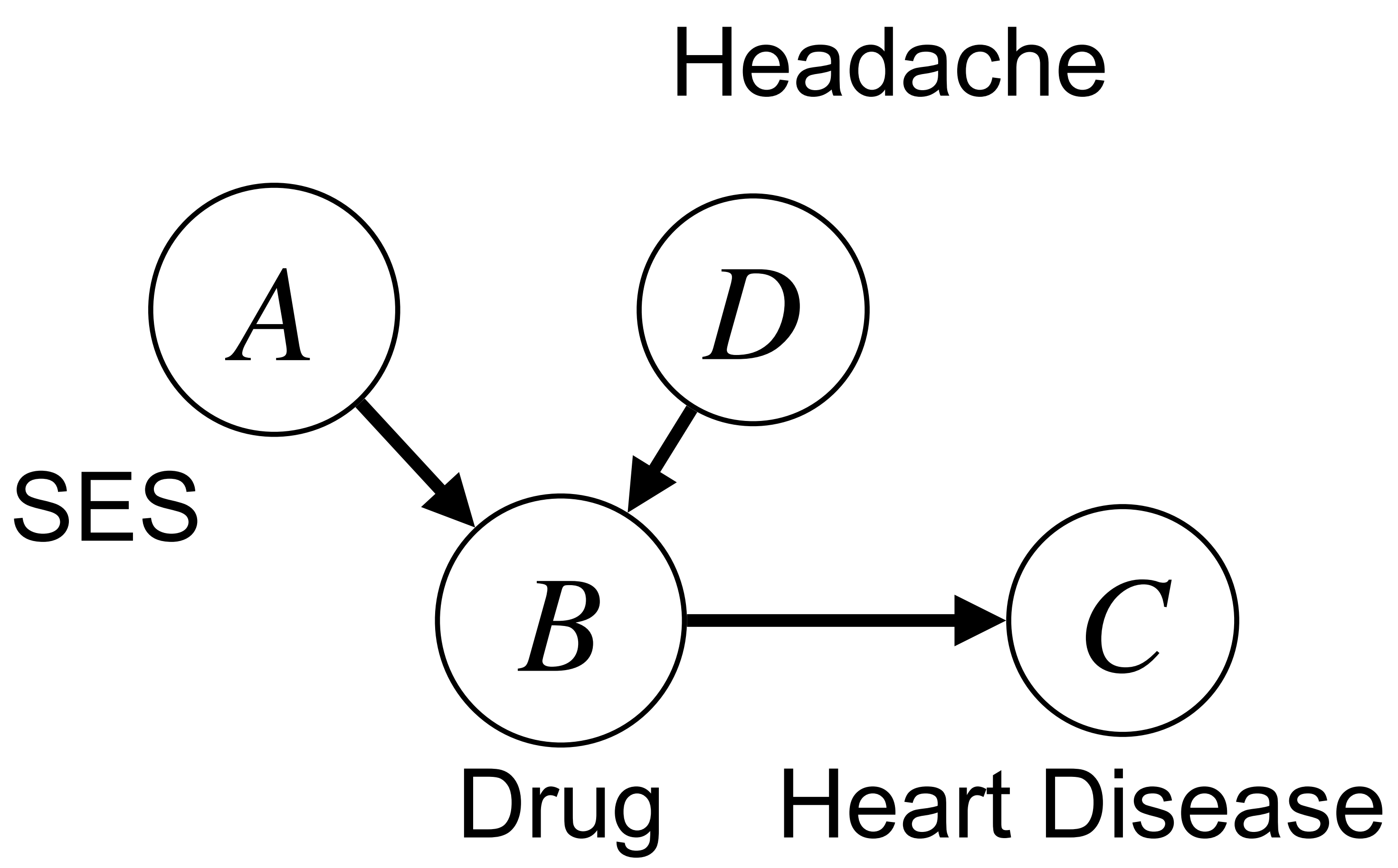
Causal Diagram: Encoder of ***Structural*** Knowledge

Structural Causal Model (SCM)

$$\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathscr{F}, P(\mathbf{u}) \rangle$$

$$\mathscr{M} = \begin{cases} \mathbf{V} = \{A, B, C, D\} \\ \mathbf{U} = \{U_A, U_B, U_C, U_D, U_{CD}\} \\ \mathscr{F} = \begin{cases} A \leftarrow f_A(U_A) \\ B \leftarrow f_B(A, D, U_B) \\ D \leftarrow f_Z(U_D, U_{CD}) \\ C \leftarrow f_X(B, U_C, U_{CD}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Induced Causal Diagram



An SCM $\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathscr{F}, P(\mathbf{u}) \rangle$ induces a causal diagram such that, **for every** $V_i, V_j \in \mathbf{V}$:

$V_i \rightarrow V_j$, if V_i appears as argument of $f_j \in \mathscr{F}$.

Causal Diagram: Encoder of ***Structural*** Knowledge

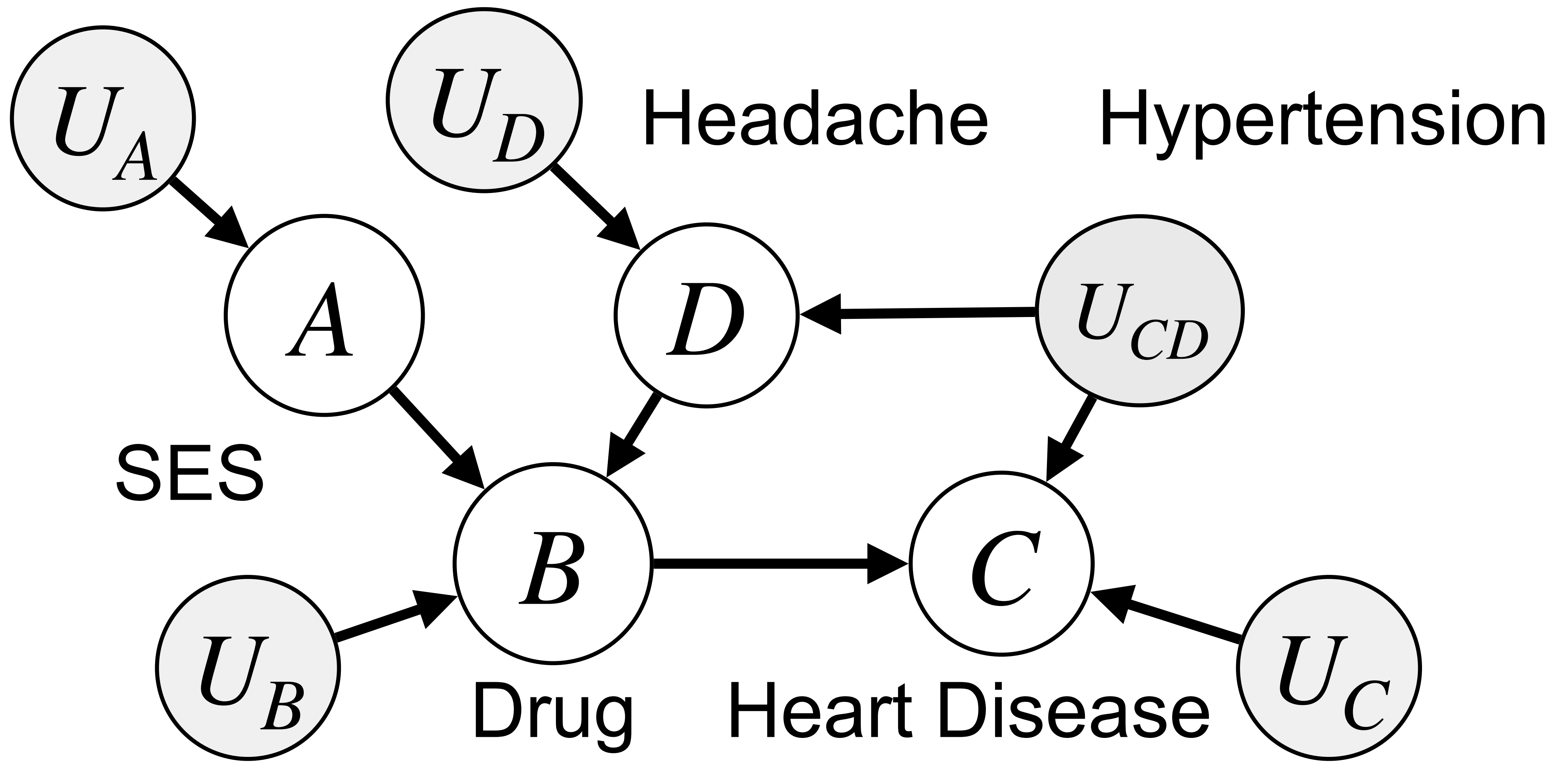
Structural Causal Model (SCM)

$$\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$$

$$\mathcal{M} = \begin{cases} \mathbf{V} = \{A, B, C, D\} \\ \mathbf{U} = \{U_A, U_B, U_C, U_D, U_{CD}\} \\ \mathcal{F} = \begin{cases} A \leftarrow f_A(U_A) \\ B \leftarrow f_B(A, D, U_B) \\ D \leftarrow f_Z(U_D, U_{CD}) \\ C \leftarrow f_X(B, U_C, U_{CD}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

$\longrightarrow$

Induced Causal Diagram



An SCM $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$ induces a causal diagram such that, **for every** $V_i, V_j \in \mathbf{V}$:

$V_i \rightarrow V_j$, if V_i appears as argument of $f_j \in \mathcal{F}$.

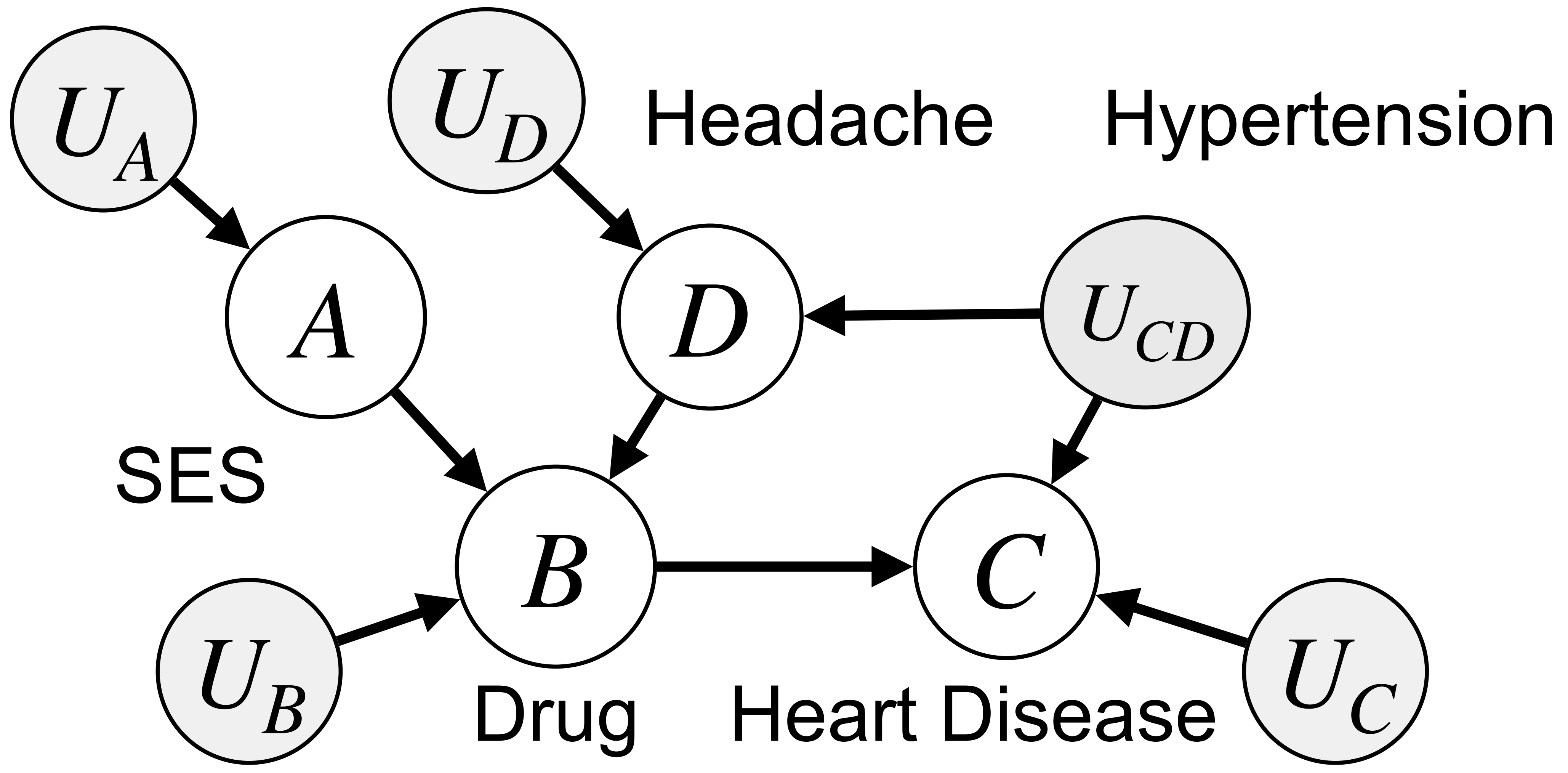
Causal Diagram: Encoder of ***Structural*** Knowledge

Structural Causal Model (SCM)

$$\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathscr{F}, P(\mathbf{u}) \rangle$$

$$\mathscr{M} = \begin{cases} \mathbf{V} = \{A, B, C, D\} \\ \mathbf{U} = \{U_A, U_B, U_C, U_D, U_{CD}\} \\ \mathscr{F} = \begin{cases} A \leftarrow f_A(U_A) \\ B \leftarrow f_B(A, D, U_B) \\ D \leftarrow f_Z(U_D, U_{CD}) \\ C \leftarrow f_X(B, U_C, U_{CD}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Induced Causal Diagram



An SCM $\mathscr{M} = \langle \mathbf{V}, \mathbf{U}, \mathscr{F}, P(\mathbf{u}) \rangle$ induces a causal diagram such that, **for every** $V_i, V_j \in \mathbf{V}$:

$V_i \rightarrow V_j$, if V_i appears as argument of $f_j \in \mathscr{F}$.

$V_i \dashleftarrow\dashrightarrow V_j$ if the corresponding $U_i, U_j \in \mathbf{U}$ are correlated or f_i , f_j share some argument $U \in \mathbf{U}$.

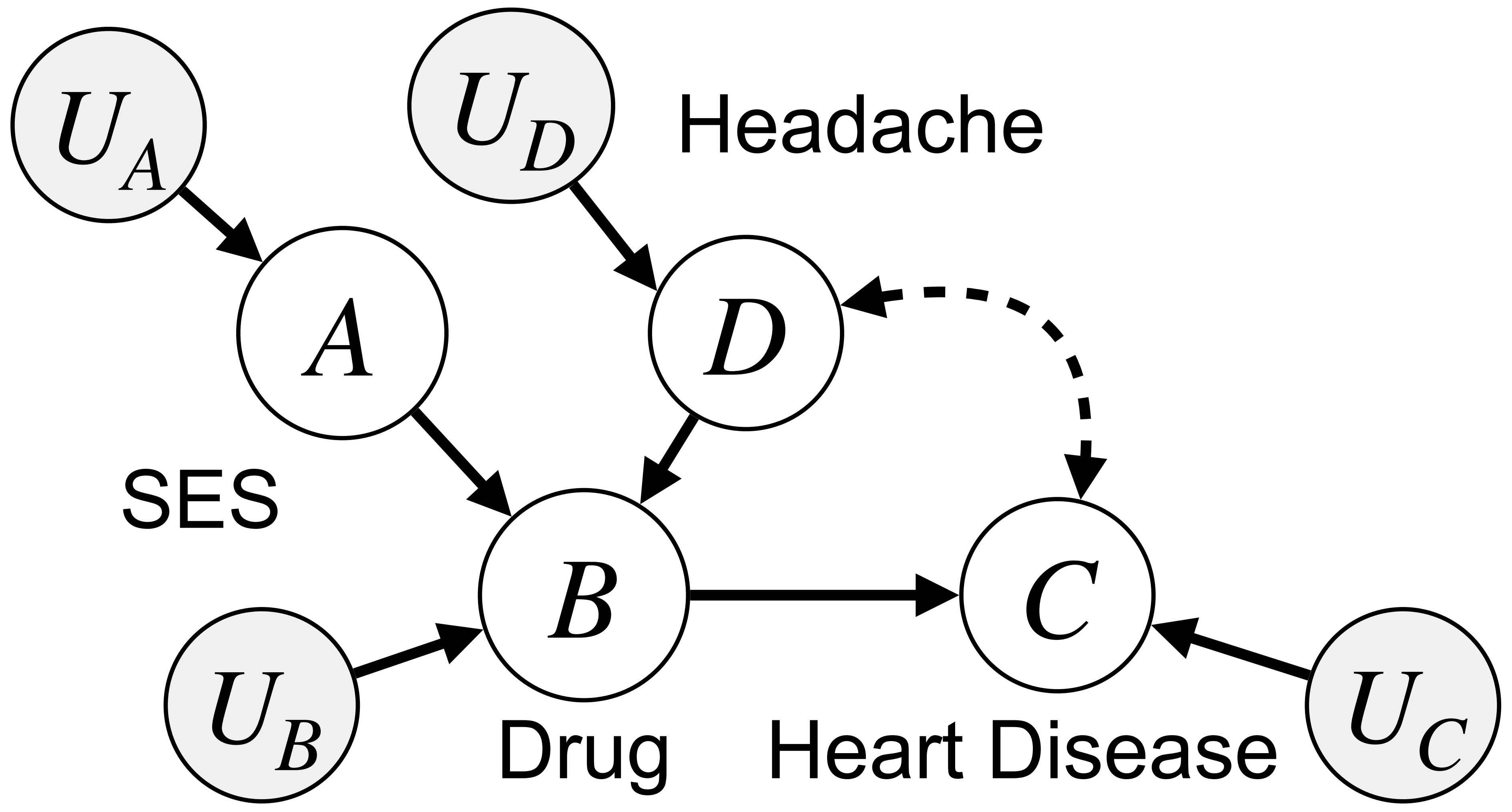
Causal Diagram: Encoder of ***Structural*** Knowledge

Structural Causal Model (SCM)

$$\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$$

$$\mathcal{M} = \begin{cases} \mathbf{V} = \{A, B, C, D\} \\ \mathbf{U} = \{U_A, U_B, U_C, U_D, U_{CD}\} \\ \mathcal{F} = \begin{cases} A \leftarrow f_A(U_A) \\ B \leftarrow f_B(A, D, U_B) \\ D \leftarrow f_Z(U_D, U_{CD}) \\ C \leftarrow f_X(B, U_C, U_{CD}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

Induced Causal Diagram



An SCM $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$ induces a causal diagram such that, **for every** $V_i, V_j \in \mathbf{V}$:

$V_i \rightarrow V_j$, if V_i appears as argument of $f_j \in \mathcal{F}$.

$V_i \leftarrow\!\!\dashrightarrow V_j$ if the corresponding $U_i, U_j \in \mathbf{U}$ are correlated or f_i , f_j share some argument $U \in \mathbf{U}$.

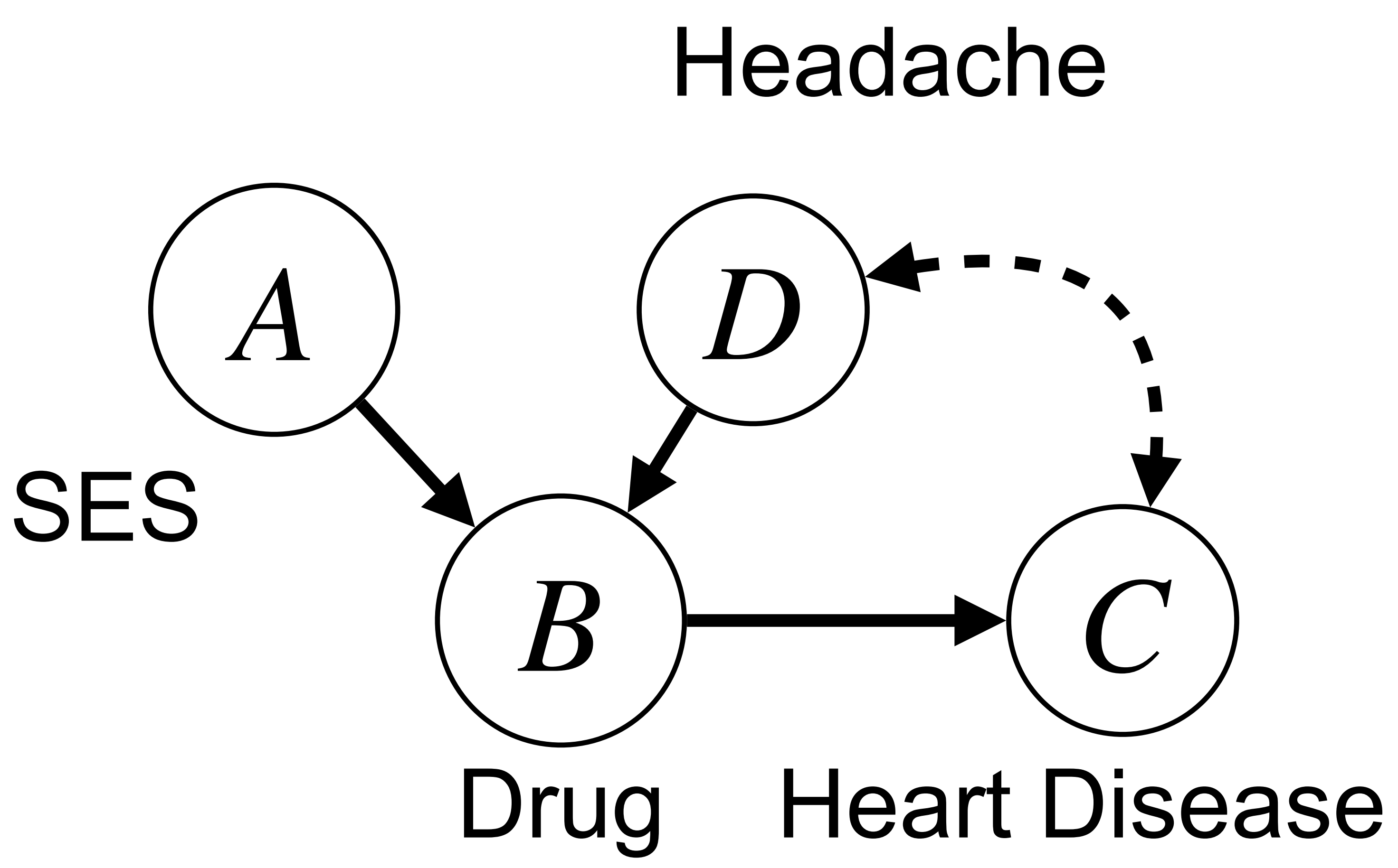
Causal Diagram: Encoder of ***Structural*** Knowledge

Structural Causal Model (SCM)

$$\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$$

$$\mathcal{M} = \begin{cases} \mathbf{V} = \{A, B, C, D\} \\ \mathbf{U} = \{U_A, U_B, U_C, U_D, U_{CD}\} \\ \mathcal{F} = \begin{cases} A \leftarrow f_A(U_A) \\ B \leftarrow f_B(A, D, U_B) \\ D \leftarrow f_Z(U_D, U_{CD}) \\ C \leftarrow f_X(B, U_C, U_{CD}) \end{cases} \\ P(\mathbf{U}) \end{cases}$$

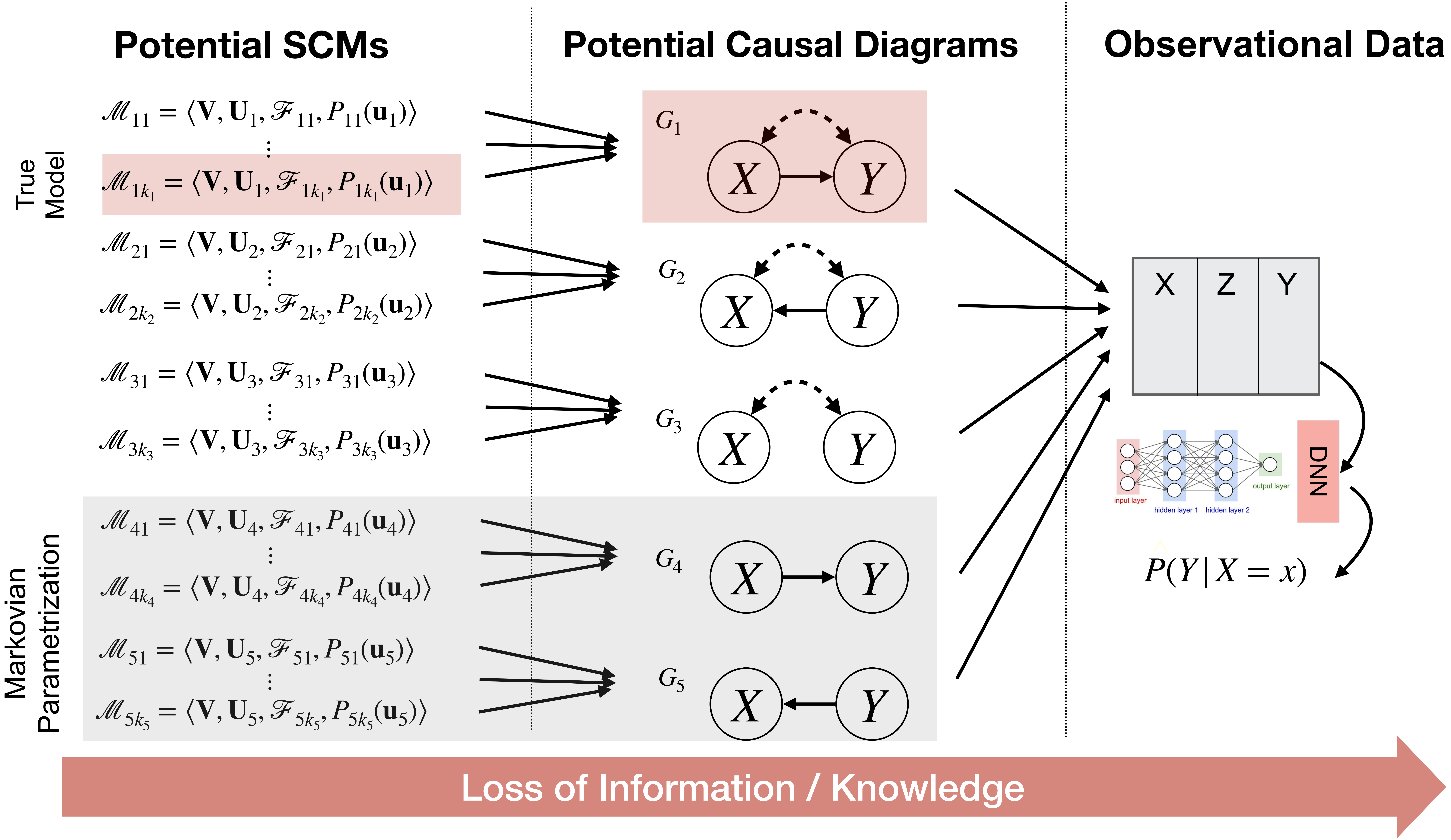
Induced Causal Diagram



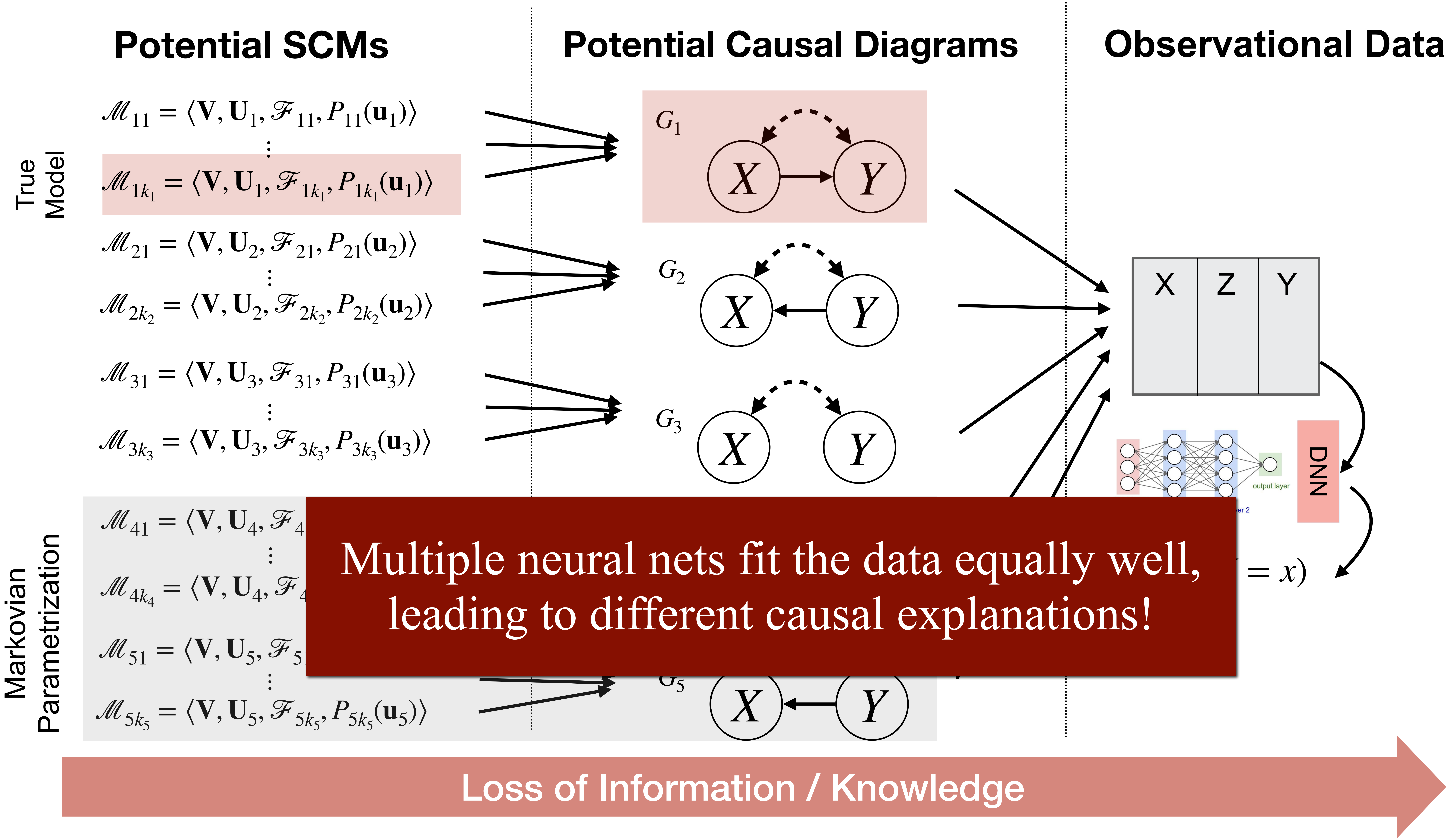
An SCM $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{u}) \rangle$ induces a causal diagram such that, **for every** $V_i, V_j \in \mathbf{V}$:

$V_i \rightarrow V_j$, if V_i appears as argument of $f_j \in \mathcal{F}$.

$V_i \dashleftarrow\dashrightarrow V_j$ if the corresponding $U_i, U_j \in \mathbf{U}$ are correlated or f_i , f_j share some argument $U \in \mathbf{U}$.



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



Potential SCMs
Potential Causal Diagrams
Observational Data
True Model
$\mathcal{M}_{11} = \langle \mathbf{V}, \mathbf{U}_1, \mathcal{F}_{11}, P_{11}(\mathbf{u}_1) \rangle$
$\vdots$
$\mathcal{M}_{1k_1} = \langle \mathbf{V}, \mathbf{U}_1, \mathcal{F}_{1k_1}, P_{1k_1}(\mathbf{u}_1) \rangle$
$\mathcal{M}_{21} = \langle \mathbf{V}, \mathbf{U}_2, \mathcal{F}_{21}, P_{21}(\mathbf{u}_2) \rangle$
$\vdots$
$\mathcal{M}_{2k_2} = \langle \mathbf{V}, \mathbf{U}_2, \mathcal{F}_{2k_2}, P_{2k_2}(\mathbf{u}_2) \rangle$
$\mathcal{M}_{31} = \langle \mathbf{V}, \mathbf{U}_3, \mathcal{F}_{31}, P_{31}(\mathbf{u}_3) \rangle$
$\vdots$
$\mathcal{M}_{3k_3} = \langle \mathbf{V}, \mathbf{U}_3, \mathcal{F}_{3k_3}, P_{3k_3}(\mathbf{u}_3) \rangle$
Markovian Parametrization
$\mathcal{M}_{41} = \langle \mathbf{V}, \mathbf{U}_4, \mathcal{F}_4$
$\vdots$
$\mathcal{M}_{4k_4} = \langle \mathbf{V}, \mathbf{U}_4, \mathcal{F}$
$\mathcal{M}_{51} = \langle \mathbf{V}, \mathbf{U}_5, \mathcal{F}_5$
$\vdots$
$\mathcal{M}_{5k_5} = \langle \mathbf{V}, \mathbf{U}_5, \mathcal{F}_{5k_5}, P_{5k_5}(\mathbf{u}_5) \rangle$
G_1
X Y
G_2
X Y
G_3
X Y
G_5
X Y
X Z Y
output layer
DNN
$= x)$
Multiple neural nets fit the data equally well, leading to different causal explanations!
Loss of Information / Knowledge

Super-Exponential Growth

The space of **Markovian Causal Diagrams** (a.k.a. Directed Acyclic Graphs, or DAGs for short) 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	$\lvert DAG(n) \rvert$
2	3
3	27
4	729
5	59.049
6	1.4349×10^7
7	1.0460×10^{10}
8	2.2877×10^{13}

Inference through enumeration is not a good idea. :)

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

The space of **Markovian Causal Diagrams** (a.k.a. Acyclic Directed Mixed Graphs, or ADMGs for short) also grows super-exponentially with the number *n* of variables, and it is much bigger than the space of DAGs:

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

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}

Now, inference seems impossible…

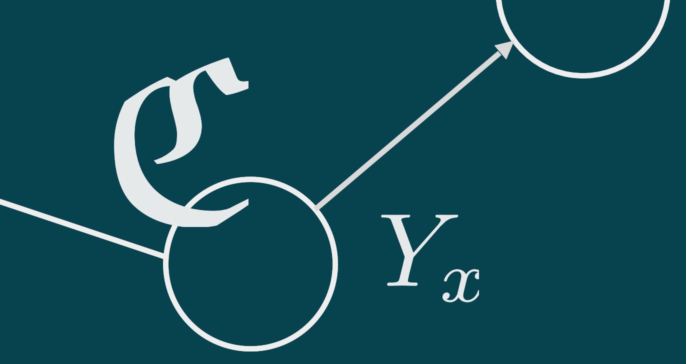
Surprisingly, that is not the case!

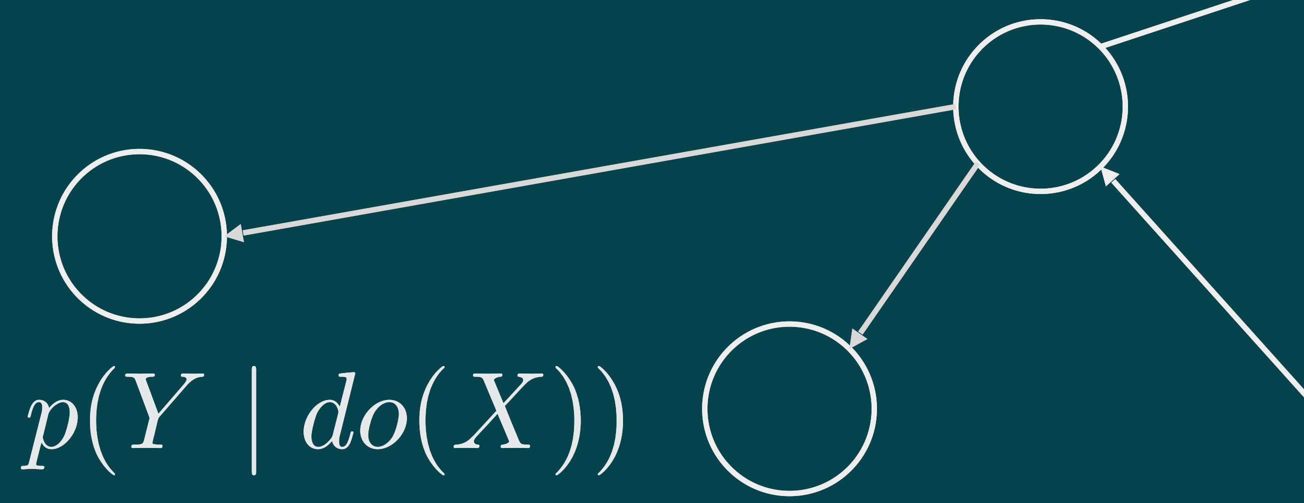
Causal Discovery

Can we learn a causal diagram $\mathscr{G}$ from observational data?

In non-parametric settings, we can't learn the true causal diagram, but **Causal Discovery** algorithms such as the **Fast Causal Inference (FCI)** can learn a graphical representation of its ***Markov Equivalence Class (MEC)***!

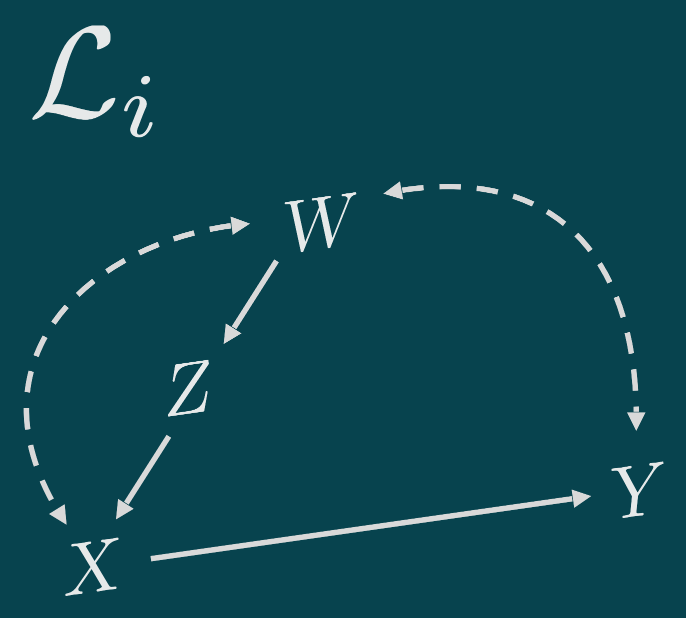
Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896. Link



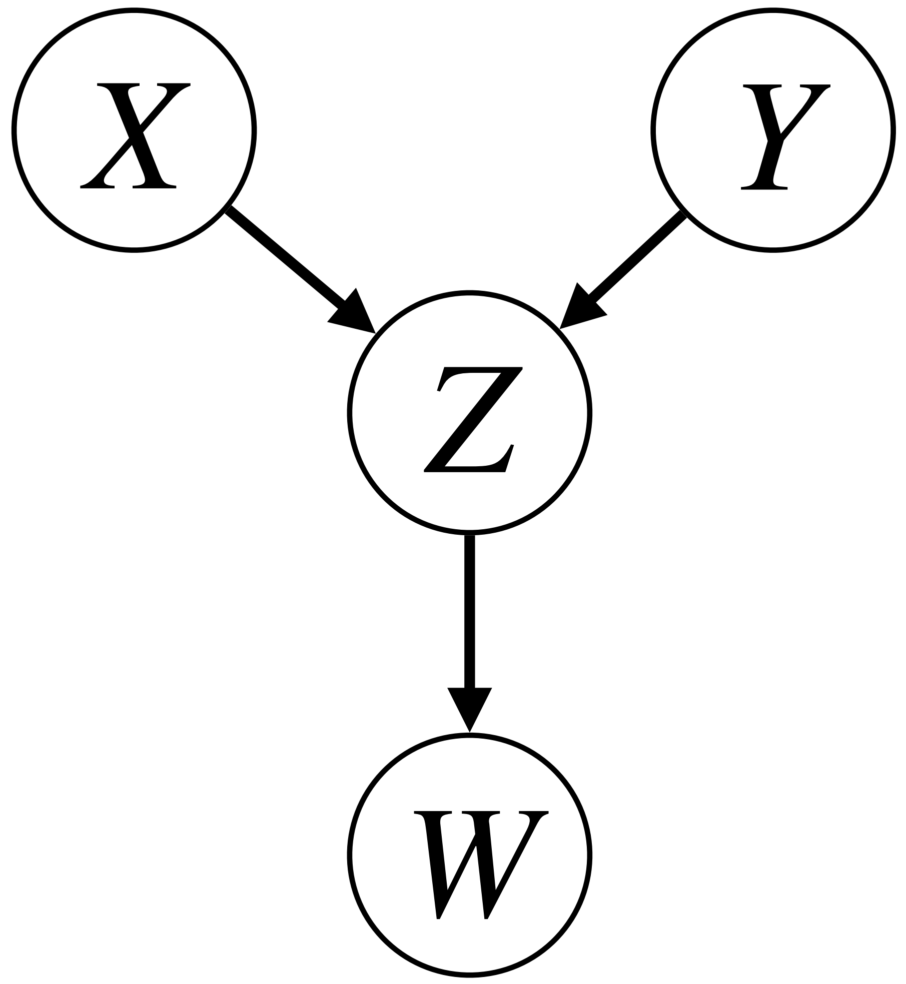


2 Encoding Conditional Independencies

D-Separation and I-Maps



Graphical Notation



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., $Y \in NDesc(X)$

$\langle X, Z, Y \rangle$ is a *collider triplet*

$\langle X, Z, W \rangle$ and $\langle Y, Z, W \rangle$ is a *non-collider triplets*

Note: In many settings, the variable itself is included in their own set of parents, ancestors, children, and descendants.

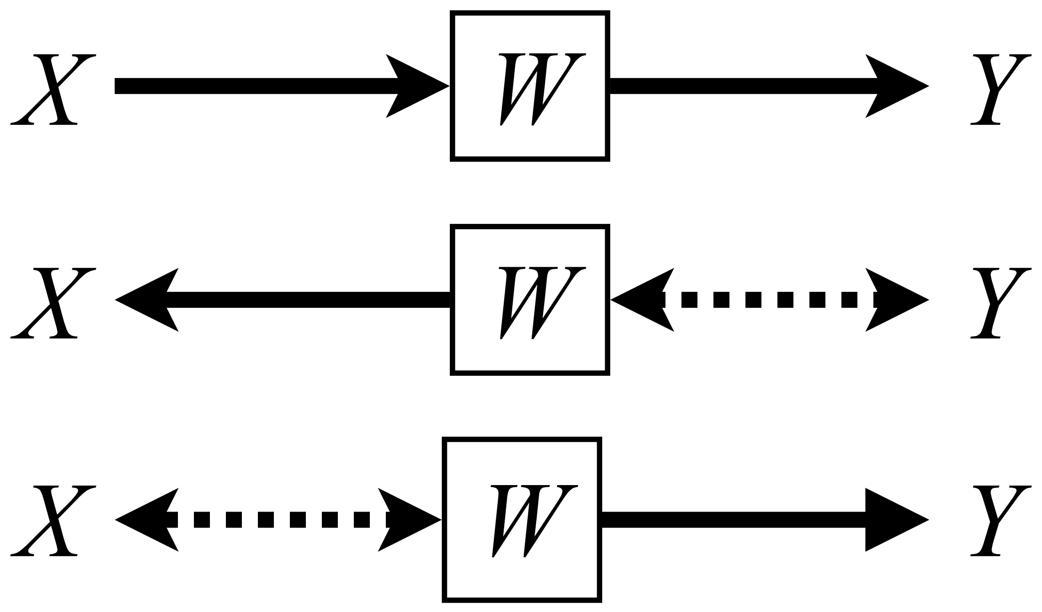
For example, $X \in Ch(X),\ An(W) = \{X, Y, Z, W\}, Pa(Z, W) = \{X, Y, Z, W\}$

D-Separation and Implied Conditional Independencies

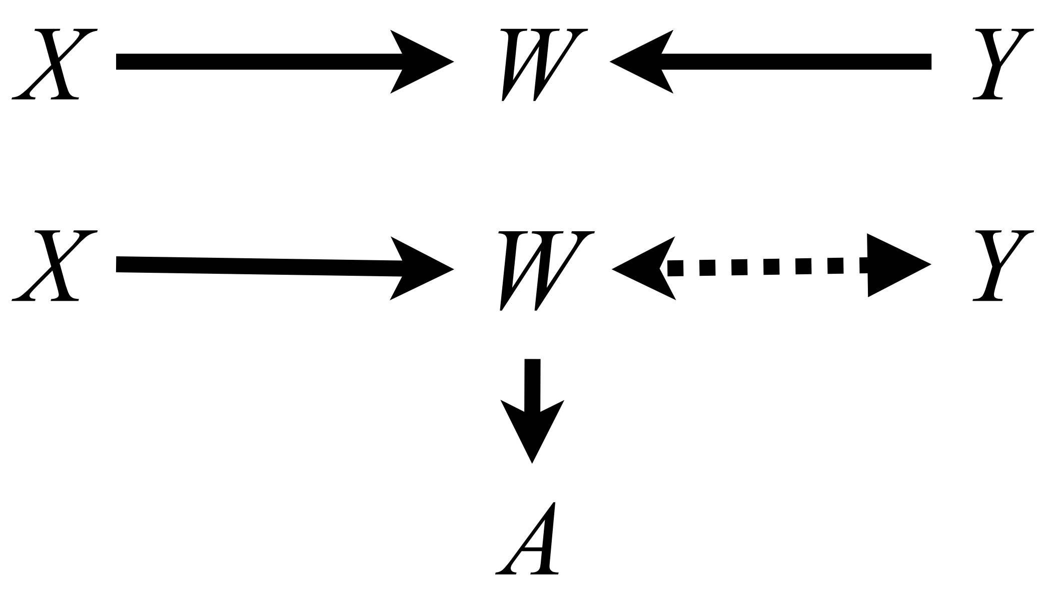
Definition (active triplet): A triplet in a subpath $\langle V_i, V_m, V_j \rangle$ is said to be ***active*** relative to a set $\mathbf{Z}$ if V_m:

1. Is a non-collider and not a member of $\mathbf{Z}$; or
2. Is a collider and an ancestor of some member of $\mathbf{Z}$.

W is non-collider
(active if $W \notin \mathbf{Z}$)



W is a collider
(active if $W \in \mathbf{Z}$ or any of its descendants is in $\mathbf{Z}$,
e.g., $A \in \mathbf{Z}$)



D-Separation and Implied Conditional Independencies

Definition (active triplet): A triplet in a subpath $\langle V_i, V_m, V_j \rangle$ is said to be ***active*** relative to a set $\mathbf{Z}$ if V_m:

1. Is a non-collider and not a member of $\mathbf{Z}$; or
2. Is a collider and an ancestor of some member of $\mathbf{Z}$.

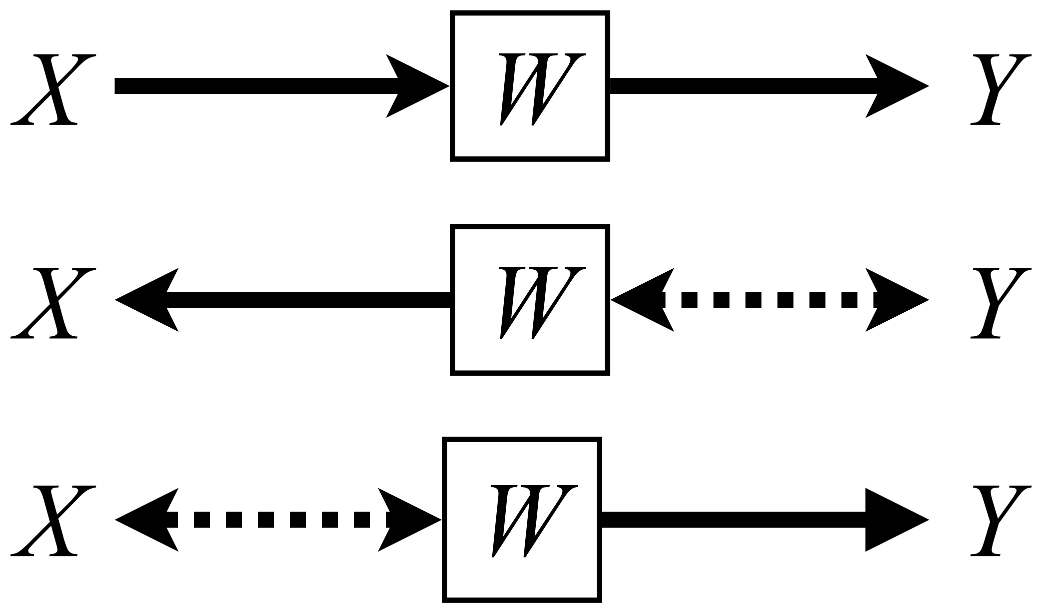
Definition (d-connecting path): A path p between X and Y in a causal diagram G is said to be ***d-connecting*** (or open/active) relative to a (possibly empty) set $\mathbf{Z}$ ($X, Y \notin \mathbf{Z}$) if and only if all triplets in it are active.

Definition (d-separation): A set $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ if and only if there is no d-connecting path between them relative to $\mathbf{Z}$. This is denoted by $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_G$.

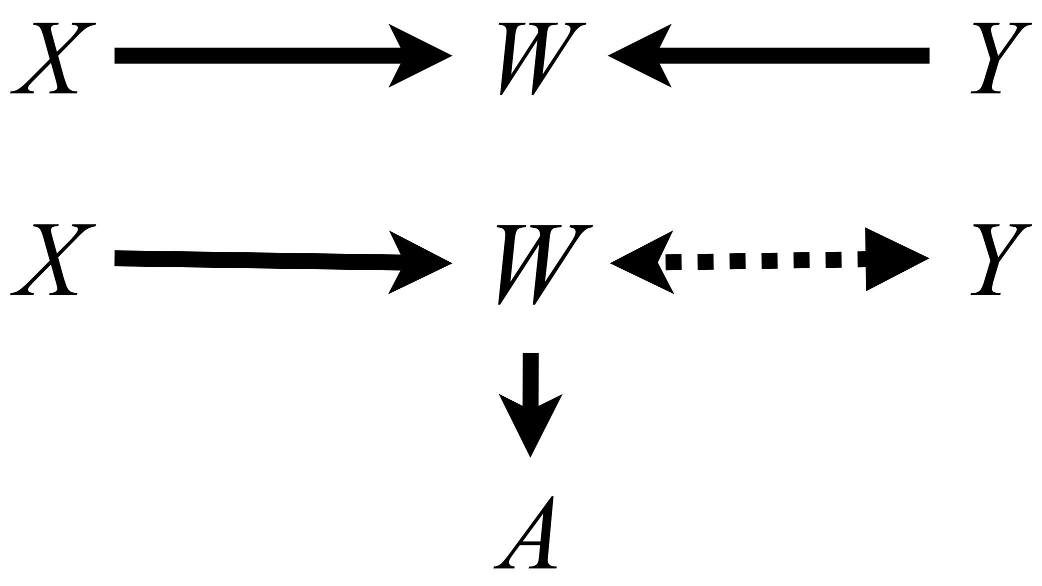
Global Markov property: $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_G \Rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_P$

D-separations in G imply conditional independencies in P

W is non-collider (active if $W \notin \mathbf{Z}$)



W is a collider (active if $W \in \mathbf{Z}$ or any of its descendants is in $\mathbf{Z}$, e.g., $A \in \mathbf{Z}$)



D-Separation and Implied Conditional Independencies

Definition (active triplet): A triplet in a subpath $\langle V_i, V_m, V_j \rangle$ is said to be ***active*** relative to a set $\mathbf{Z}$ if V_m:

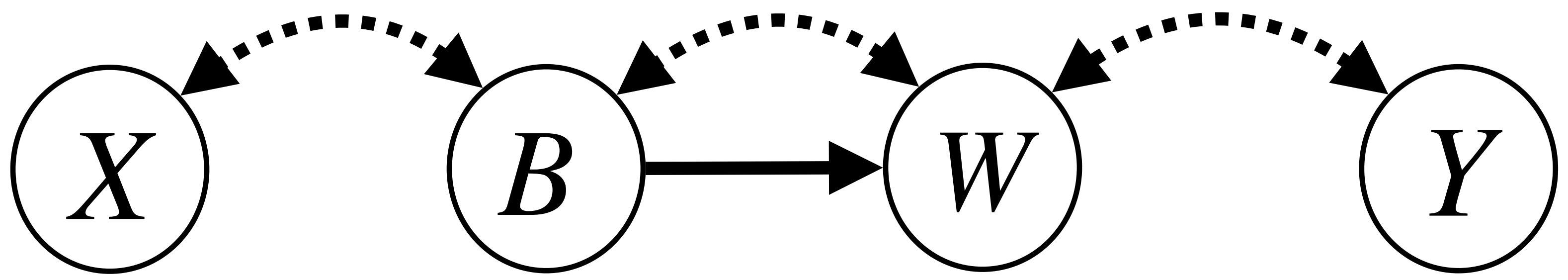
1. Is a non-collider and not a member of $\mathbf{Z}$; or
2. Is a collider and an ancestor of some member of $\mathbf{Z}$.

Definition (d-connecting path): A path p between X and Y in a causal diagram G is said to be ***d-connecting*** (or open/active) relative to a (possibly empty) set $\mathbf{Z}$ $(X, Y \notin \mathbf{Z})$ if and only if all triplets in it are active.

Definition (d-separation): A set $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ if and only if there is no d-connecting path between them relative to $\mathbf{Z}$. This is denoted by $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_G$.

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

D-separations in G imply conditional independencies in P



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

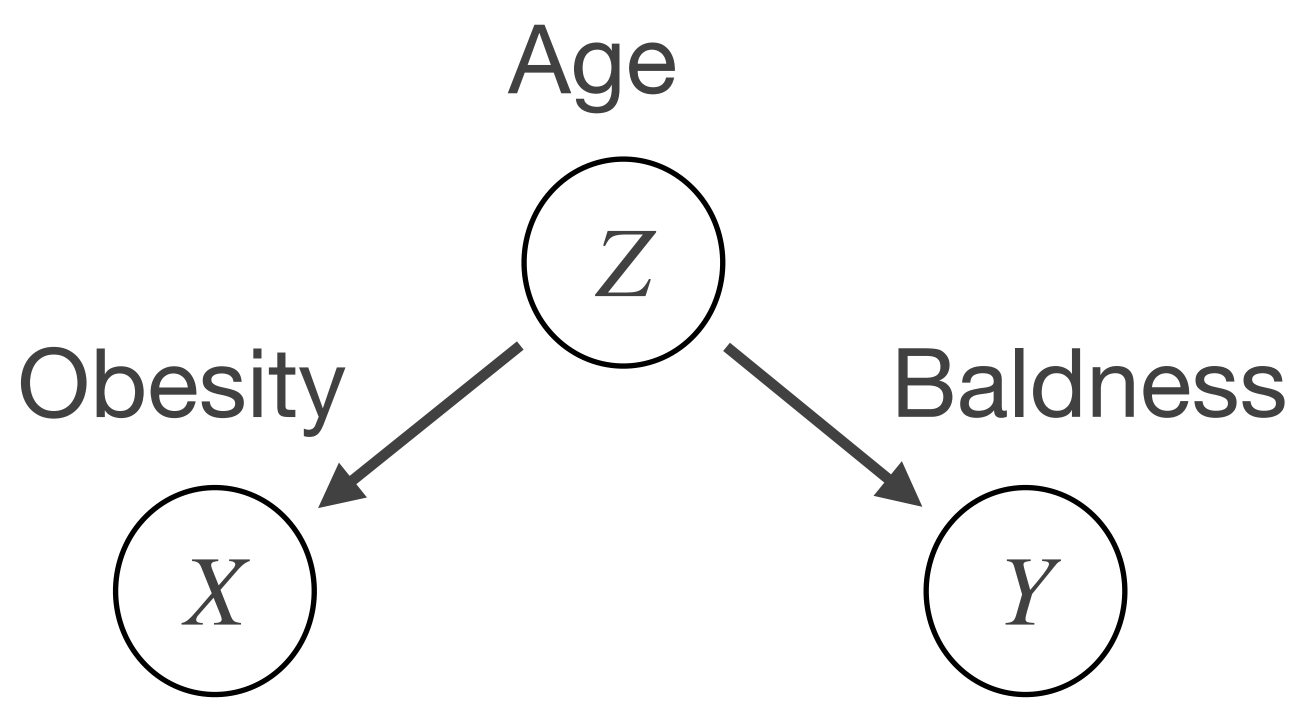
$\mathbf{Z}$: ✓ $\{\}$ ✓ $\{B\}$ ✗ $\{W\}$ ✗ $\{B, W\}$

We have that $(X \perp\!\!\!\perp Y)$ and $(X \perp\!\!\!\perp Y \mid B)$, but $(X \not\perp\!\!\!\perp Y \mid W)$ and $(X \not\perp\!\!\!\perp Y \mid B, W)$

Special Triplets and Formations

Fork

Z **as a common cause**

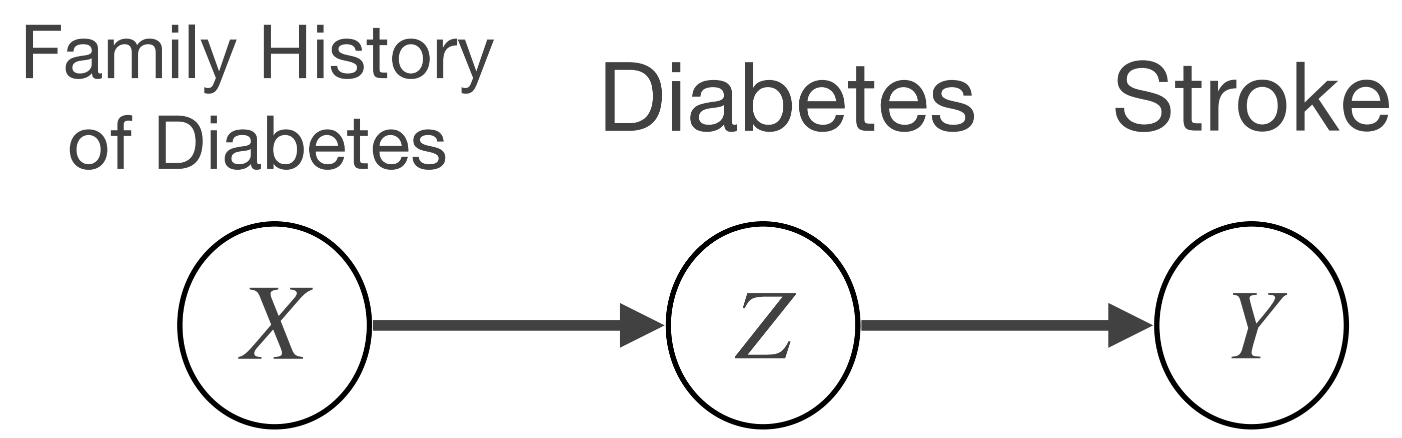


$$X \not\perp\!\!\!\perp Y$$

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

Chain

Z **as a mediator**



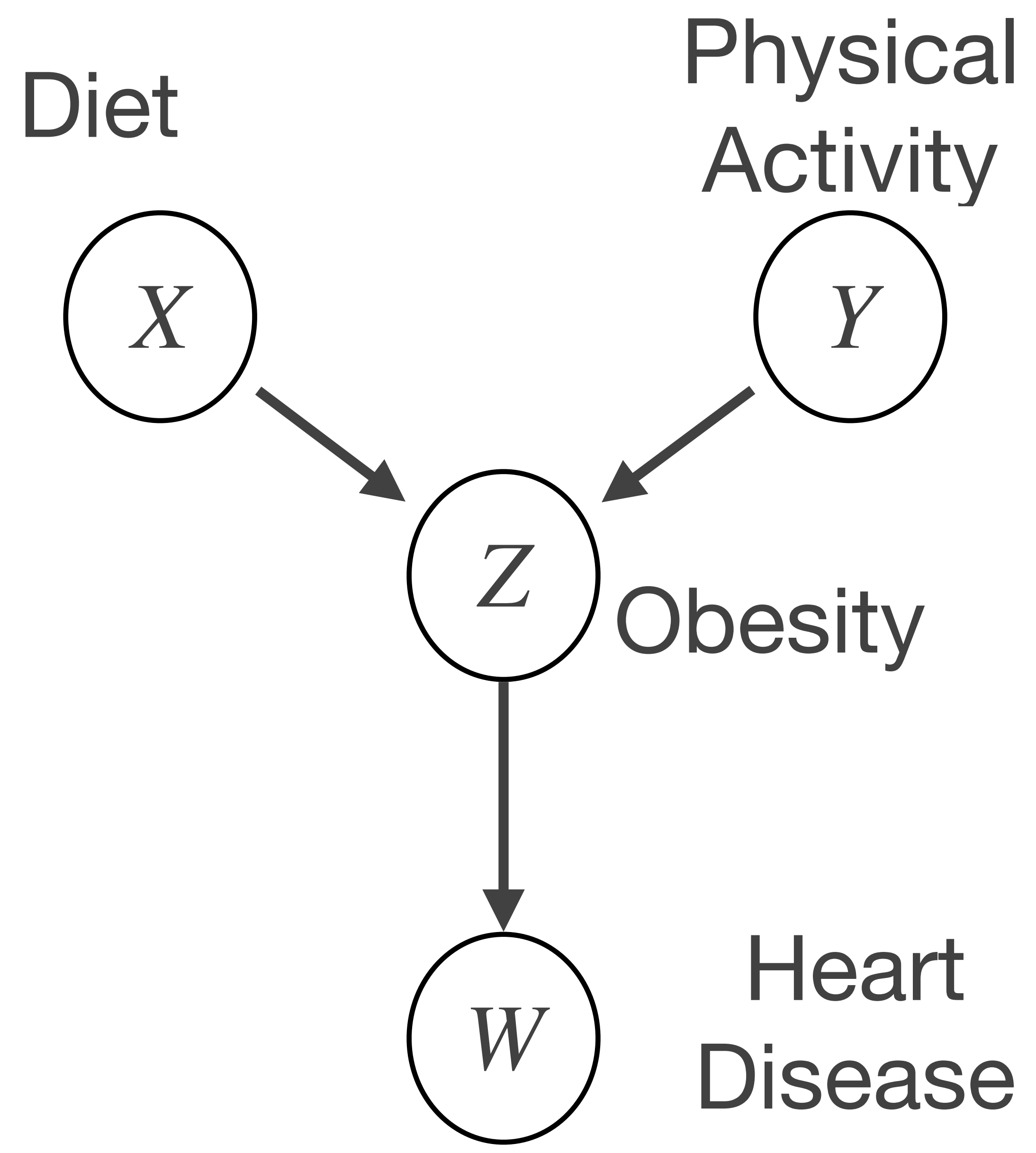
$$X \not\perp\!\!\!\perp Y$$

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

In both cases, Z is a non-collider!

V-Structure

Z **as a collider or common effect**



$$X \perp\!\!\!\perp Y$$

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

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

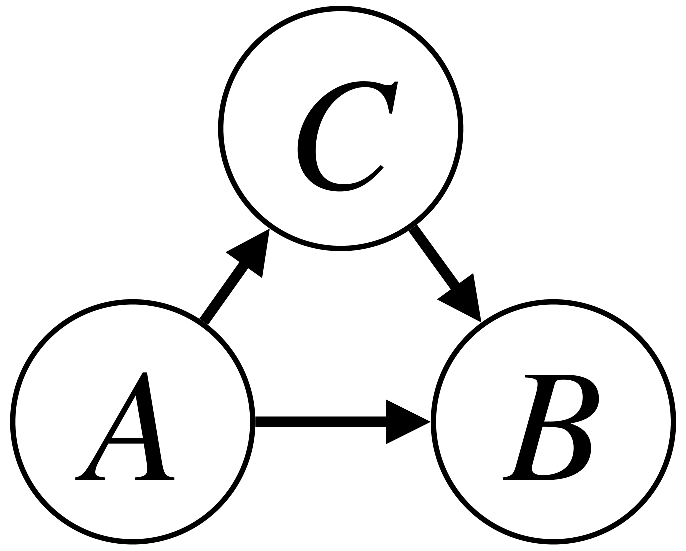
Independence Maps (I-Maps)

Definition (I-Map): If, for a distribution $P(\mathbf{V})$, and any sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, it holds that

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

then G is an ***I-Map of*** P.

And, P is Markov Relative to G



G is I-Map of any $P(A, B, C)$

A complete graph trivially satisfies any distribution

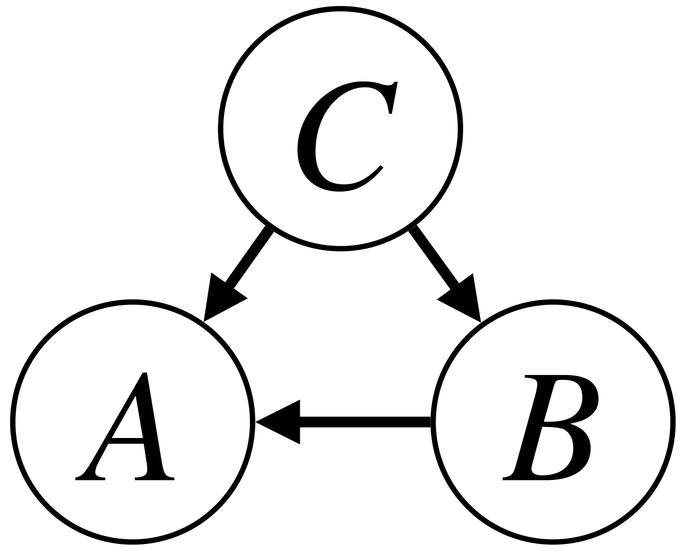
Minimal I-Maps and Bayesian Networks

Definition (Minimal I-Map): If G is an I-Map of P and none of its edges can be removed without ceasing its I-Map property of P, then G is a ***minimal I-Map of*** P.

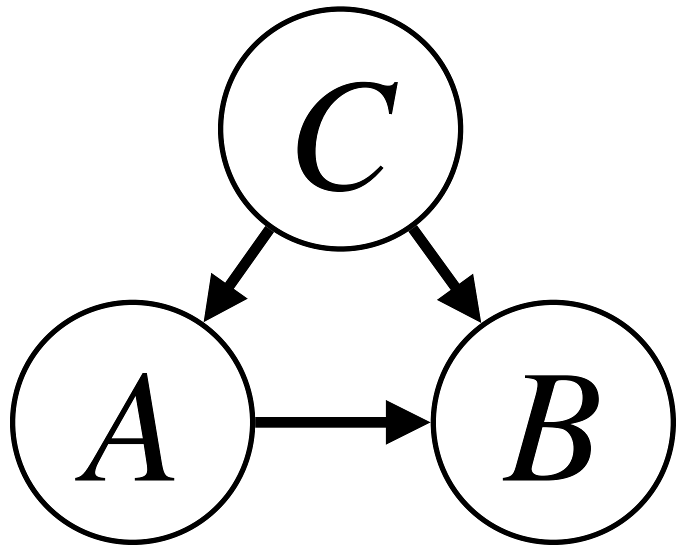
Definition (Bayesian Network, BN for short): A a Bayesian Network is a directed acyclic graph (DAG) or acyclic directed mixed graph (ADMG) G over $\mathbf{V}$ that is a minimal I-map of $P(\mathbf{V})$.

Consider $P(A, B, C)$ with $A \perp\!\!\!\perp C$ being the **only** independence relation.

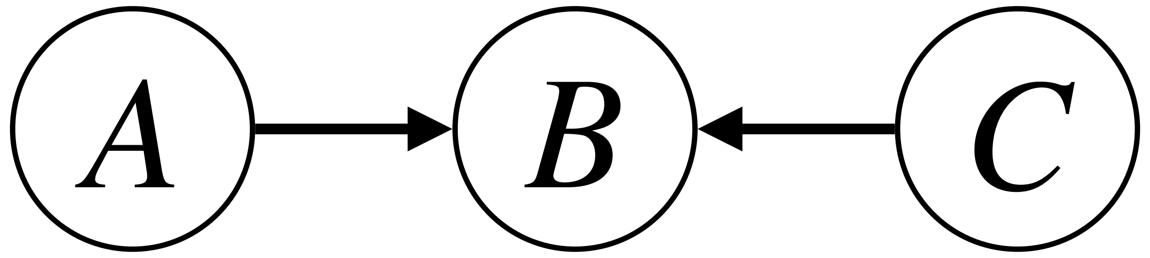
$A \perp\!\!\!\perp C$
$A \not\!\perp\!\!\!\perp C \mid B$
$A \not\!\perp\!\!\!\perp B$
$A \not\!\perp\!\!\!\perp B \mid C$
$B \not\!\perp\!\!\!\perp C$
$B \not\!\perp\!\!\!\perp C \mid A$



G_1 is I-Map of $P(A, B, C)$



G_2 is I-Map of $P(A, B, C)$

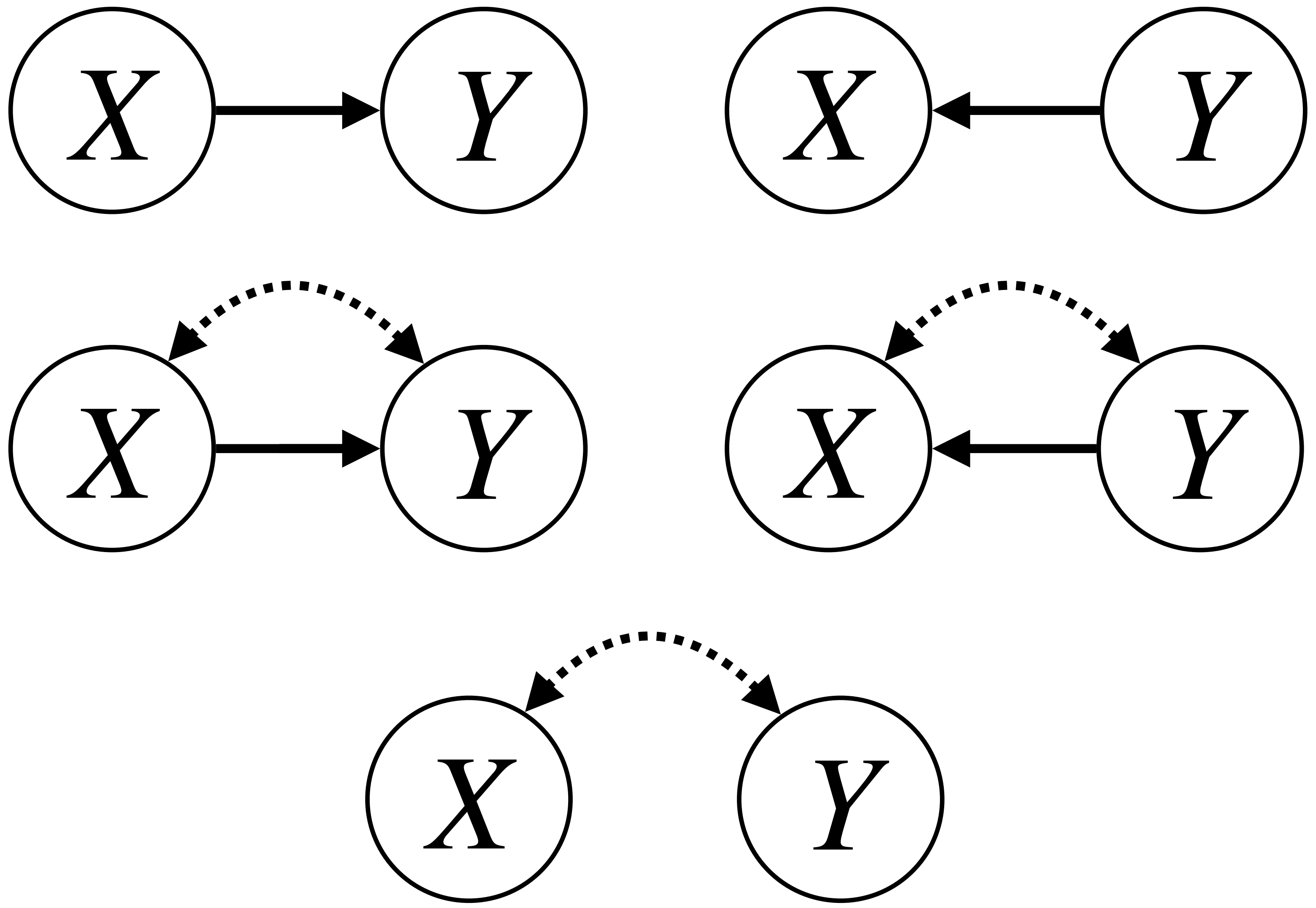


G_3 is minimal I-Map of $P(A, B, C)$

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.

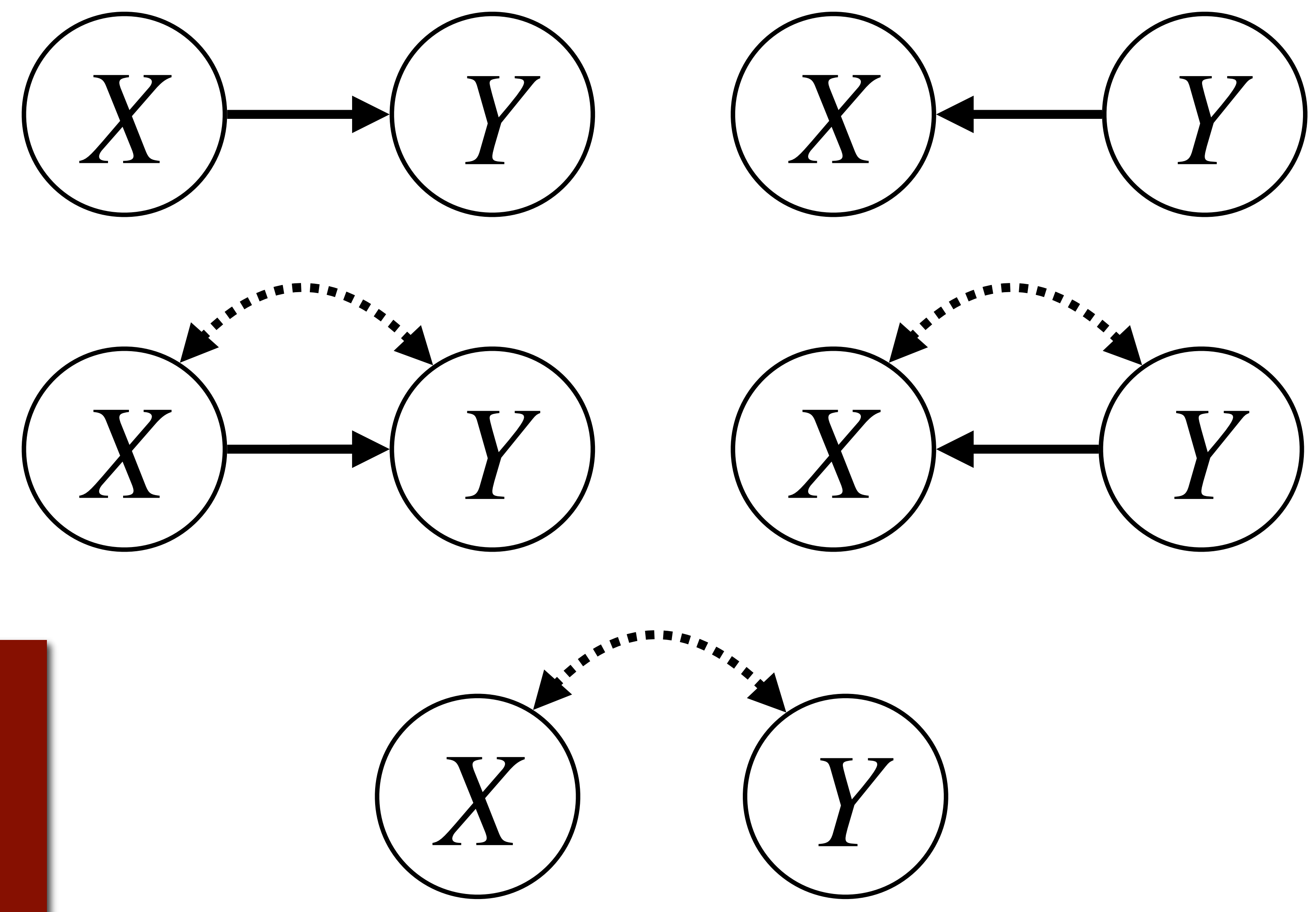
Markov Equivalent BNs

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

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 $P(Y \mid X) \neq P(Y)$ 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

$$P(X, Y, Z)$$

with $P(Y \mid X, Z) = P(Y \mid X)$

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

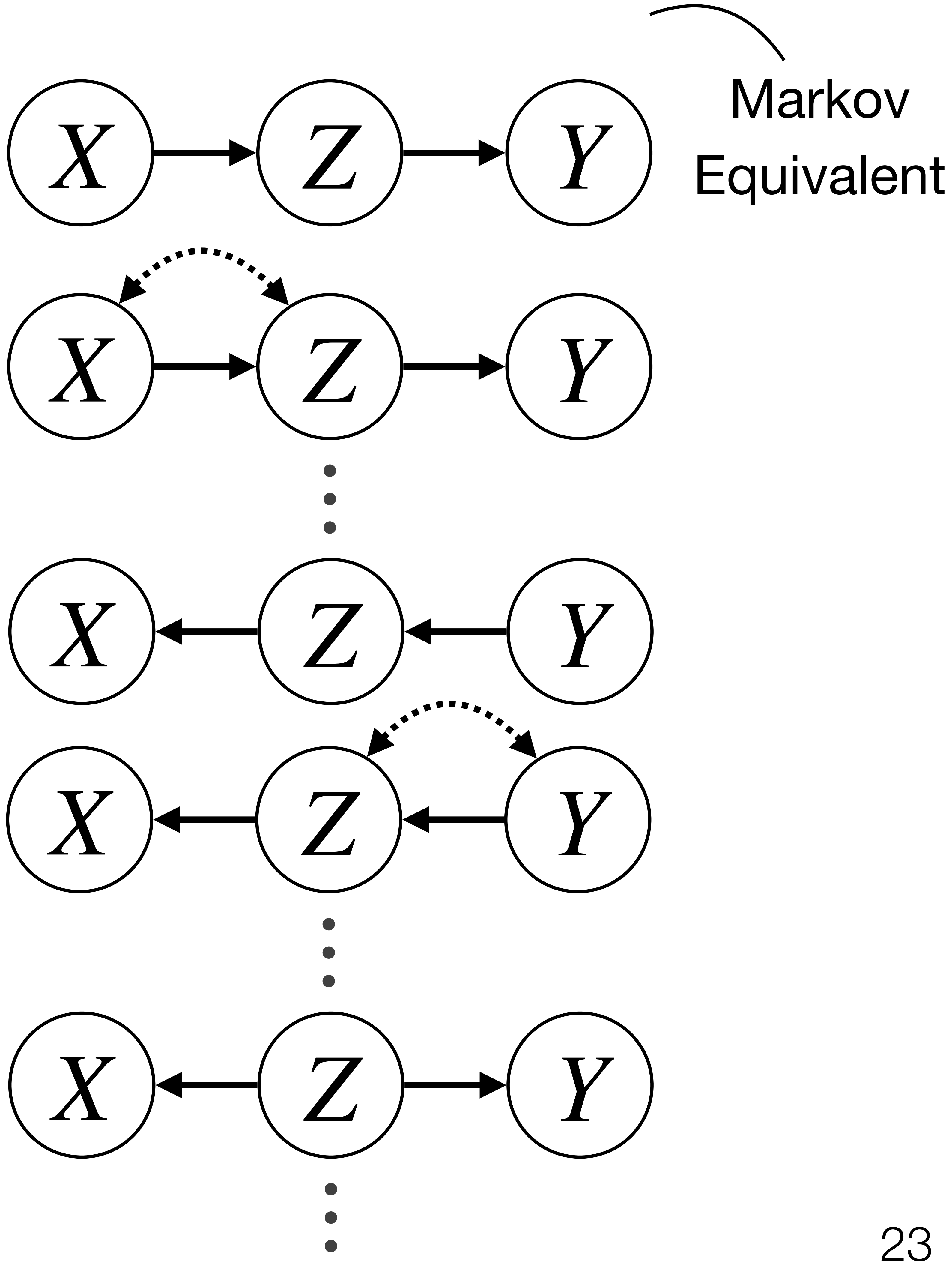
Factorization

$$\begin{aligned} P(x, y, z) &= P(y \mid x, z)P(z \mid x)P(x) \\ &= P(y \mid z)P(z \mid x)P(x) \end{aligned}$$

$$\begin{aligned} P(x, y, z) &= P(x \mid y, z)P(y \mid z)P(z) \\ &= P(x \mid z)P(z \mid y)P(y) \end{aligned}$$

$$\begin{aligned} P(x, y, z) &= P(y \mid x, z)P(x \mid z)P(z) \\ &= P(y \mid z)P(x \mid z)P(z) \end{aligned}$$

Bayesian Networks



Markov Equivalence Class

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

$P(X, Y, Z)$

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

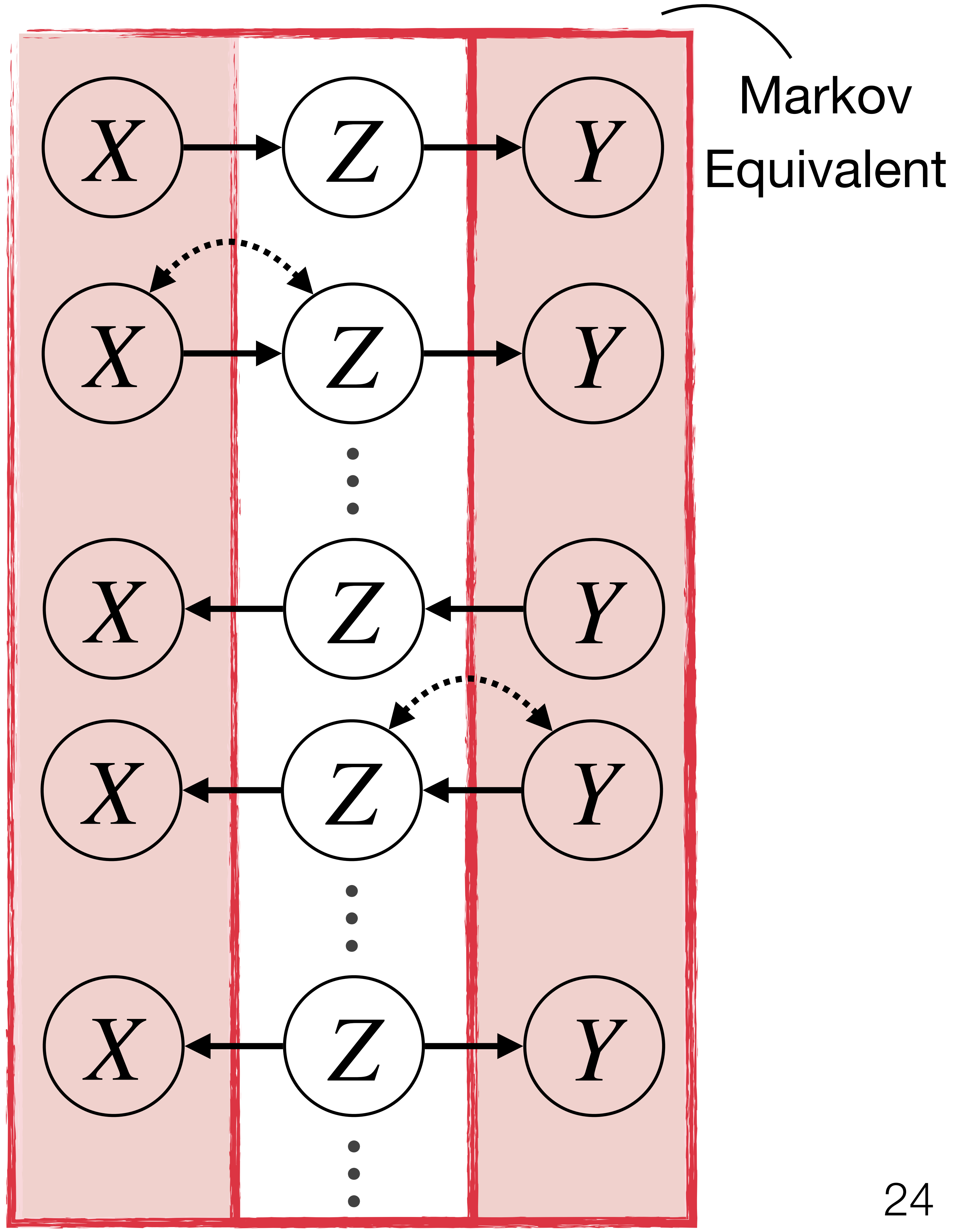
i.e., $X \perp\!\!\!\perp Y | Z$

$$P(x, y, z) = {\color{red}P(y|x, z)}P(z|x)P(x)$$
$$= {\color{red}P(y|z)}P(z|x)P(x)$$

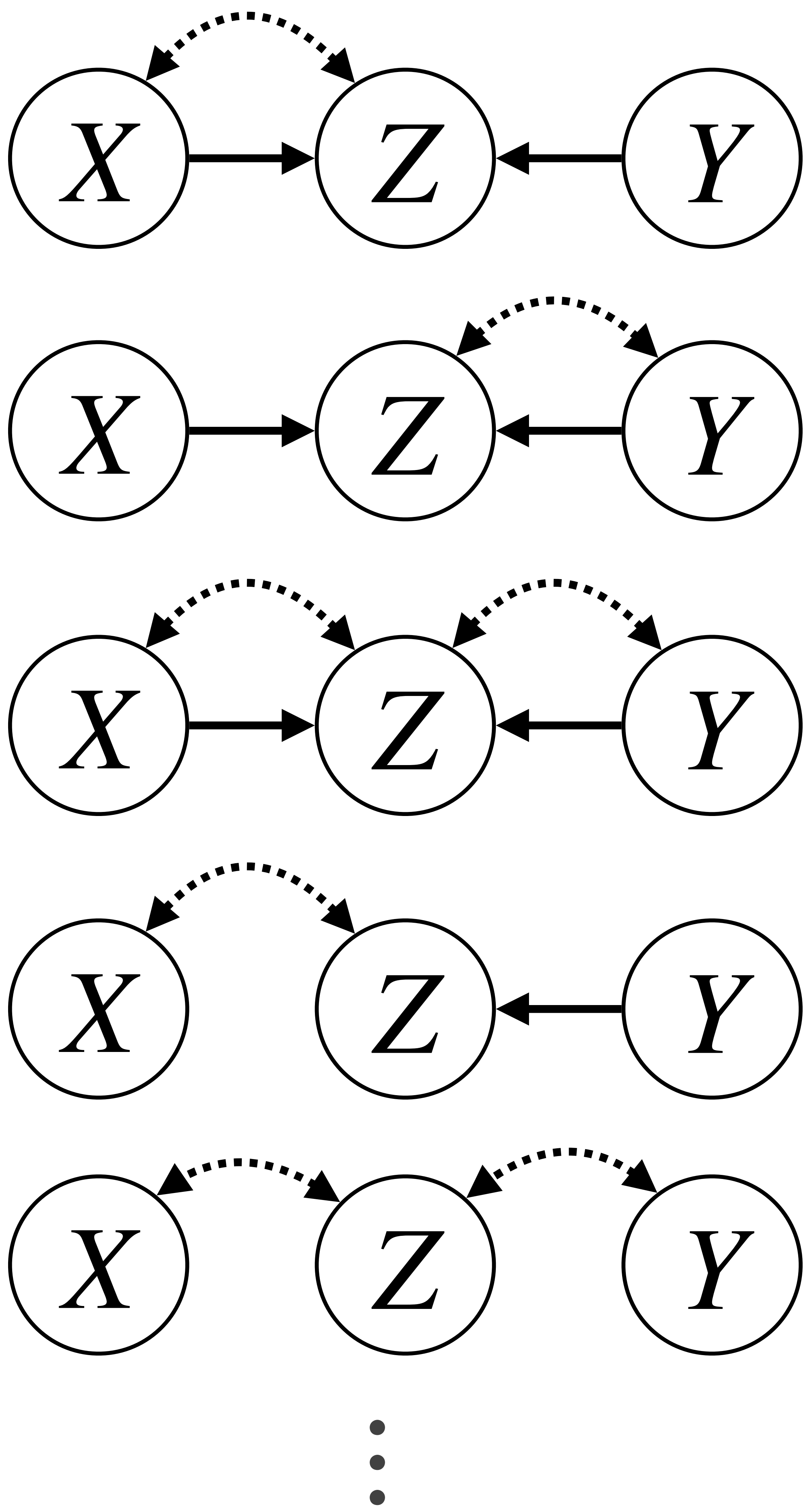
$$P(x, y, z) = {\color{red}P(x|y, z)}P(y|z)P(z)$$
$$= {\color{red}P(x|z)}P(z|y)P(y)$$

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

$$= {\color{red}P(y|z)}P(x|z)P(z)$$



Markov Equivalence Class

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

Markov Equivalent

Markov Equivalence Class

Distribution

$$P(X, Y, Z)$$

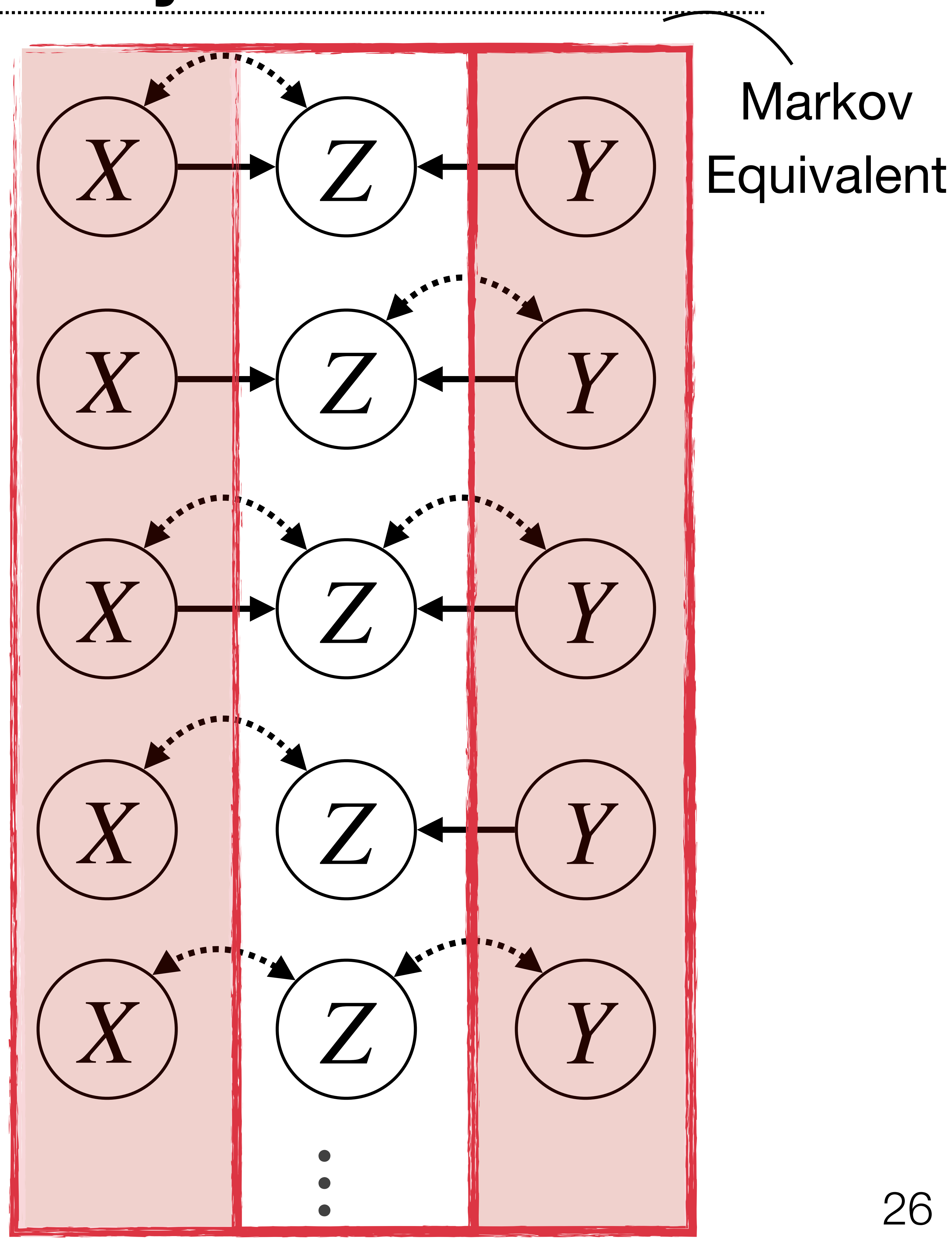
with $P(Y|X) = P(Y)$

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

Factorization

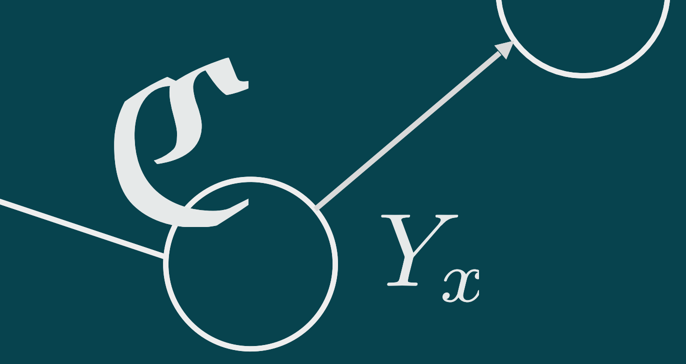
$$\begin{aligned} P(x, y, z) &= P(z|x, y)\color{red}{P(x|y)}P(y) \\ &= P(z|x, y)\color{red}{P(x)}P(y) \end{aligned}$$

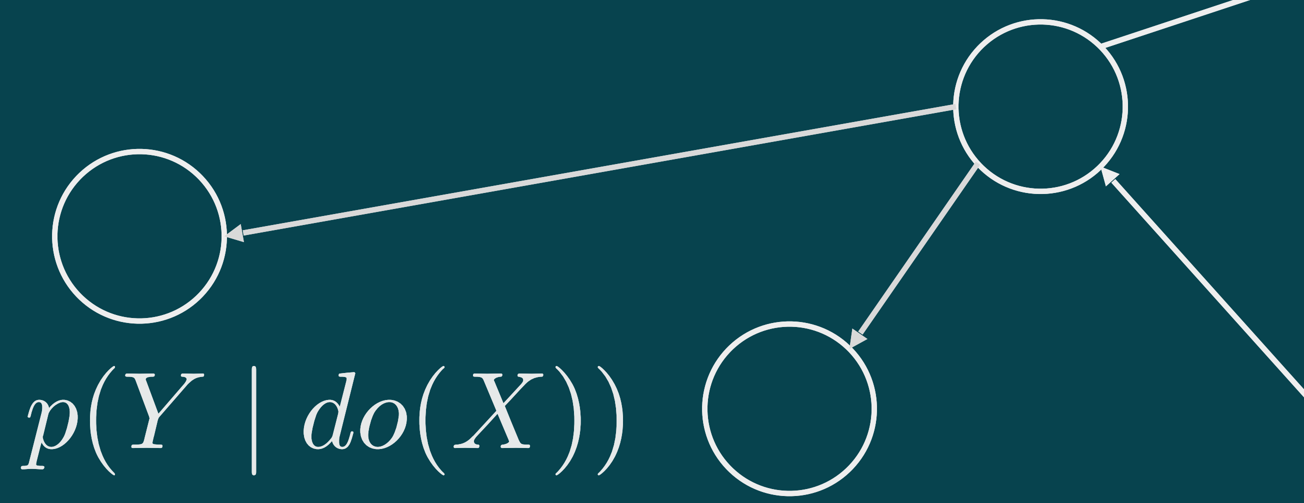
Bayesian Networks



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

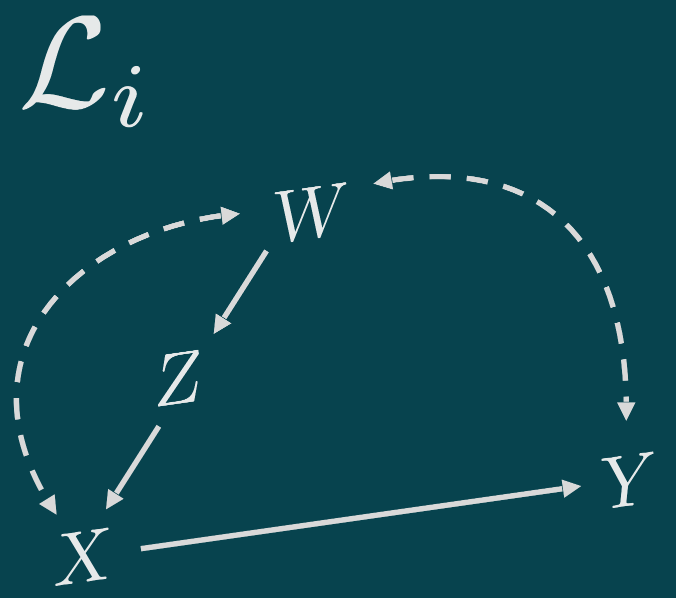
Note: Z is never an ancestor of X or Y





3 Representing Markov Equivalence Classes

Partial Ancestral Graphs



Inducing Paths

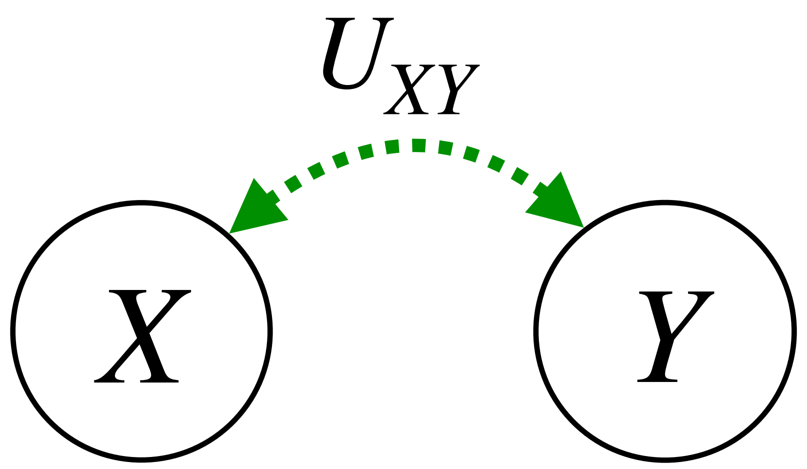
For an Bayesian Network over $\mathbf{V} \cup \mathbf{U} \cup \mathbf{S}$, where

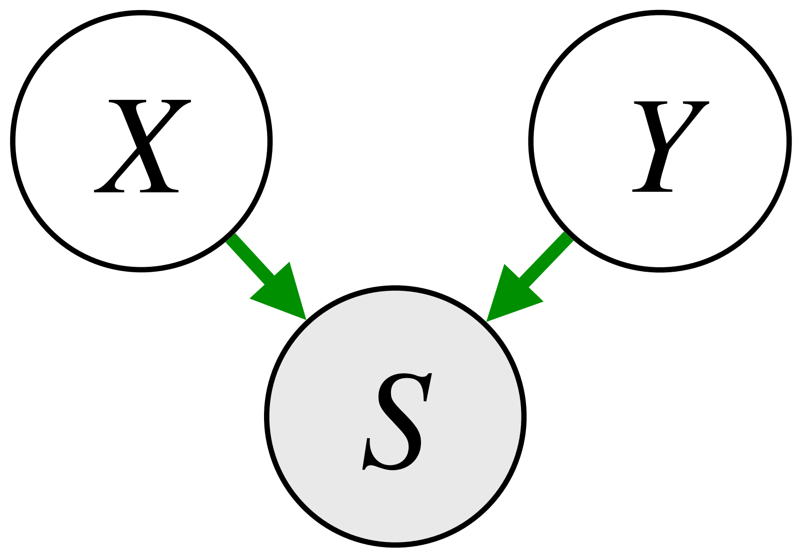
V: set of observed variables

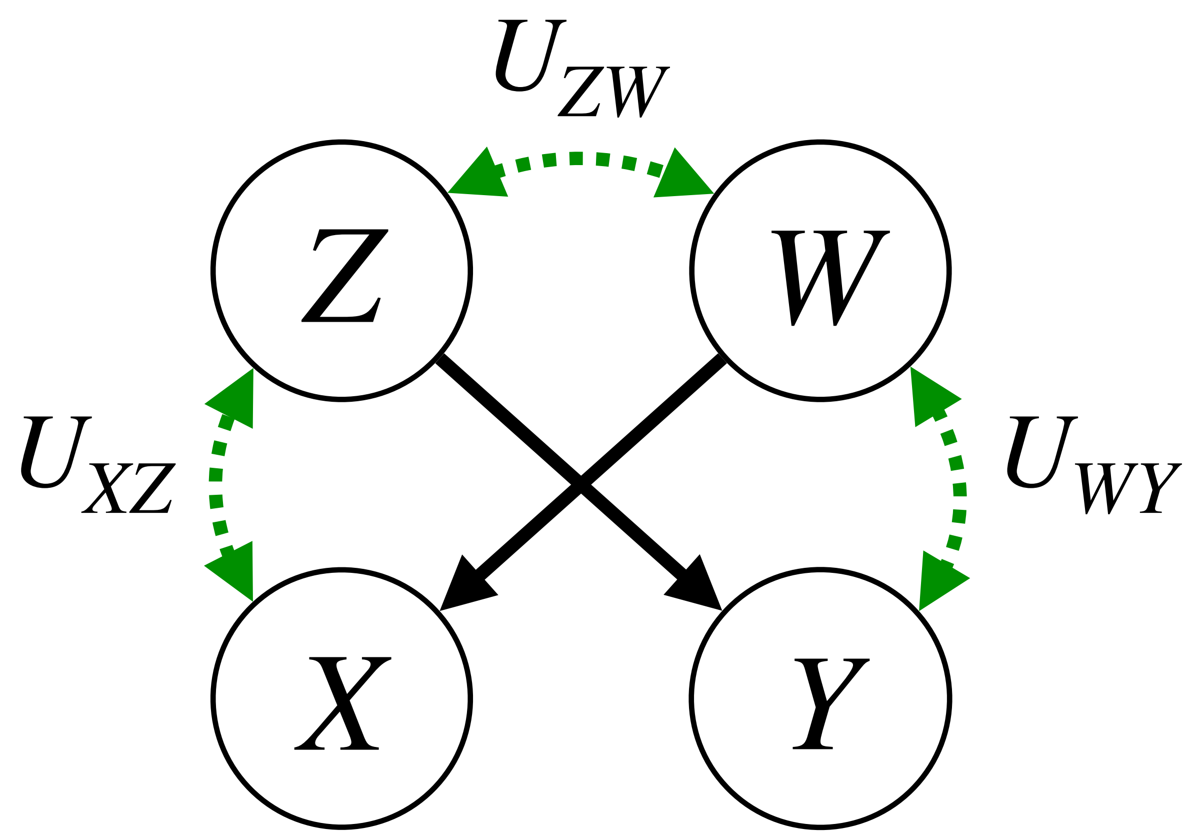
U: set of latent or unobserved variables, and

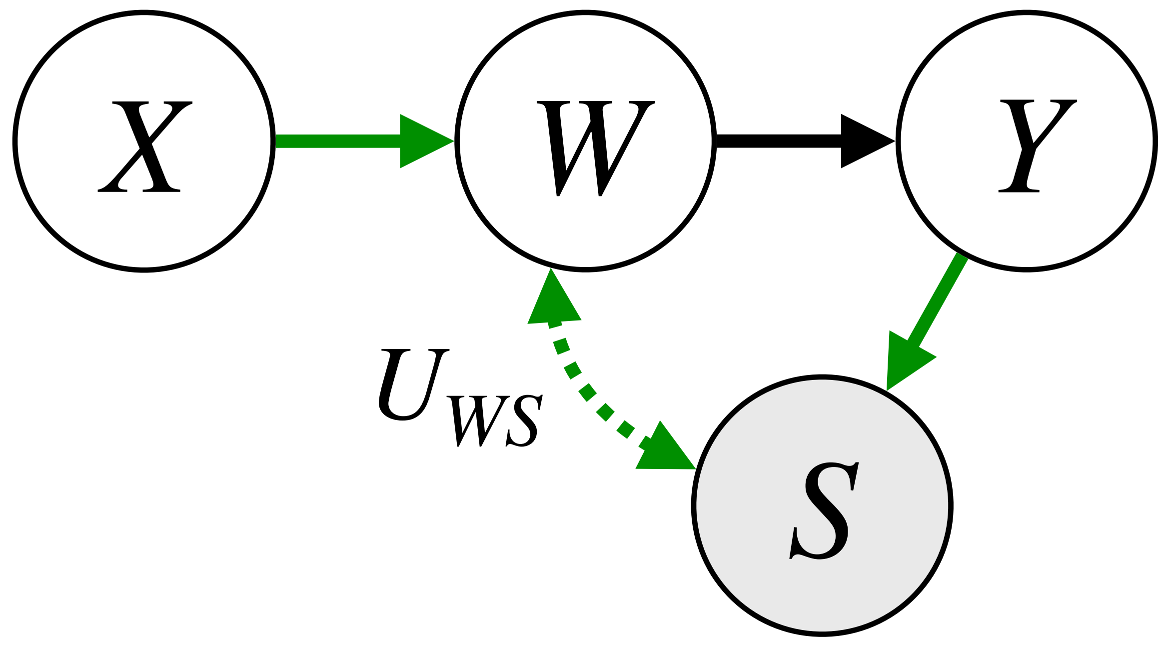
S: set of unobserved selection variables,

A path *p* between X and Y is called an **inducing path** if it every non-endpoint vertex on *p* is a collider that is either an ancestor of X or Y, or a member of $\mathbf{S}$.









$X, Y, Z, W \in \mathbf{V}$

$S \in \mathbf{S}$

$U_{XY}, U_{XZ}, U_{ZW}, U_{WY}, U_{WS} \in \mathbf{U}$

X and Y are not separable when there is an inducing path connecting them!

Partial Ancestral Graphs

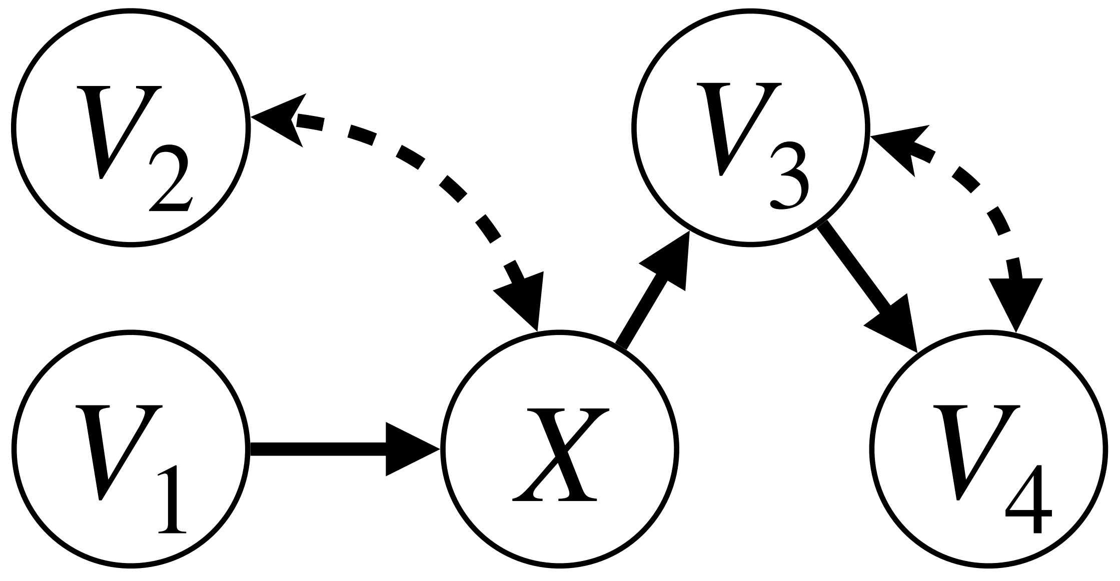
A **partial ancestral graph (PAG)** for a BN G is a graph $\mathscr{P}$ with six kinds of edges ($-$, $\rightarrow$, $\leftrightarrow$, $\circ\!-$, $\circ\!-\!\circ$, $\circ\!\!\rightarrow$), such that

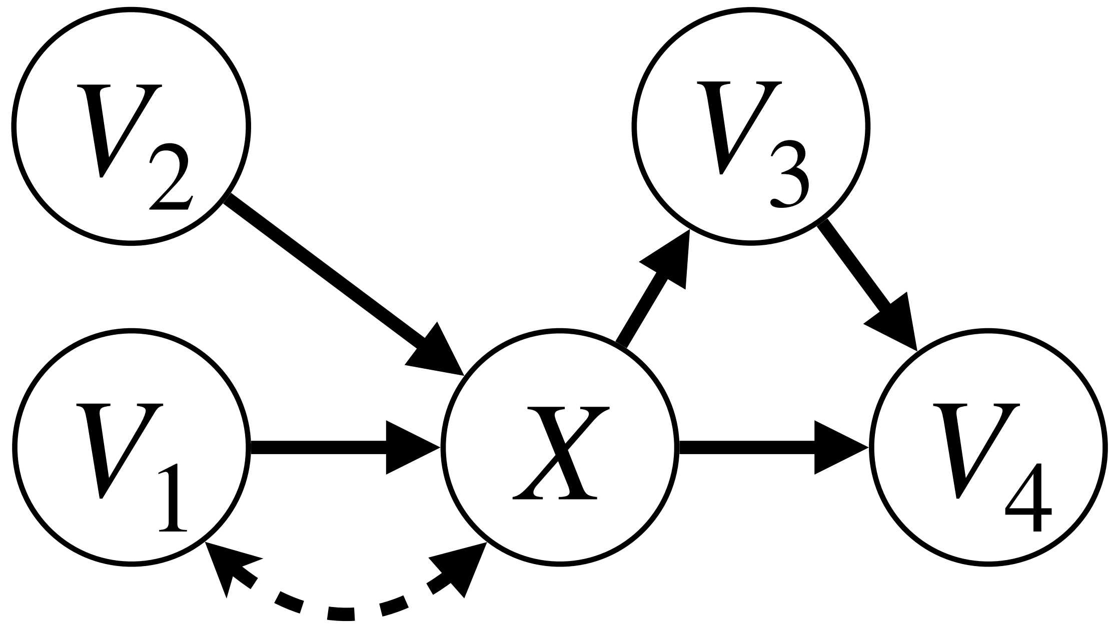
(1) Every edge in $\mathscr{P}$ corresponds to an inducing path in any member of the MEC of G;

(2) Every non-circle edge mark represents an **invariant ancestral or non-ancestral relationship** in the MEC of G

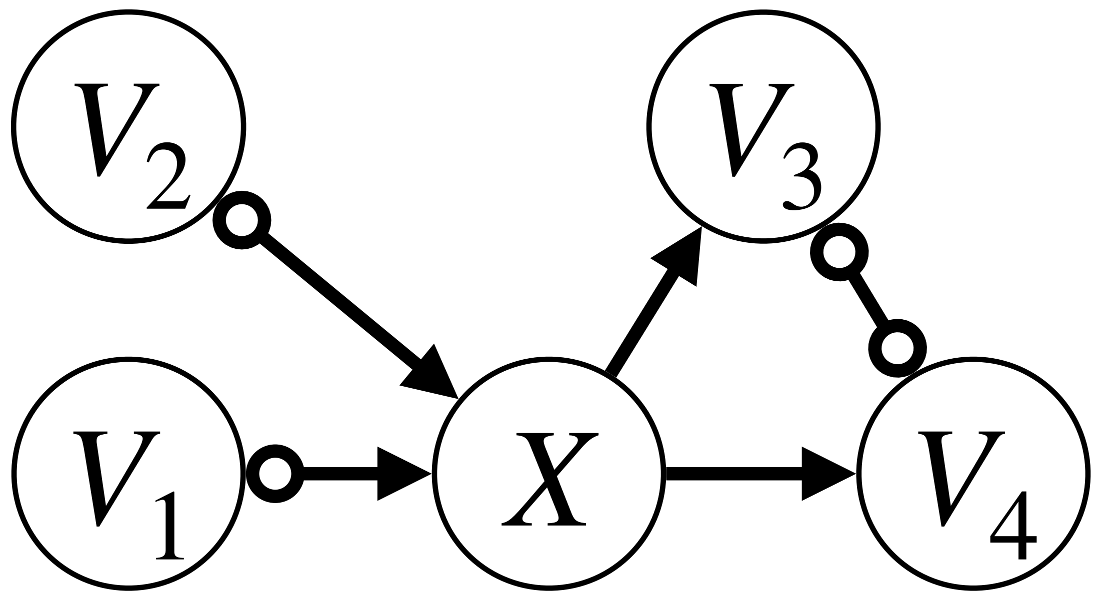
Arrowhead $\Longrightarrow$ non-ancestrality

Tail $\Longrightarrow$ ancestrally





PAG

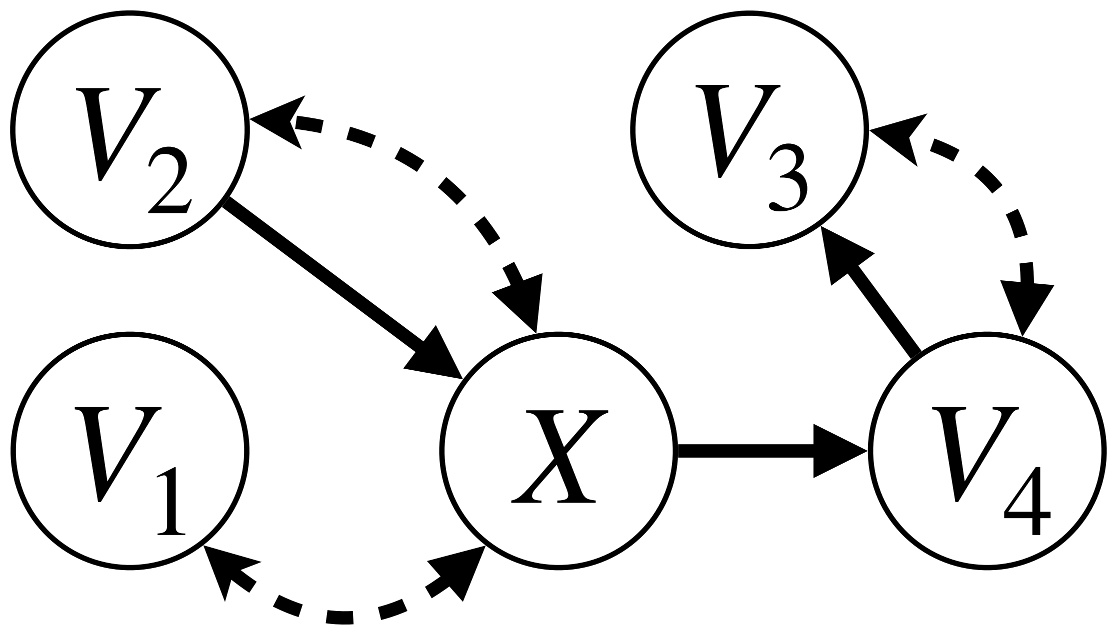


$X \longrightarrow V_3 \Longrightarrow X$ ancestral of V_3

$X \longrightarrow V_4 \Longrightarrow X$ ancestral of V_4

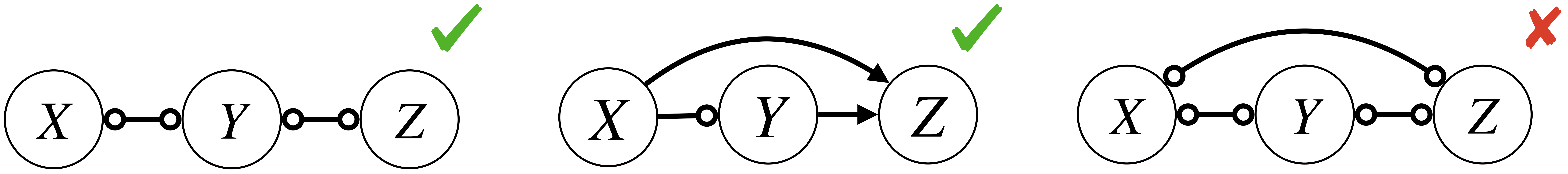
$V_1 \circ\!\!\longrightarrow X \Longrightarrow X$ non-ancestral of V_1

$V_2 \circ\!\!\longrightarrow X \Longrightarrow X$ non-ancestral of V_2

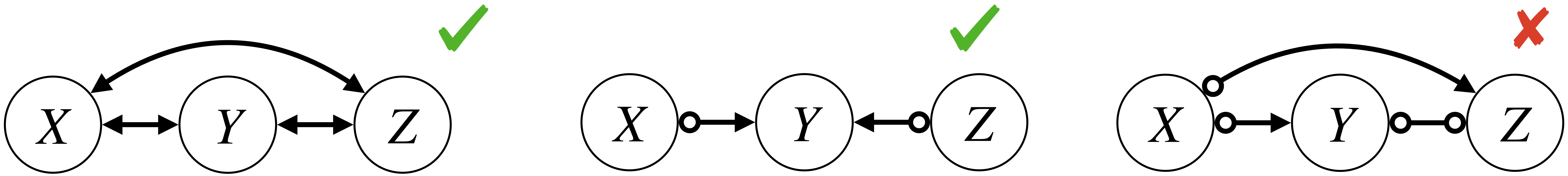


Definite Triplets in PAGs

Definition (definite non-collider): A node Y of a triplet $\langle X, Y, Z\rangle$ in a PAG is a definite non-collider if the edge between X and Y or the edge between Y and Z is out of Y, or both edges have a circle mark at Y and X and Z are not adjacent.



Definition (collider): As before, a node Y of a triplet $\langle X, Y, Z\rangle$ in a PAG is a definite collider if both the between X and Y and the edge between Y and Z are into Y.



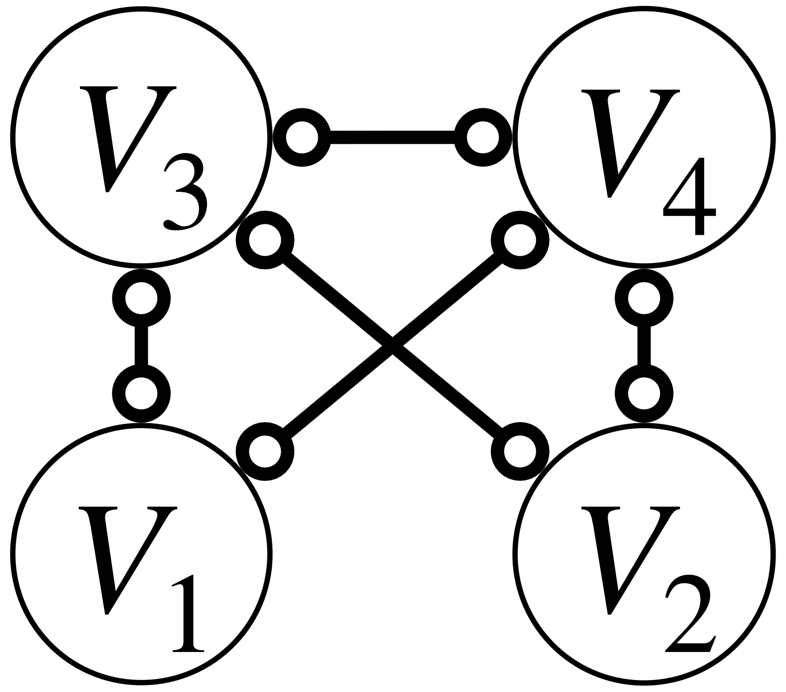
M-Separation in PAGs

Definition (definite m-connecting path): In a PAG, a path p between X and Y is a *definite m-connecting path* relative to a (possibly empty) set $\mathbf{Z}$ ($X, Y \notin \mathbf{Z}$) if every non-endpoint vertex on p is either a definite non-collider or a collider and:

i. Every definite non-collider on p is not a member of $\mathbf{Z}$;

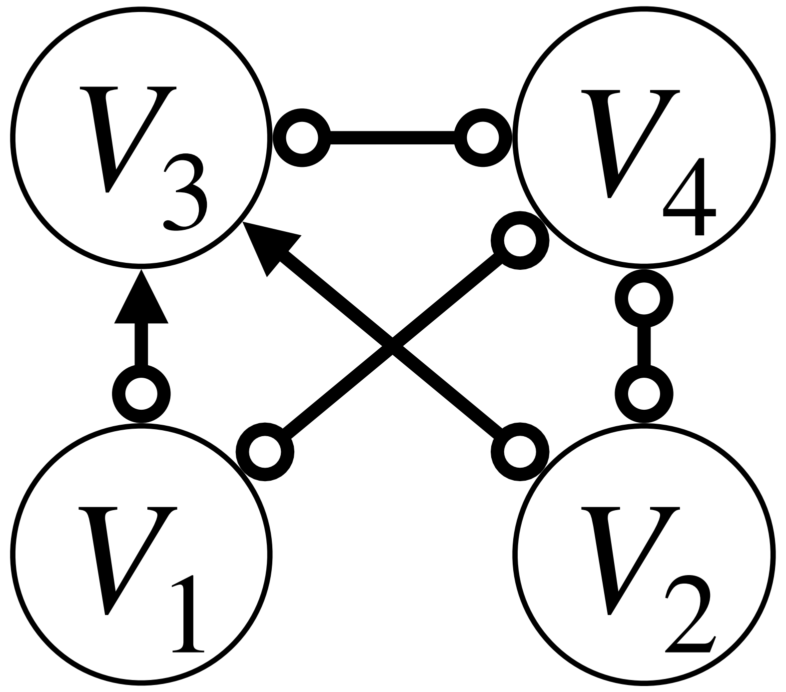
ii. Every collider on p is an ancestor of some member of $\mathbf{Z}$.

Definition (m-separation): In a PAG, X and Y are m-separated by $\mathbf{Z}$ if there is no definite m-connecting path between them relative to $\mathbf{Z}$.



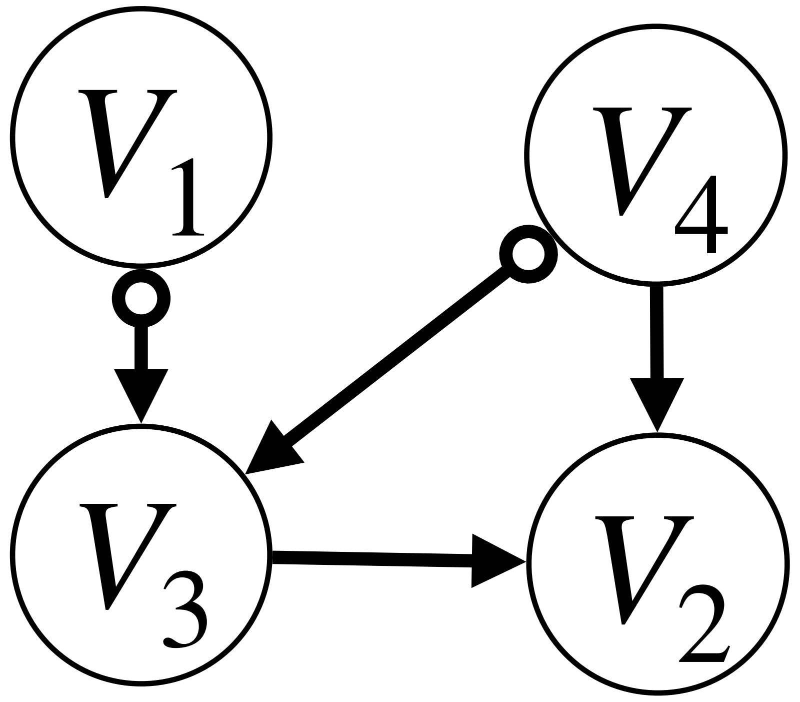
$V_1 \perp\!\!\!\perp V_2 \mid V_3, V_4$

$V_1 \not\perp\!\!\!\perp V_2 \mid V_3$ $\quad$ $V_1 \not\perp\!\!\!\perp V_2 \mid V_4$



$V_1 \perp\!\!\!\perp V_2 \mid V_4$

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



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

$V_1 \not\perp\!\!\!\perp V_4 \mid V_3$

M-Separation in PAGs

Definition (definite m-connecting path): In a PAG, a path p between X and Y is a *definite m-connecting path* relative to a (possibly empty) set $\mathbf{Z}$ ($X, Y \notin \mathbf{Z}$) if every non-endpoint vertex on p is either a definite non-collider or a collider and:

i. Every definite non-collider on p is not a member of $\mathbf{Z}$;

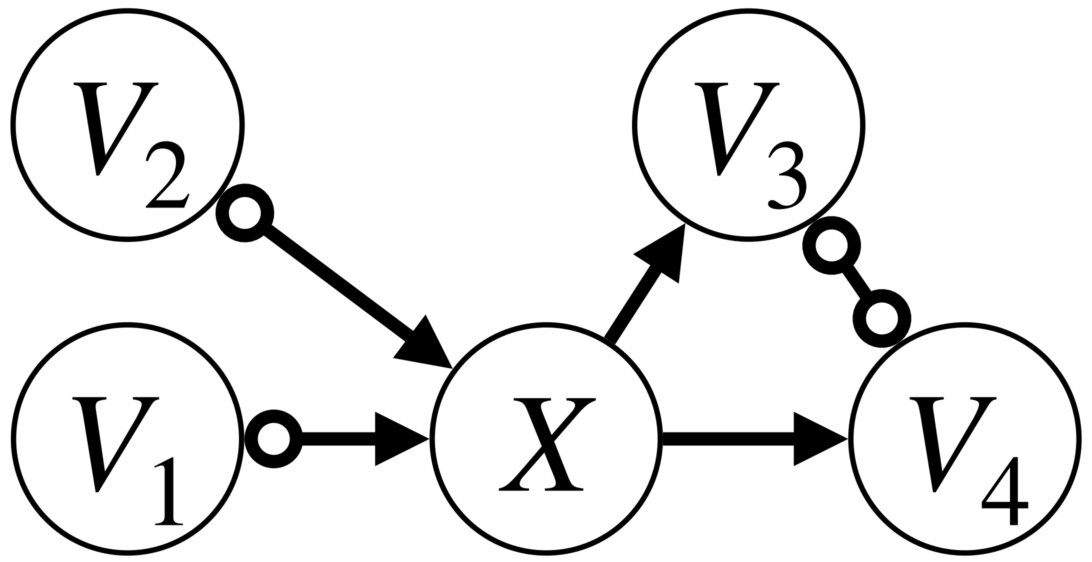
ii. Every collider on p is an ancestor of some member of $\mathbf{Z}$.

Definition (m-separation): In a PAG, X and Y are m-separated by $\mathbf{Z}$ if there is no definite m-connecting path between them relative to $\mathbf{Z}$.

A PAG $\mathscr{P}$ represents all BNs the **Markov Equivalence Class (MEC)**, i.e.:

$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z})_{\mathscr{P}} \Leftrightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z})_G$$

PAG $\mathscr{P}$



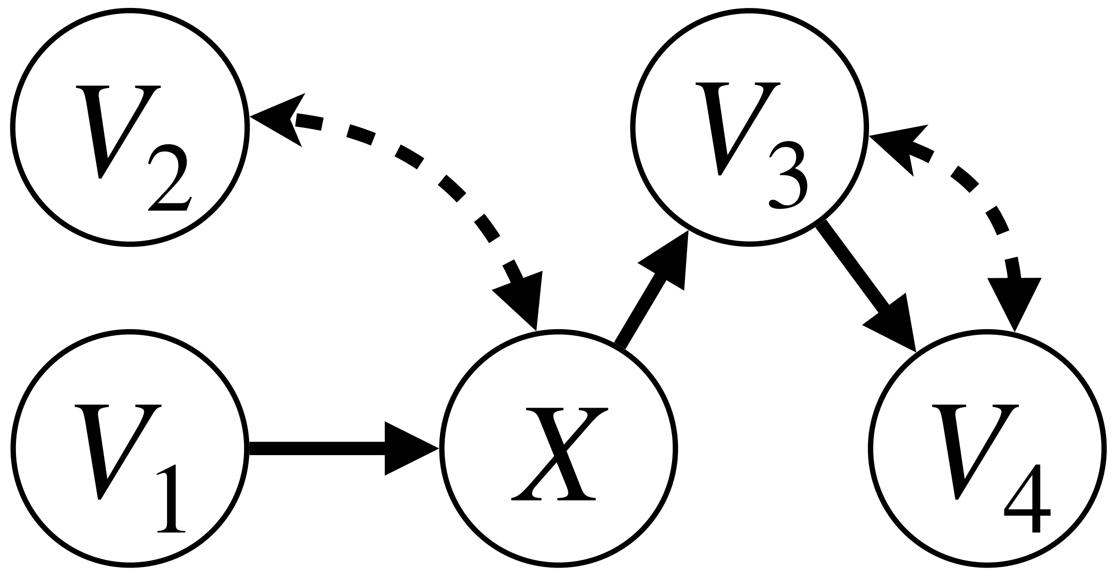
$V_1 \perp\!\!\!\perp V_2$

$V_1 \perp\!\!\!\perp V_3 \,|\, X$

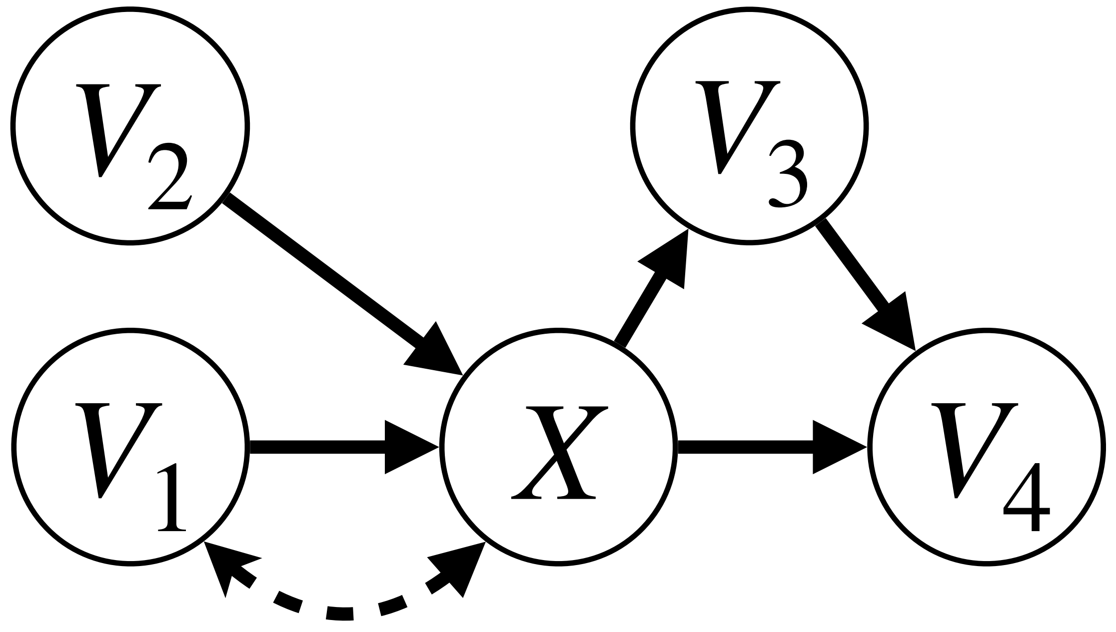
$V_1 \perp\!\!\!\perp V_4 \,|\, X$

$V_2 \perp\!\!\!\perp V_3 \,|\, X$

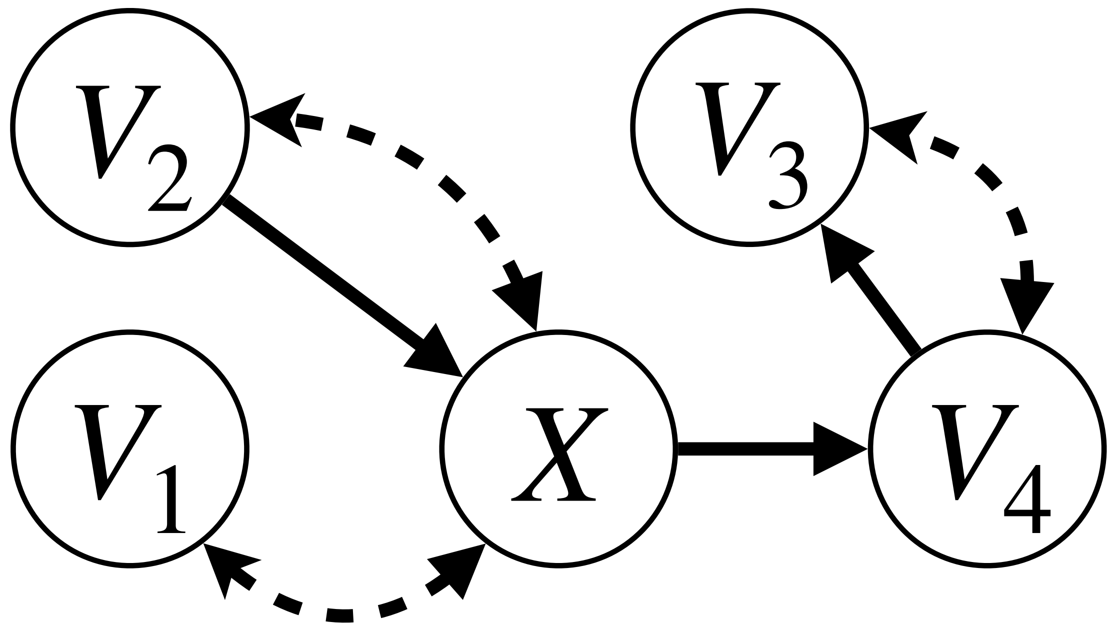
$V_2 \perp\!\!\!\perp V_4 \,|\, X$



BN G_1



BN G_2



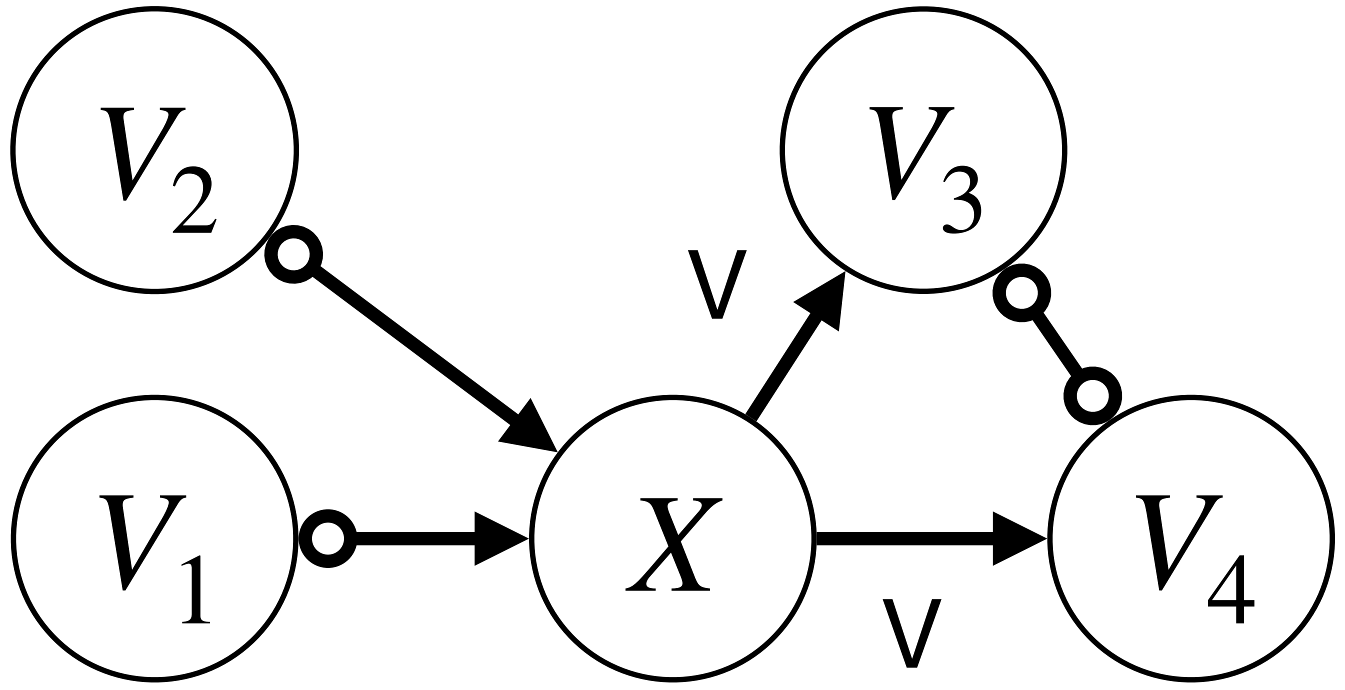
BN G_3

⋮

Edge Visibility

Definition (Visibility of an Edge): Given a PAG $\mathscr{P}$, a directed edge $X \to Y$ is visible if:

1. there is a node V not adjacent to Y such that there is an edge between V and X that is into X, or
2. if there is a collider path from V to X that is into X and every non-endpoint node on the path is a parent of Y. Otherwise, $X \to Y$ is said to be invisible.



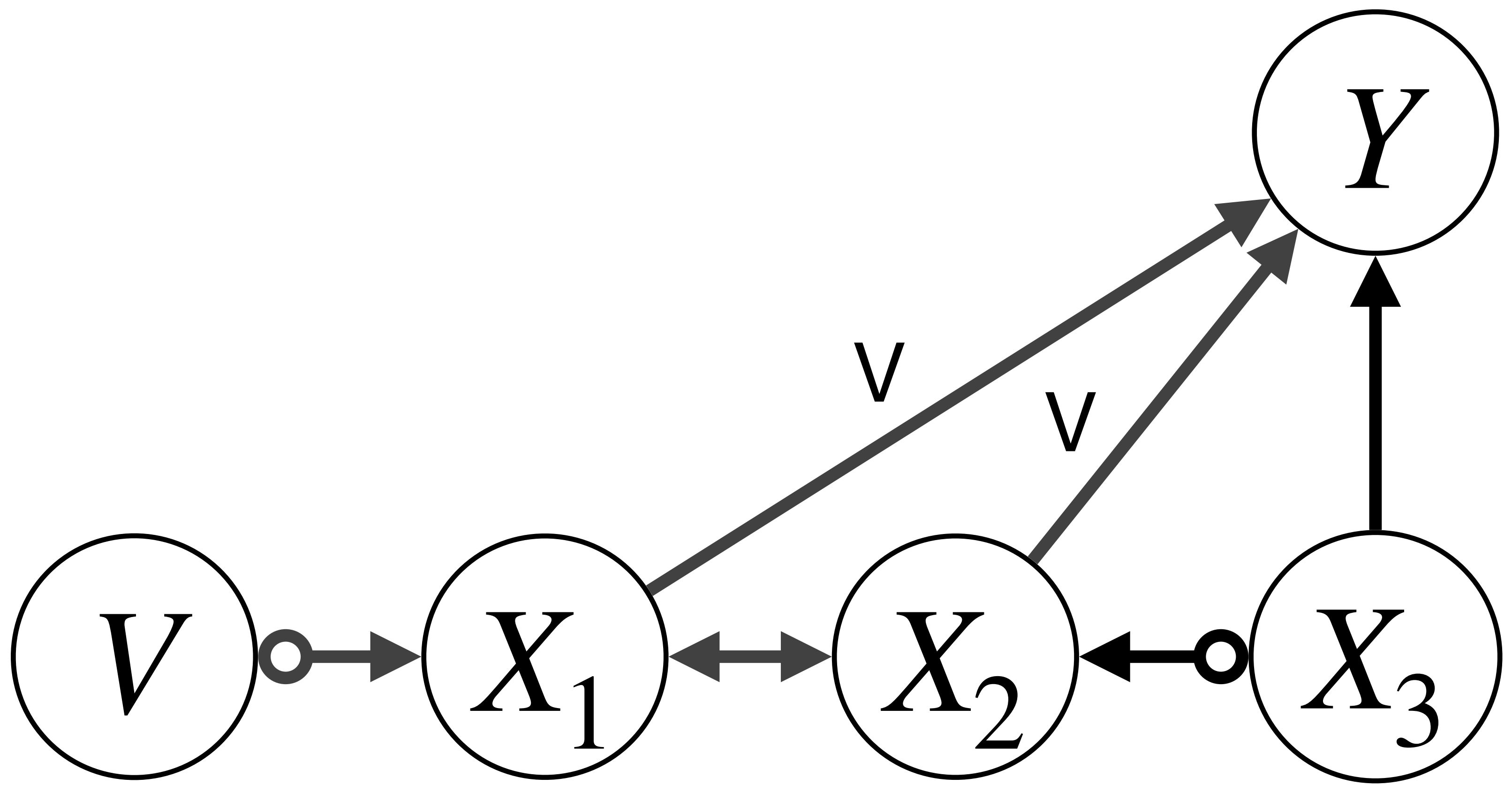
Edges labeled with ‘v’ are referred to as visible edges

Visibility of an edge denotes the absence of a hidden confounder in every member of the equivalence class.

Edge Visibility

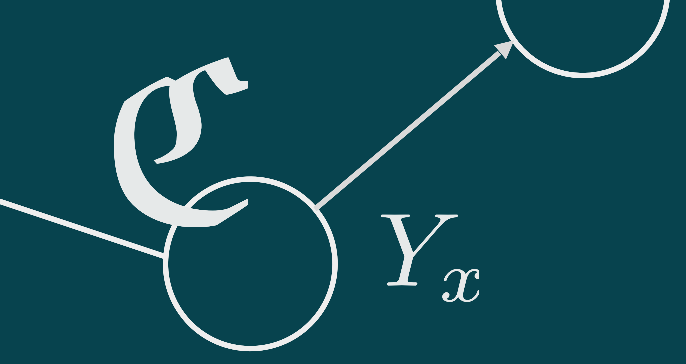
Definition (Visibility of an Edge): Given a PAG $\mathscr{P}$, a directed edge $X \to Y$ is visible if:

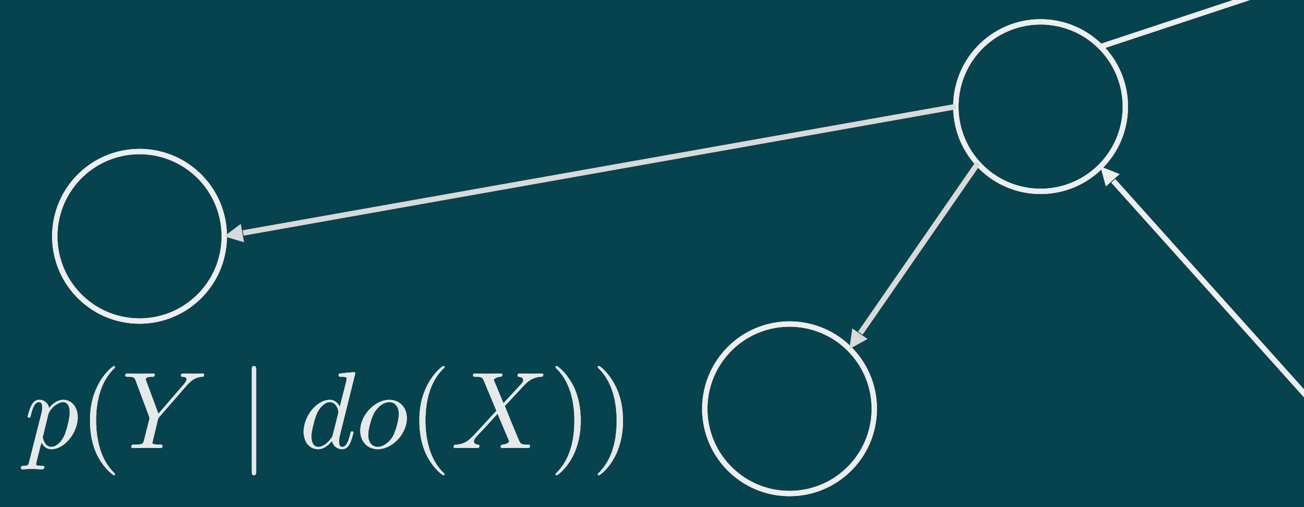
1. there is a node V not adjacent to Y such that there is an edge between V and X that is into X, or
2. if there is a collider path from V to X that is into X and every non-endpoint node on the path is a parent of Y. Otherwise, $X \to Y$ is said to be invisible.



Edges labeled with ‘v’ are referred to as visible edges

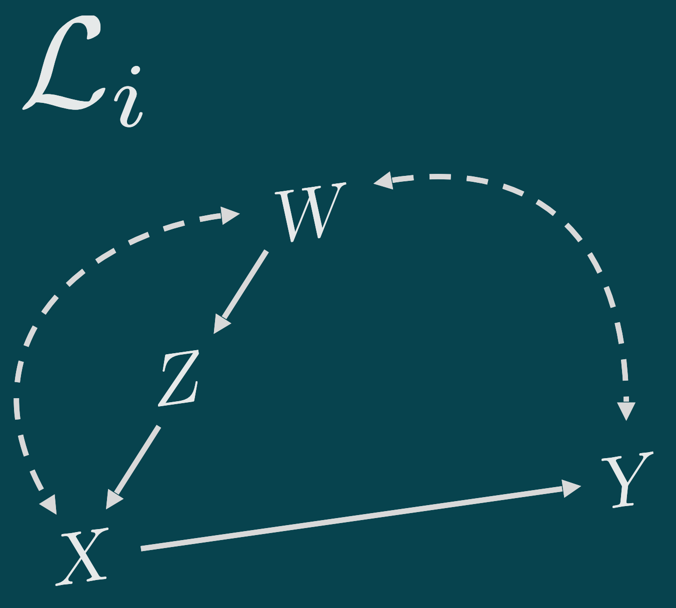
Visibility of an edge denotes the absence of a hidden confounder in every member of the equivalence class.





4 Causal Discovery

Fast Causal Inference (FCI) Algorithm



Fast Causal Inference (FCI) Algorithm

Fast Causal Inference (FCI) Algorithm: Learn a graphical representation of the Markov Equivalence Class of causal diagrams (ADMGs) from observational data.

Assumptions: the observed distribution is the marginal of a distribution P that satisfies the following conditions for the true causal diagram G (an **ADMG**):

1) **I-Map / Semi-Markov Condition:** for any disjoint subsets $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$:
$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z})_G \Rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z})_P.$$

2) **Faithfulness Condition:** for any disjoint subsets $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$:
$$(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z})_P \Rightarrow (\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z})_G.$$

G is an ***I-Map of*** P

P is ***semi-Markov relative*** to G.

P is ***faithful*** to G

Note: Estimation of the marginal distribution from limited data requires and **additional assumption**:

3) An adequate *conditional independence test* is available.

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

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

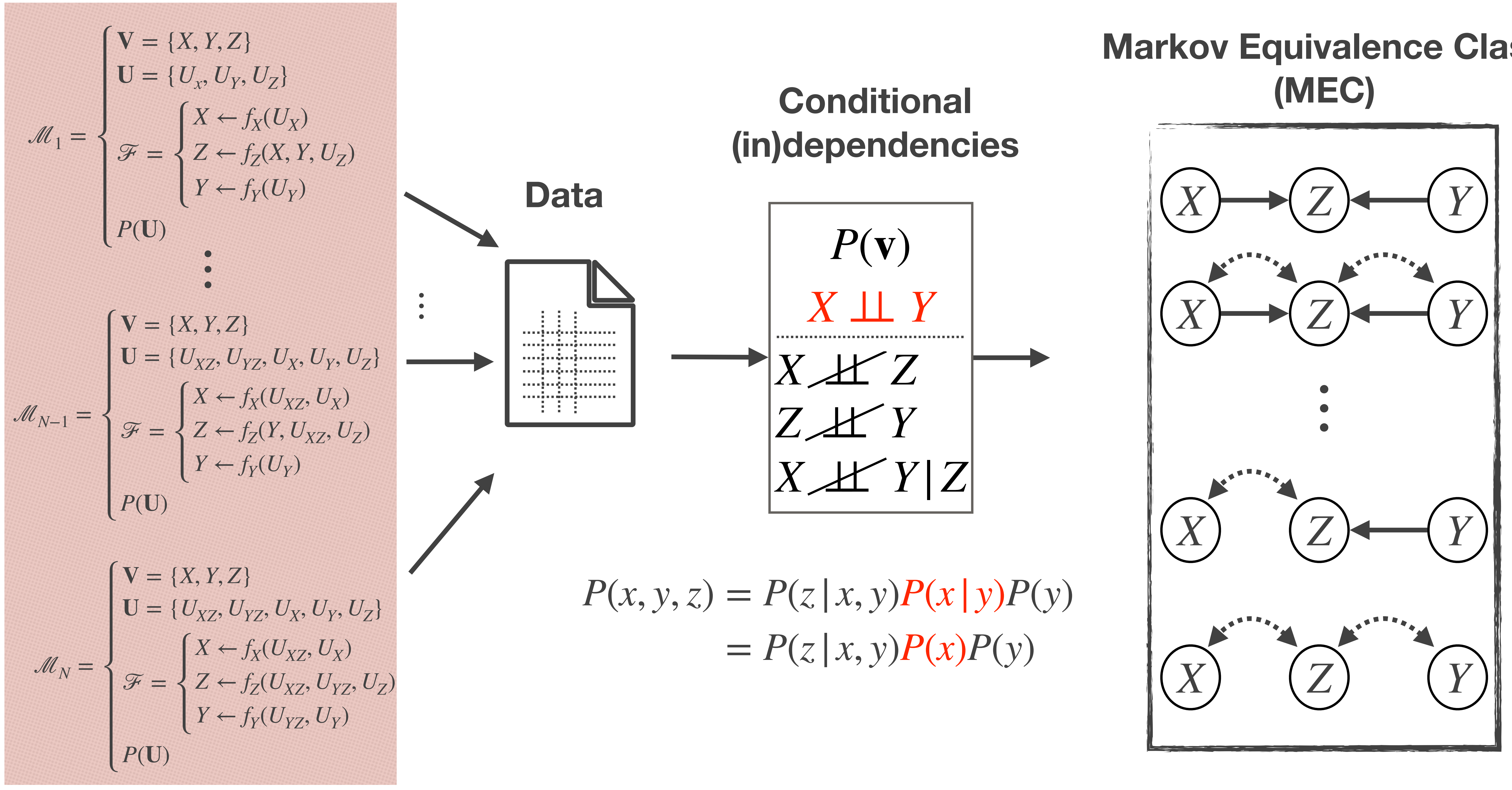
Continuous (conditional Gaussian) or Discrete (Binary, Ordinal, Multinomial) - Linear Regression

- 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) :

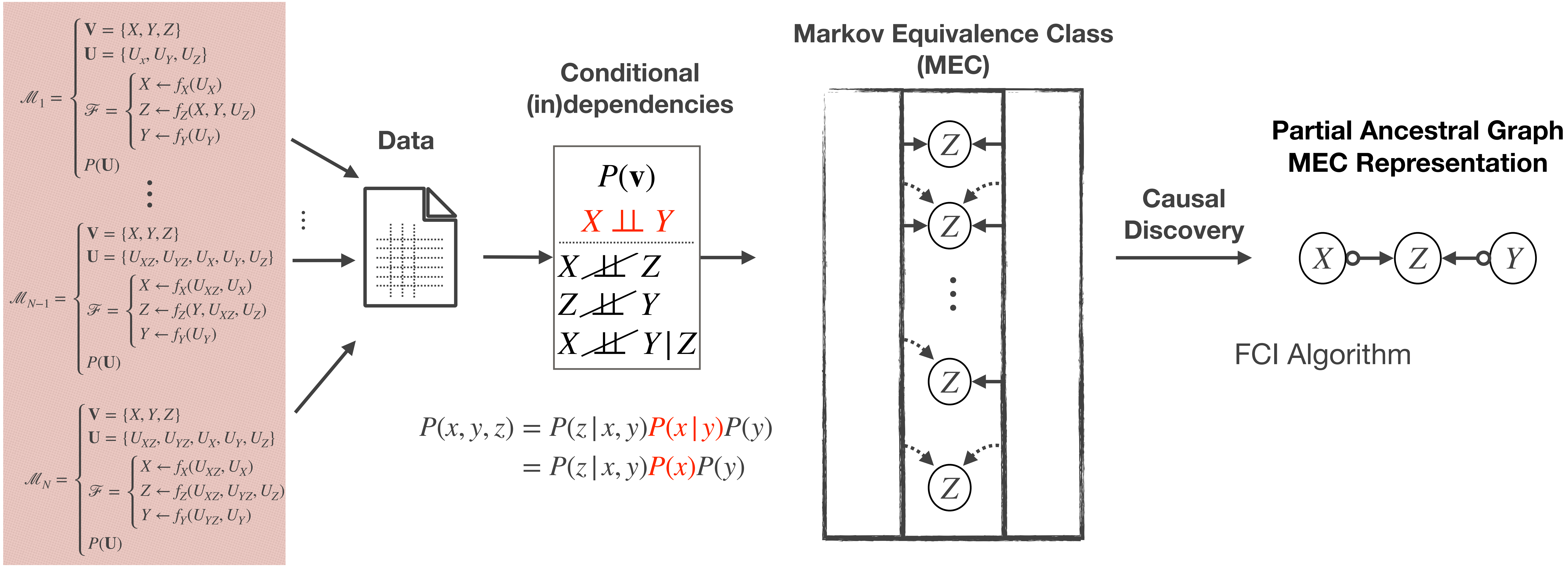
Ribeiro A.H.**,** Soler J.M.P. (2020). *Learning Genetic and environmental graphical models from family data*, Statistics in Medicine.
R package: https://github.com/adele/FamilyBasedPGMs

Learning Structural Invariances



$$P(x, y, z) = P(z \mid x, y)P(x \mid y)P(y)$$
$$= P(z \mid x, y)P(x)P(y)$$

Learning Structural Invariances

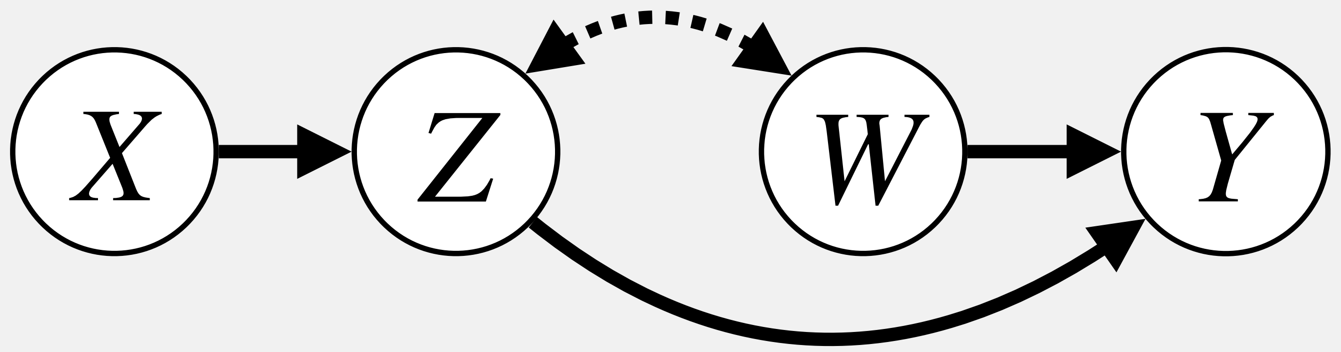


$$
\begin{aligned}
P(x, y, z) &= P(z \mid x, y)P(x \mid y)P(y) \\
&= P(z \mid x, y)P(x)P(y)
\end{aligned}
$$

FCI Algorithm - Pipeline

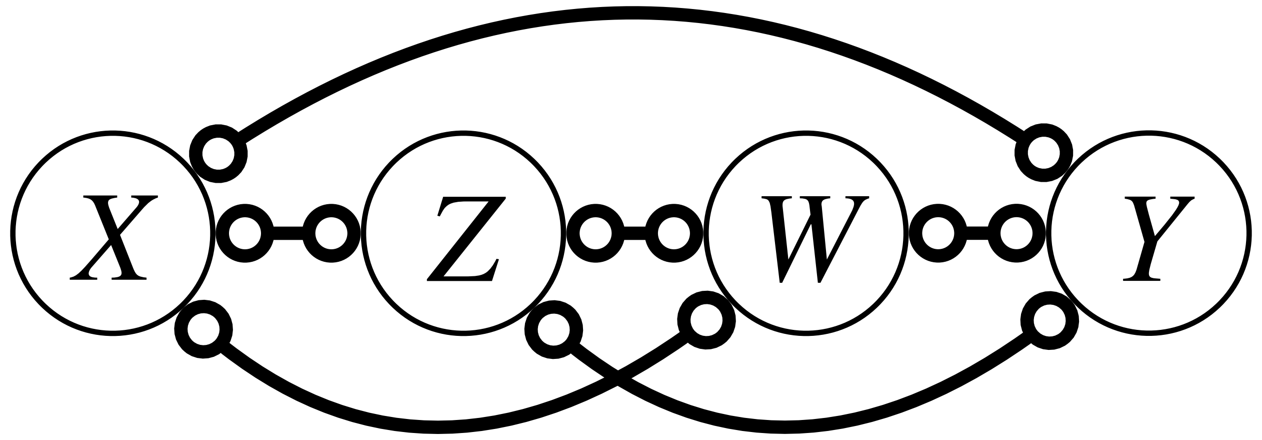
Unknown Reality

True causal diagram



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

Implied by the ADMG using d-separation

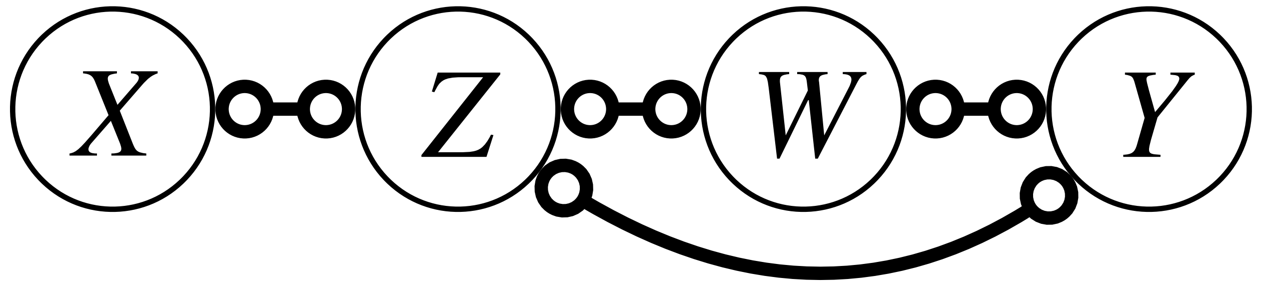


Complete Graph

Conditional Independence Tests

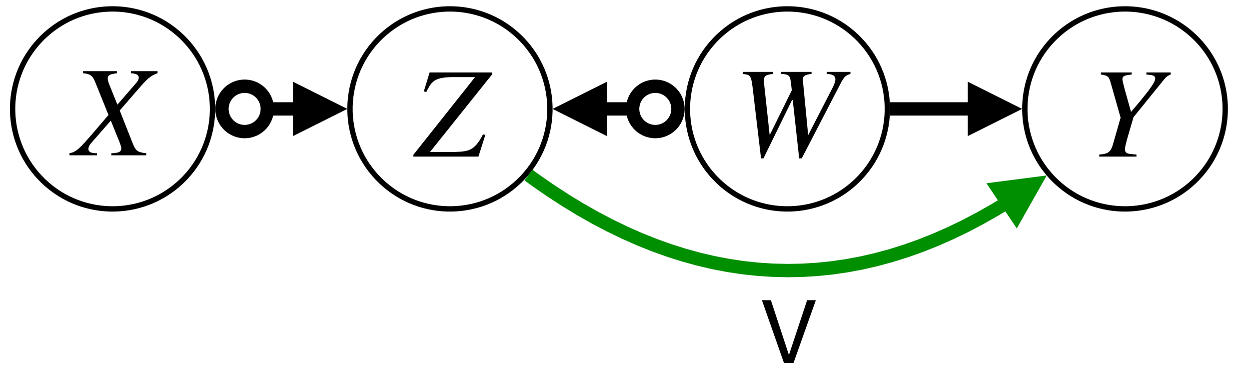
$X \perp\!\!\!\perp W$
$X \perp\!\!\!\perp Y \mid Z, W$

By faithfulness, are observed in the data



Skeleton

FCI Rules
$(R1) - (R10)$



Partial Ancestral Graph (PAG)

Implied by the PAG using m-separation

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

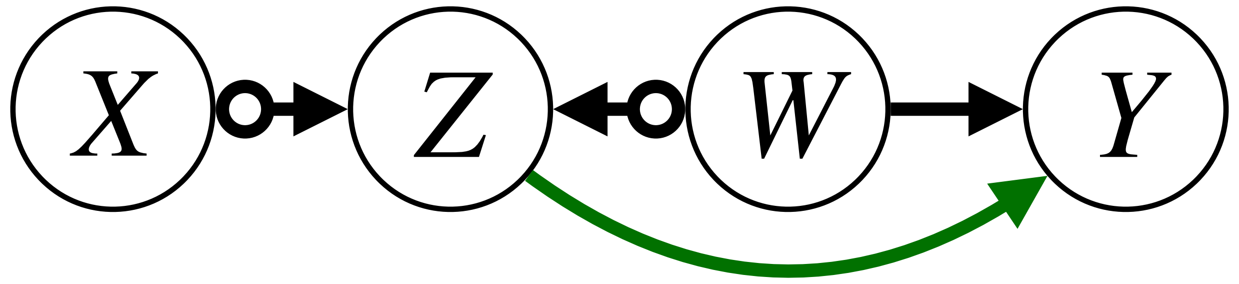
$A \longrightarrow B \implies$ ancestrally
$A \circ\!\!\longrightarrow B \implies$ non-ancestrality
$A \longleftrightarrow B \implies$ spurious association
$A \vdash\!\!\!\dashv B \implies$ selection bias

Z is not an ancestor of X or W.

Z and W are ancestors of Y.

Z is not confounded with Y.

PAG: Representation of the Markov Equivalence Class



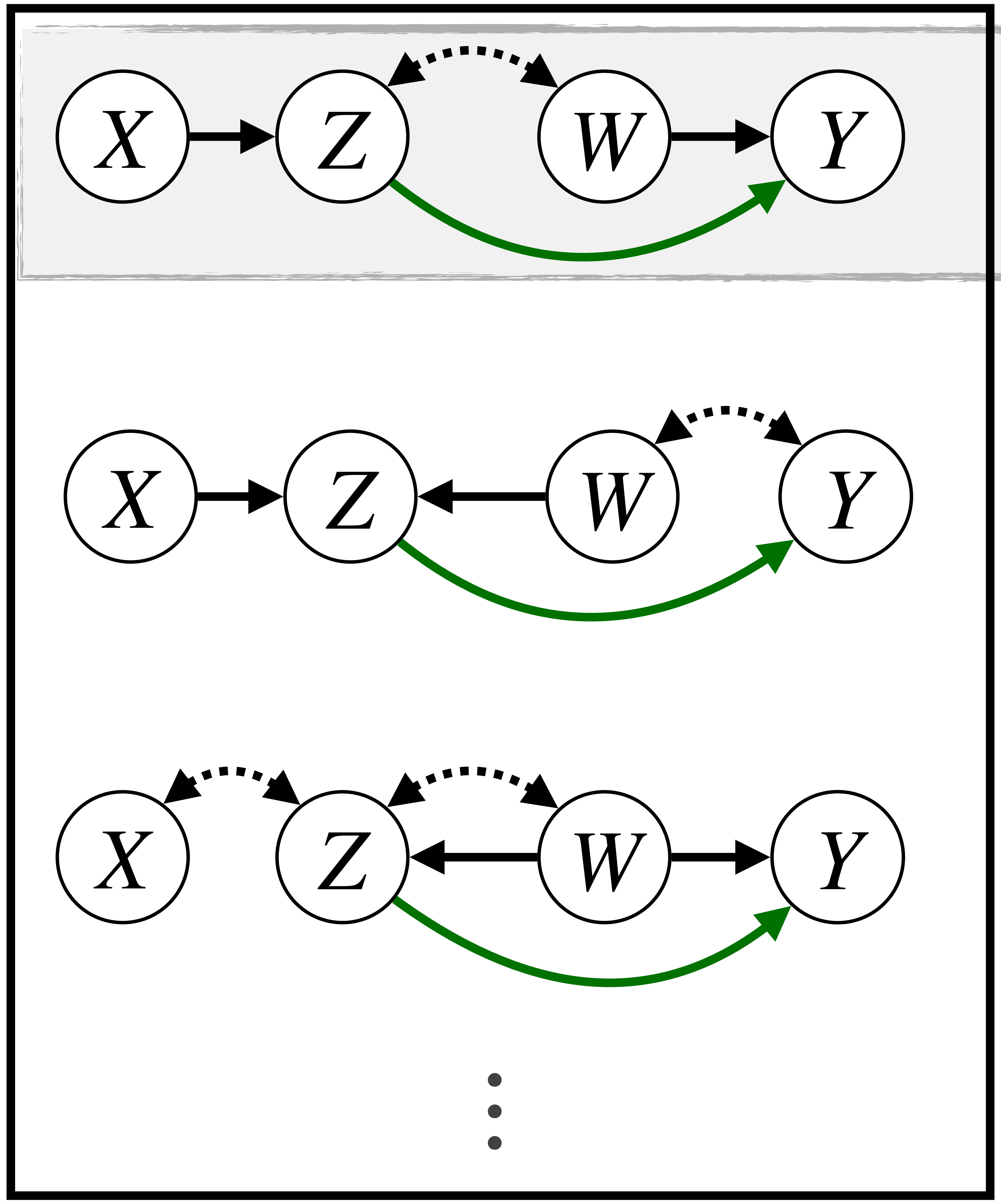
V

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**.**

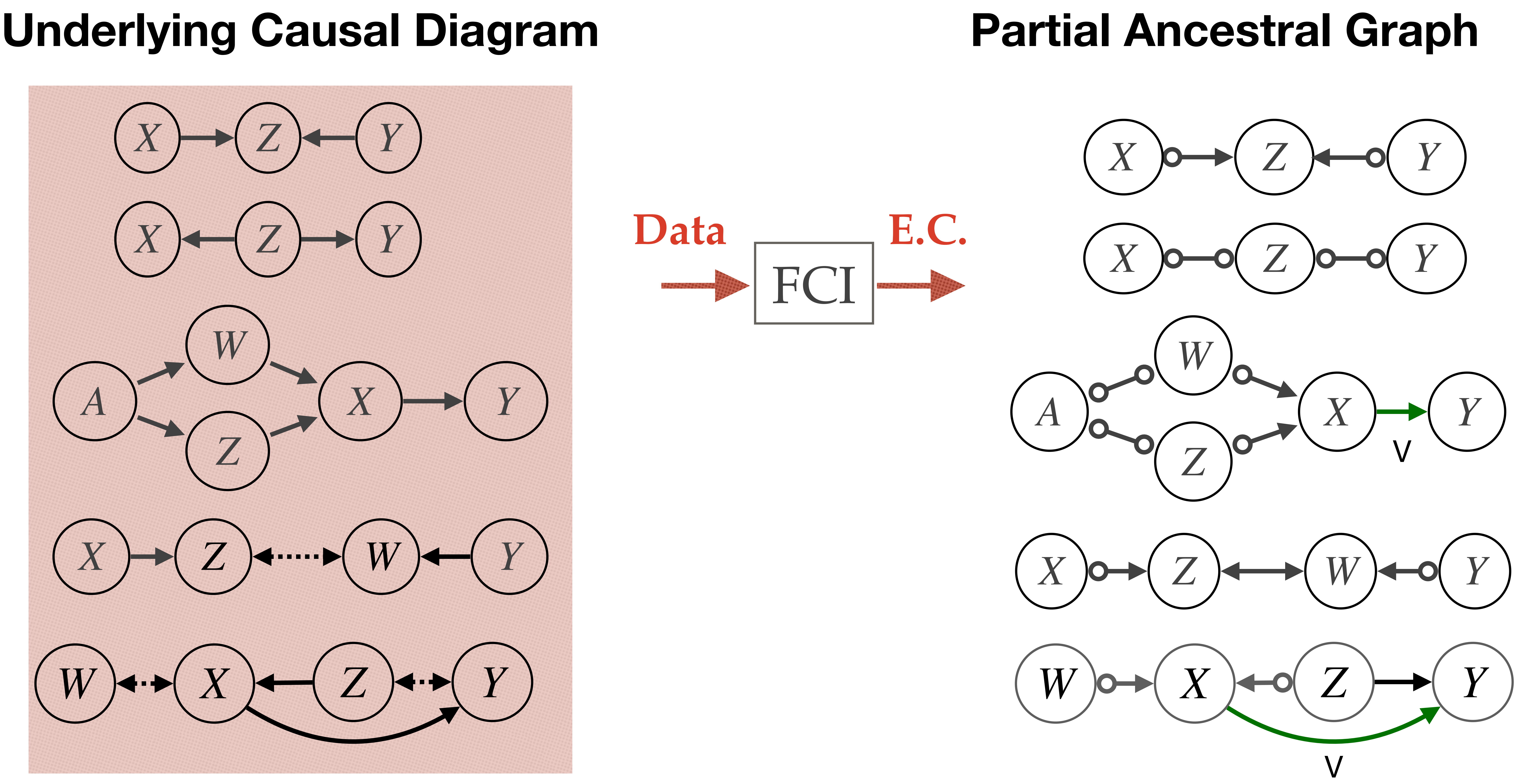


True (unknown) causal diagram

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

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

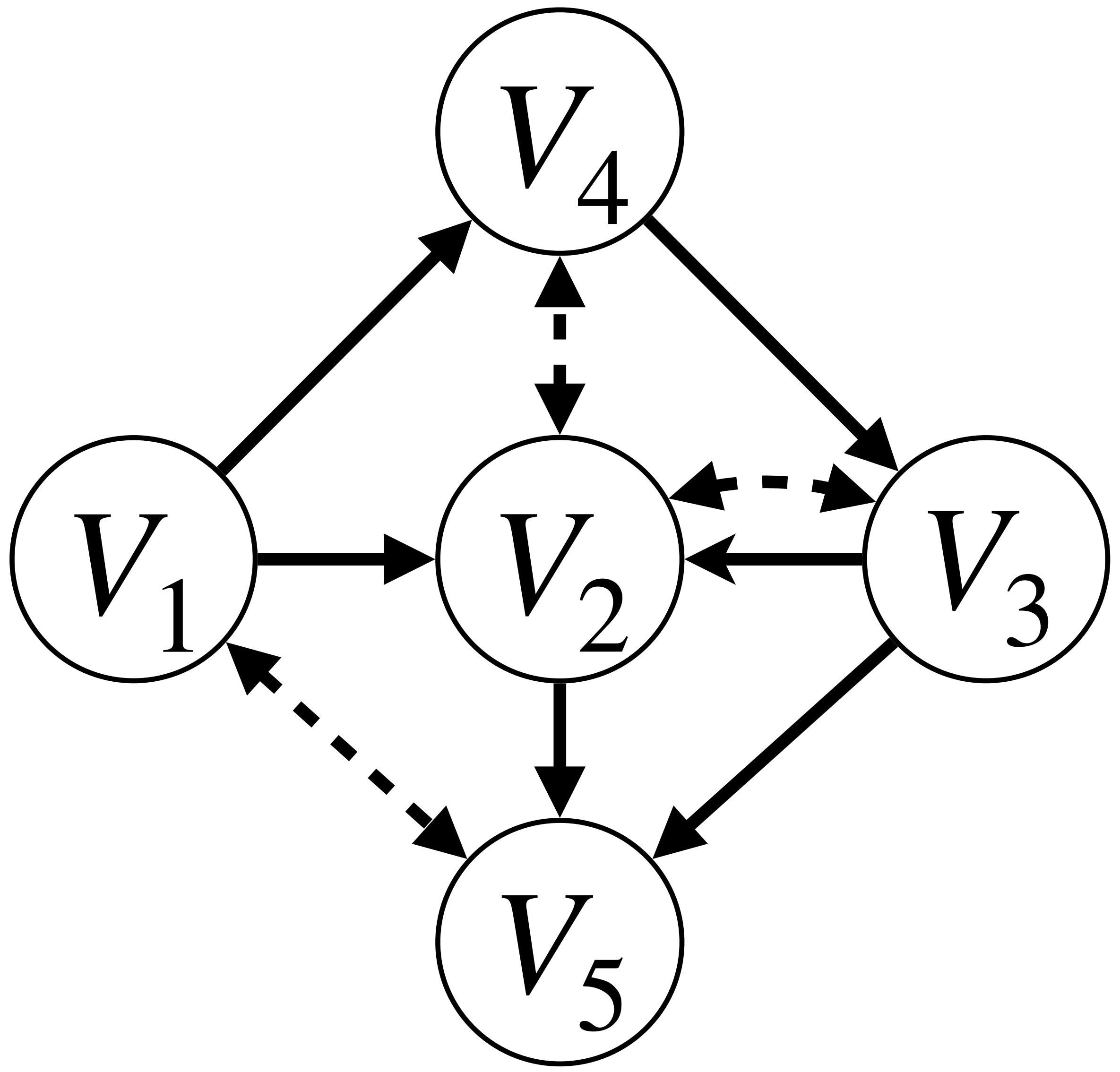
Other Examples



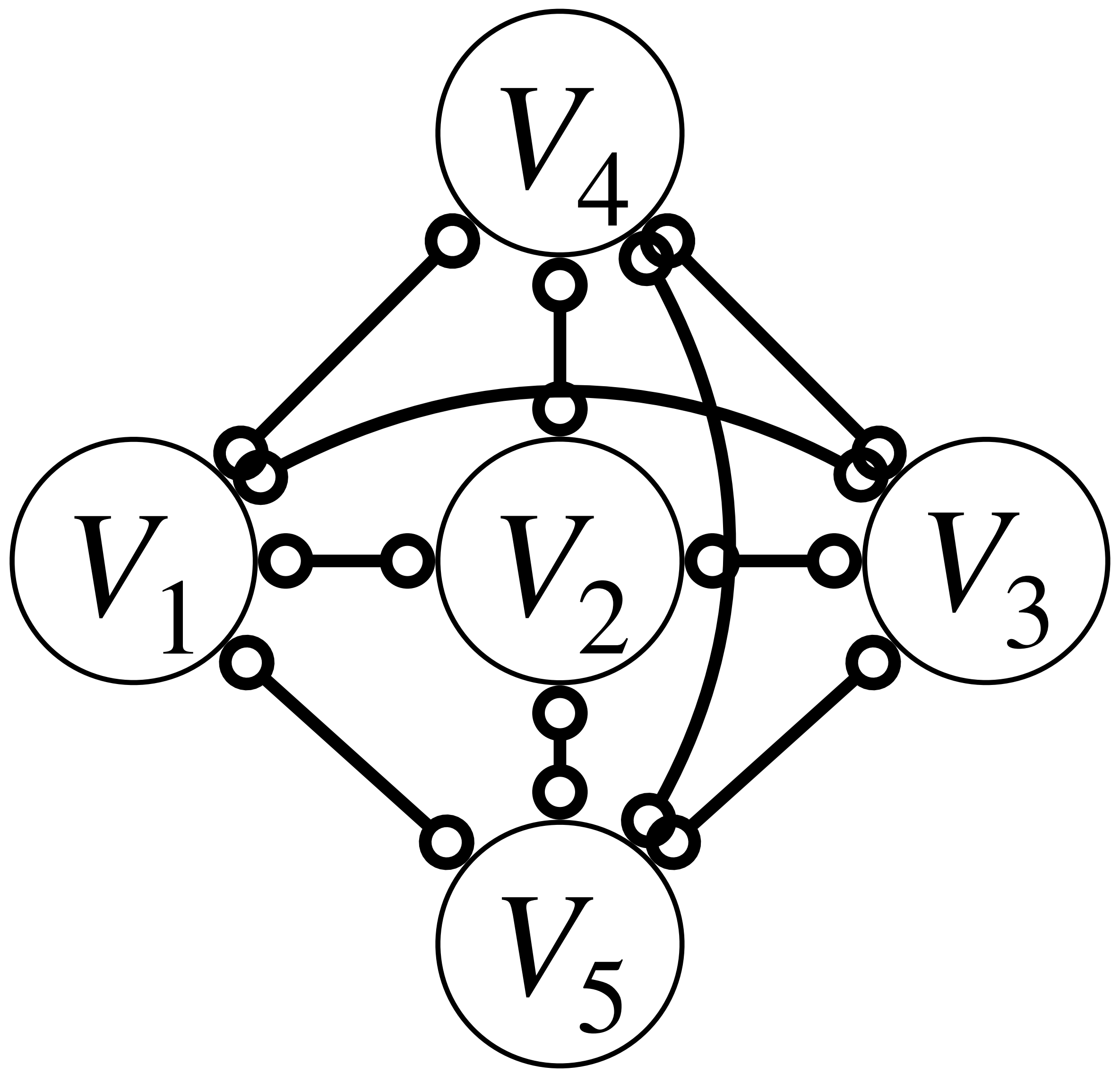
Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896. [Link]

FCI - Skeleton

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

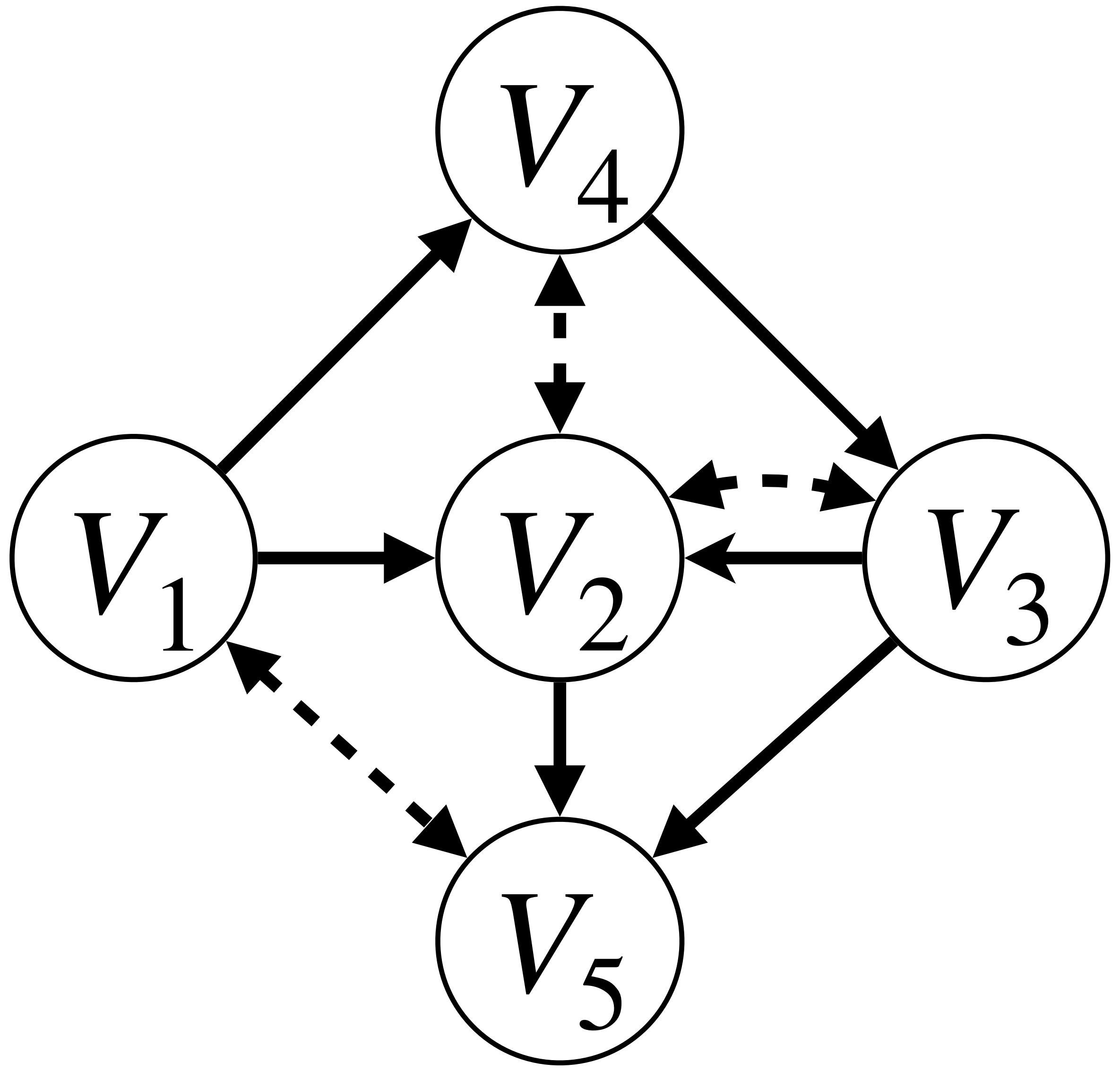


True, unknown ADMG

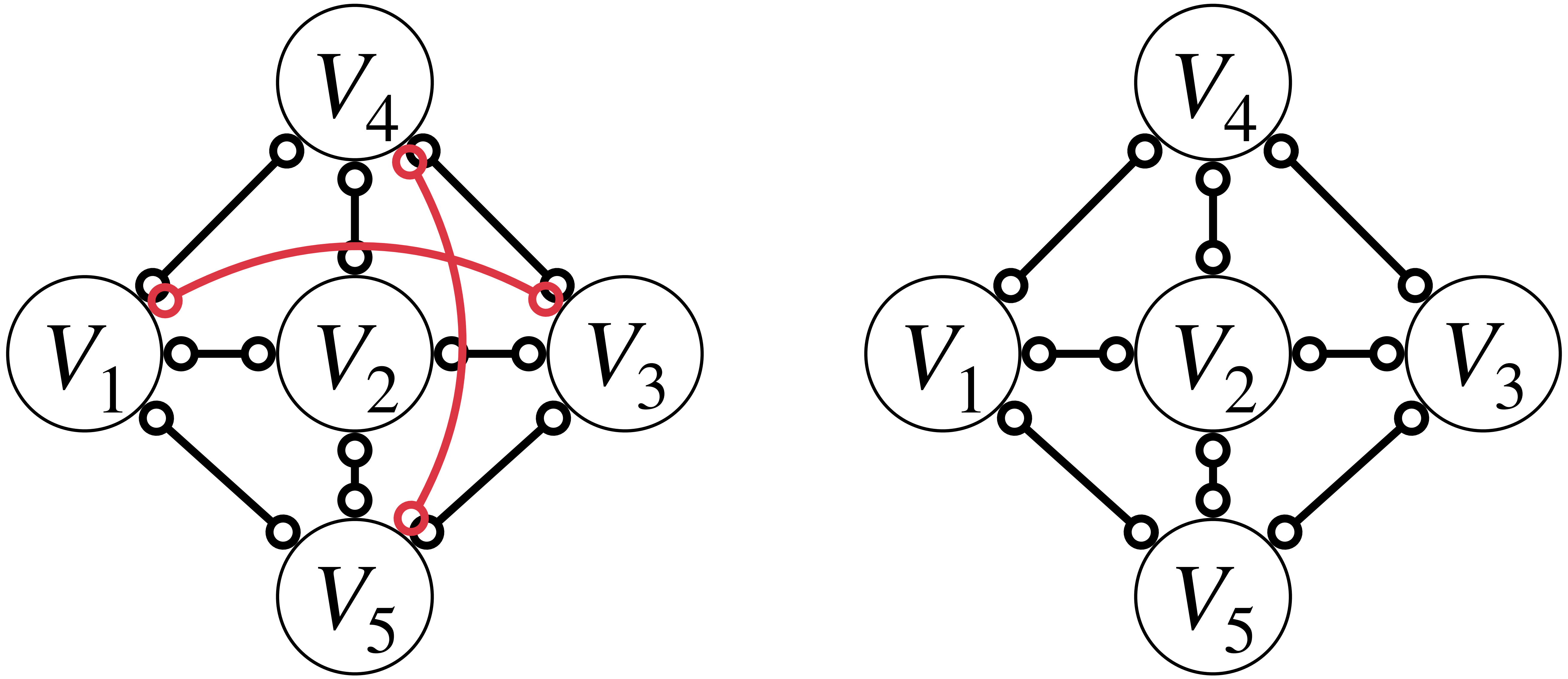


FCI - Skeleton

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).



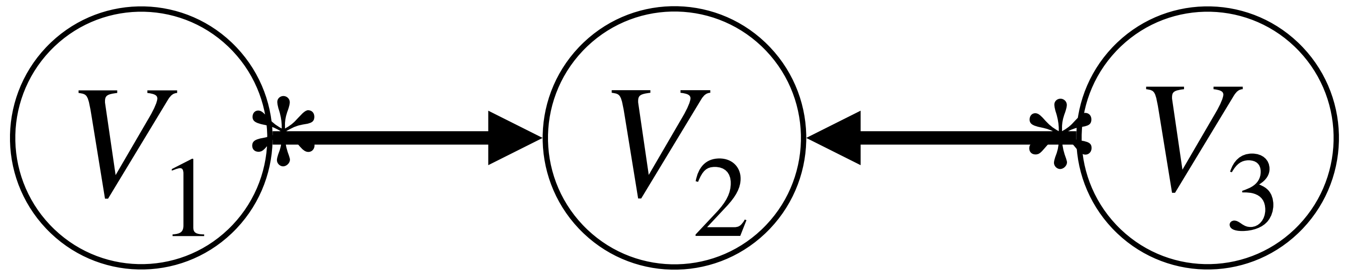
True, unknown ADMG



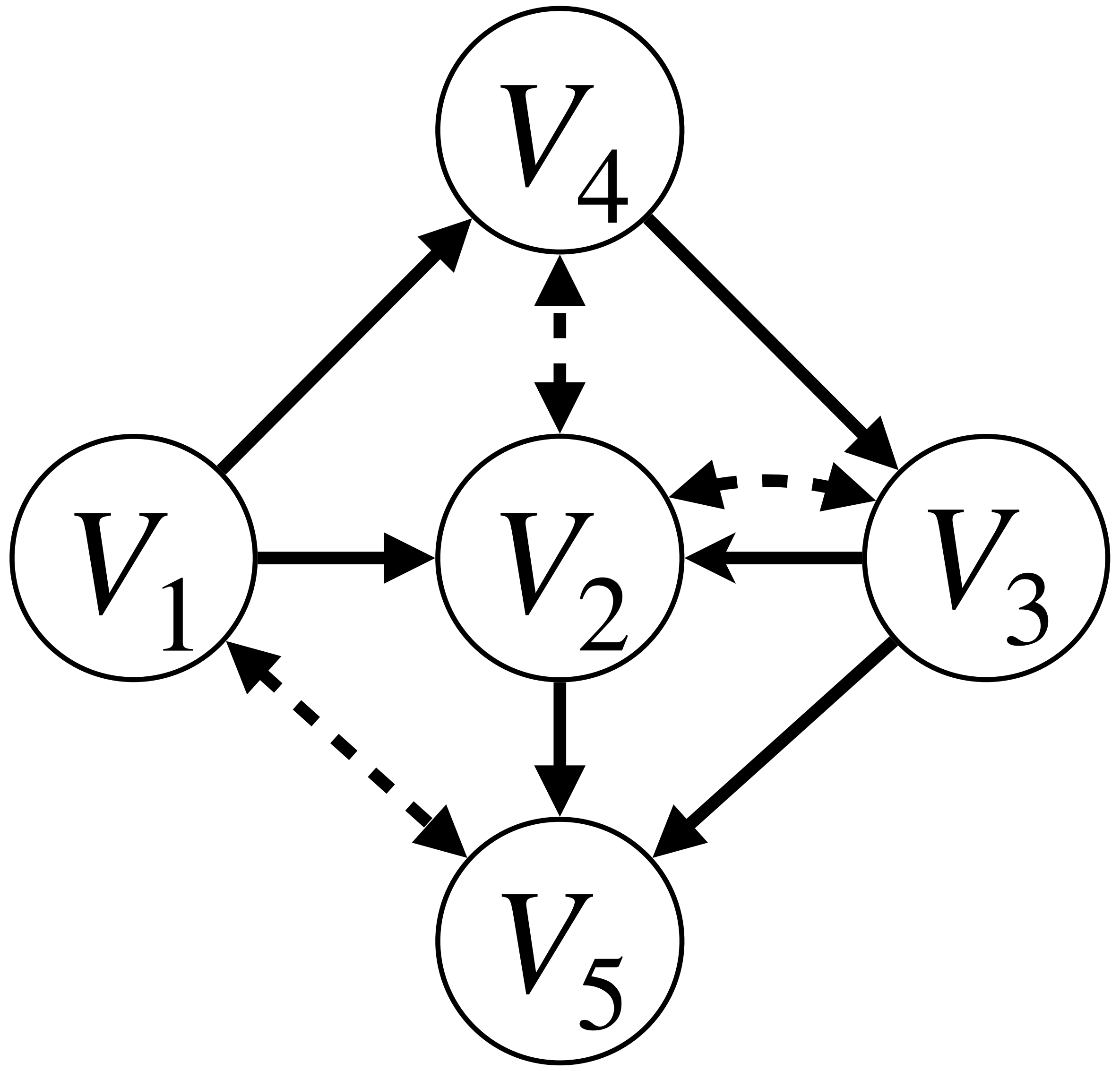
$V_1 \perp\!\!\!\perp V_3 \mid V_4$ and $V_4 \perp\!\!\!\perp V_5 \mid V_1, V_2, V_3$

FCI - Orienting the Colliders

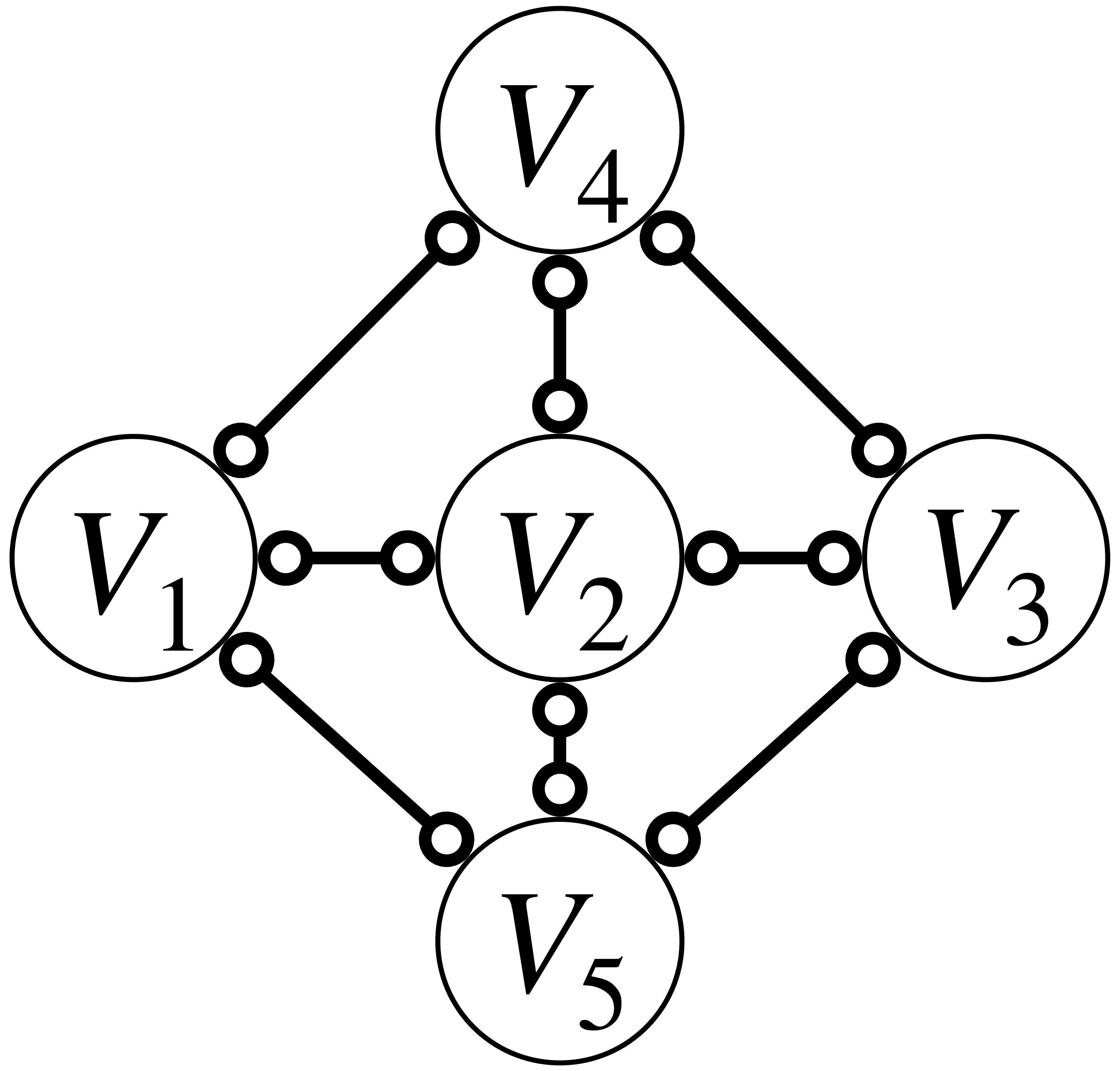
R0: If $\langle V_1, V_2, V_3 \rangle$ is unshielded and $V_2 \notin \mathit{Sepset}(V_1, V_3)$, then

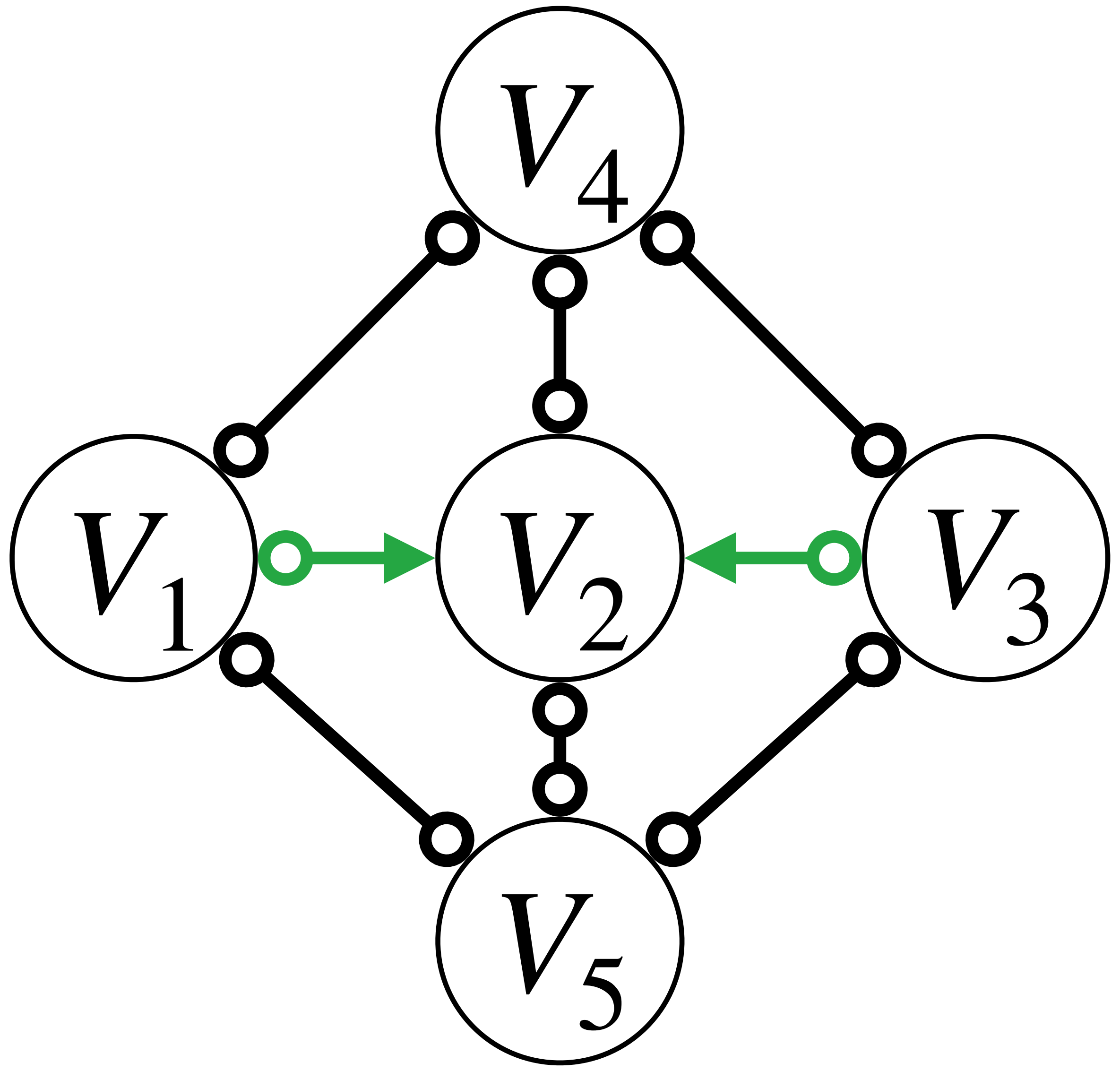


That is the only way for the path between V_1 and V_3 to be blocked when not conditioning on V2



True, unknown ADMG

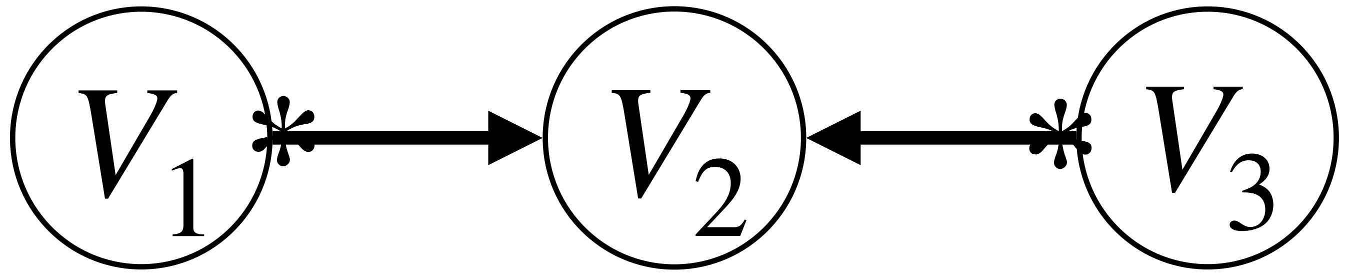




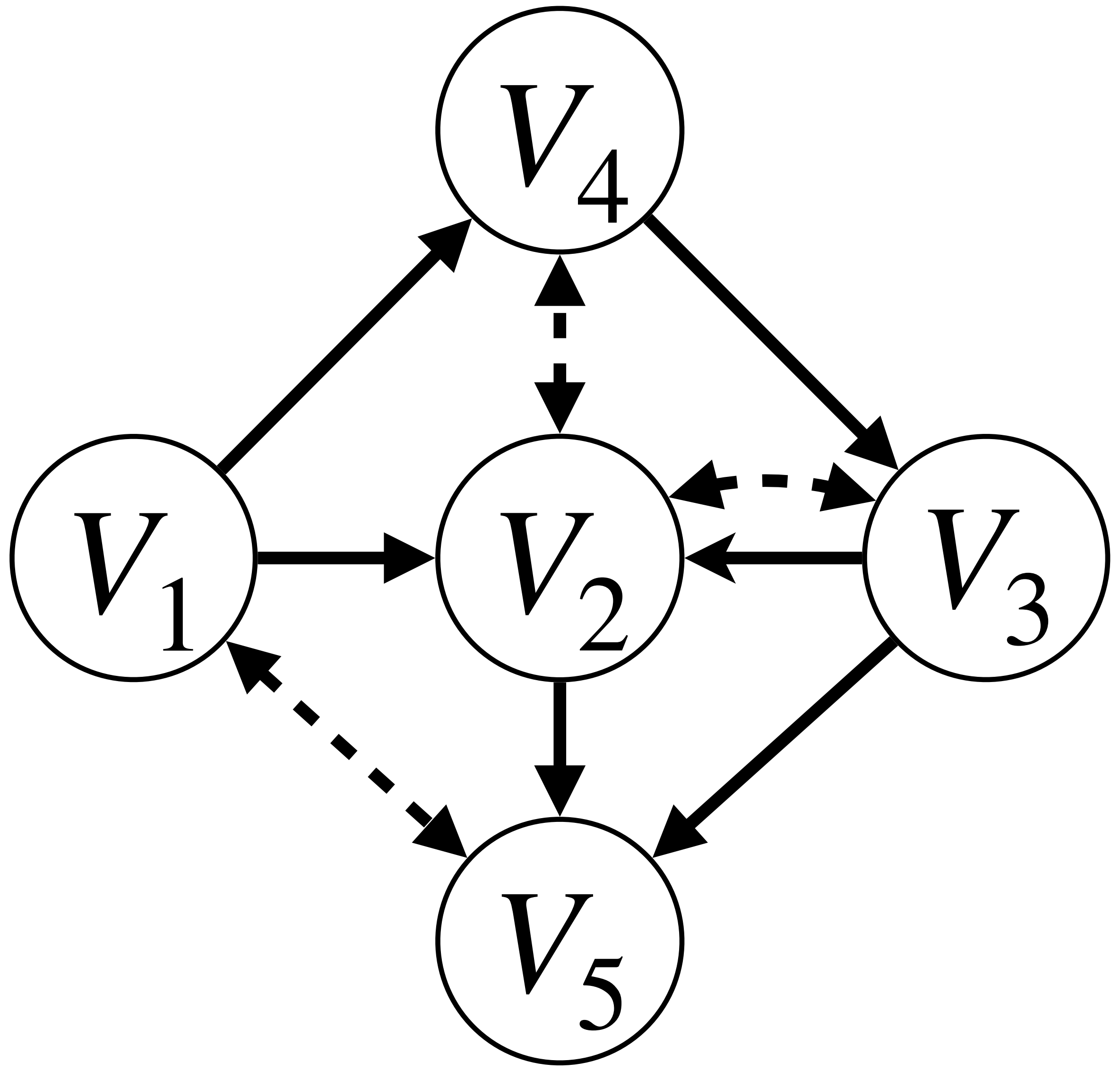
$V_1 \perp\!\!\!\perp V_3 \mid V_4$ and $V_1 \not\perp\!\!\!\perp V_3 \mid V_4, V_2$

FCI - Orienting the Colliders

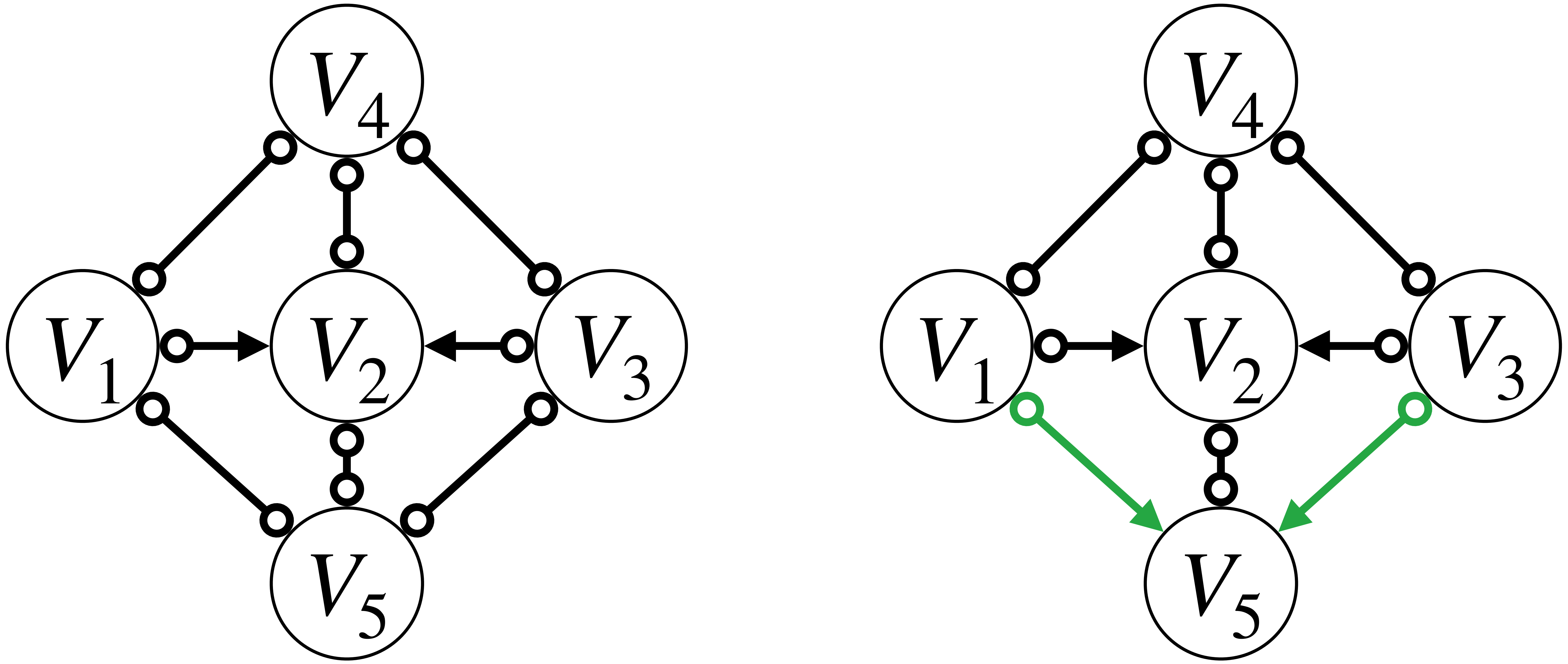
R0: If $\langle V_1, V_2, V_3 \rangle$ is unshielded and $V_2 \notin Sepset(V_1, V_3)$, then



We apply R0 until no more collider can be oriented!



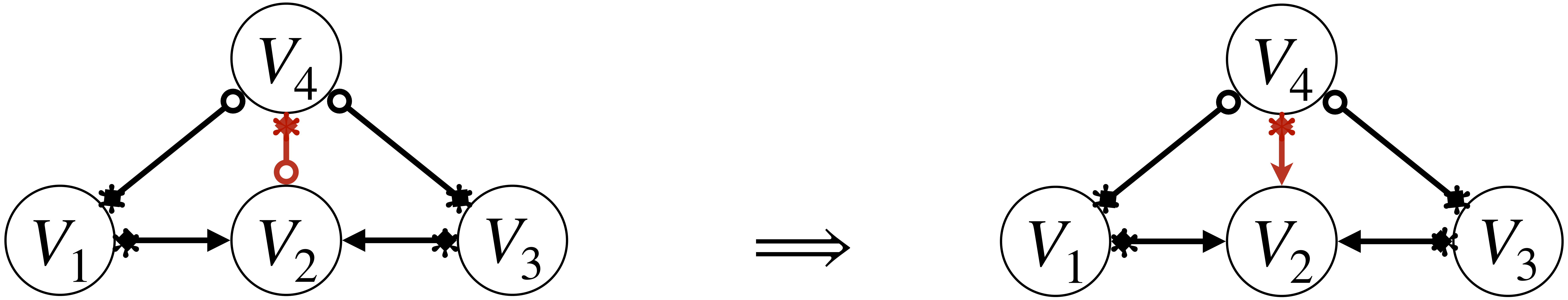
True, unknown ADMG



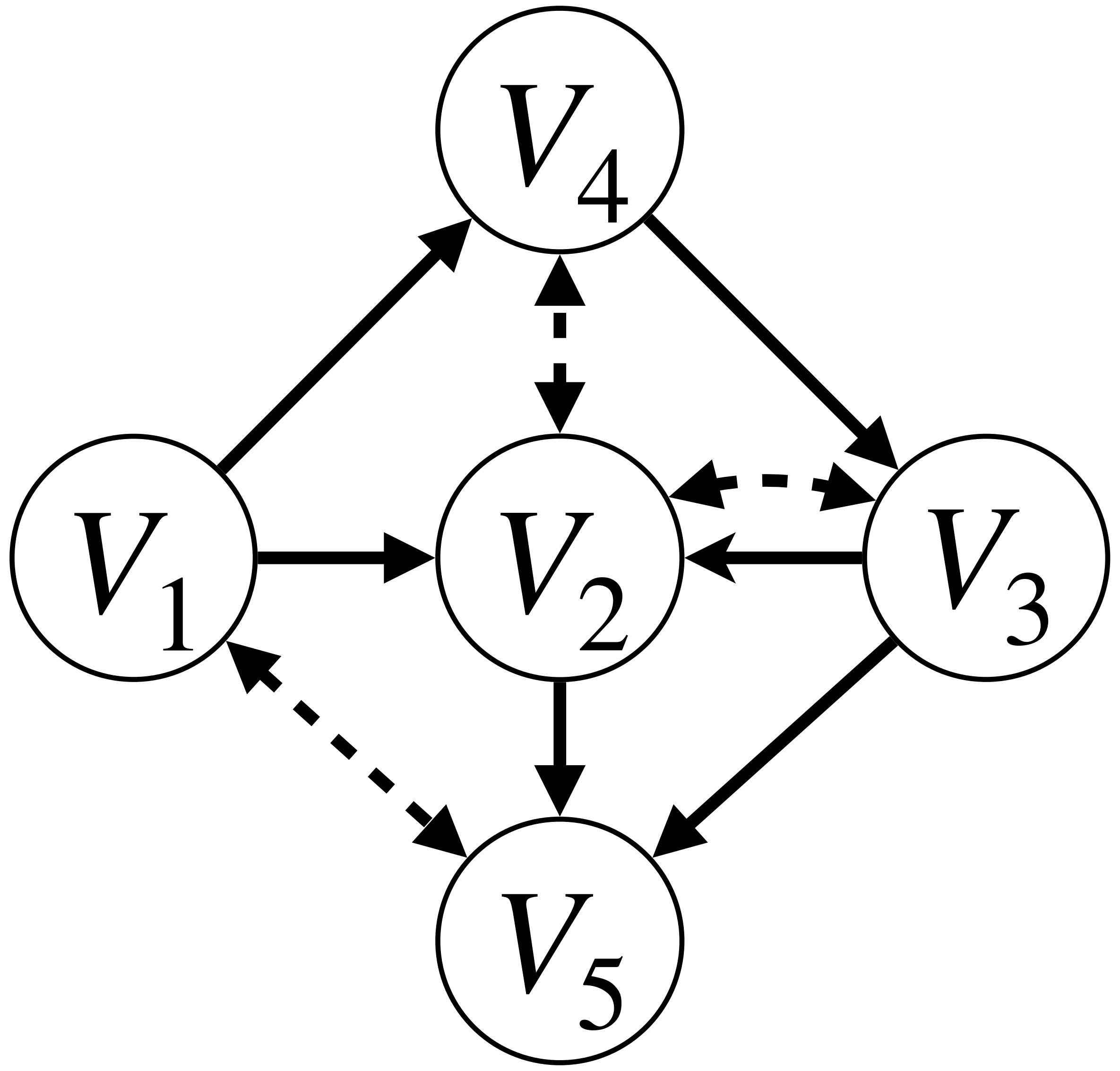
$V_1 \perp\!\!\!\perp V_3 \mid V_4$ and $V_1 \not\perp\!\!\!\perp V_3 \mid V_4, V_5$

Applying Mark Inference Rules

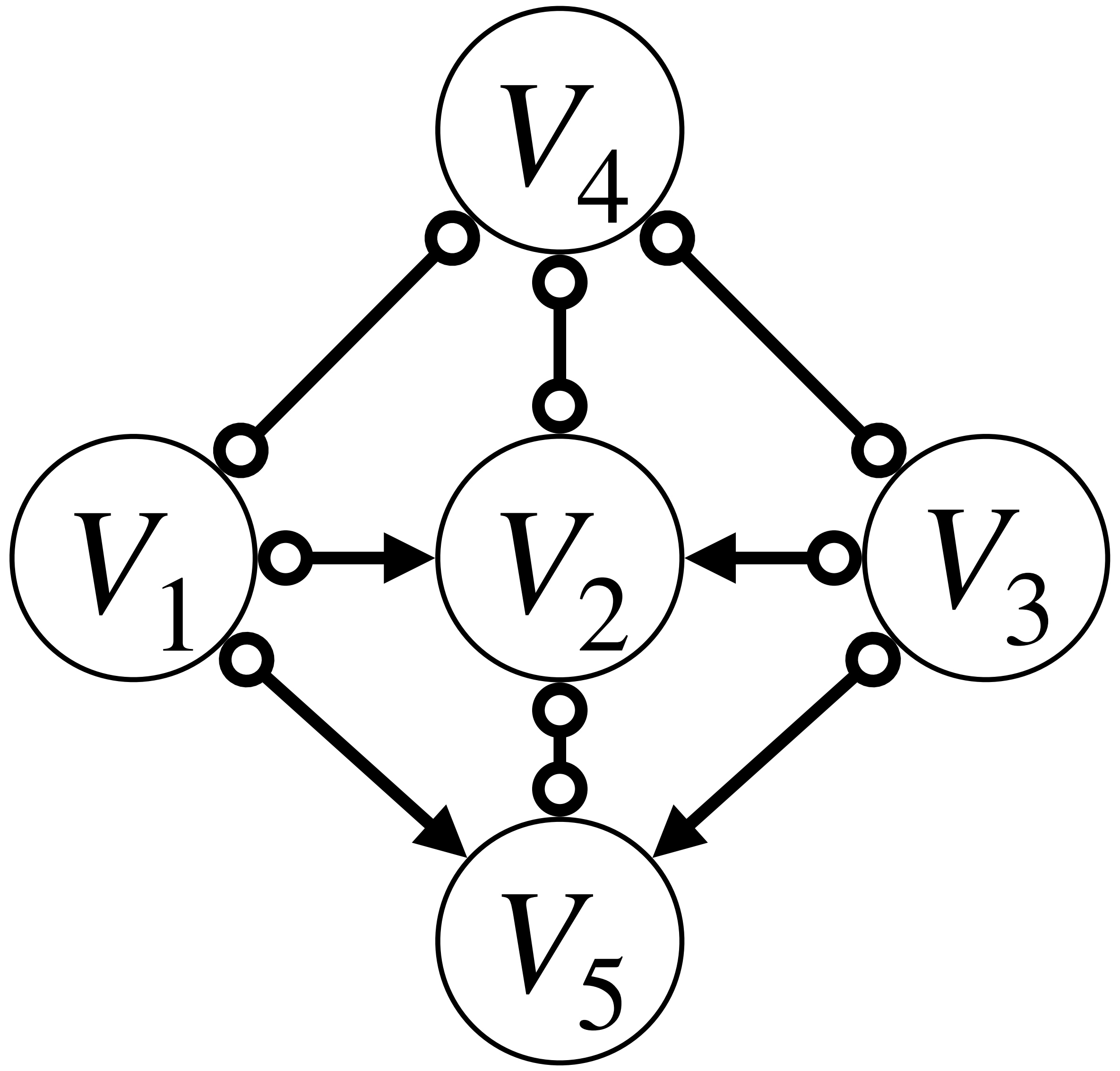
R3:



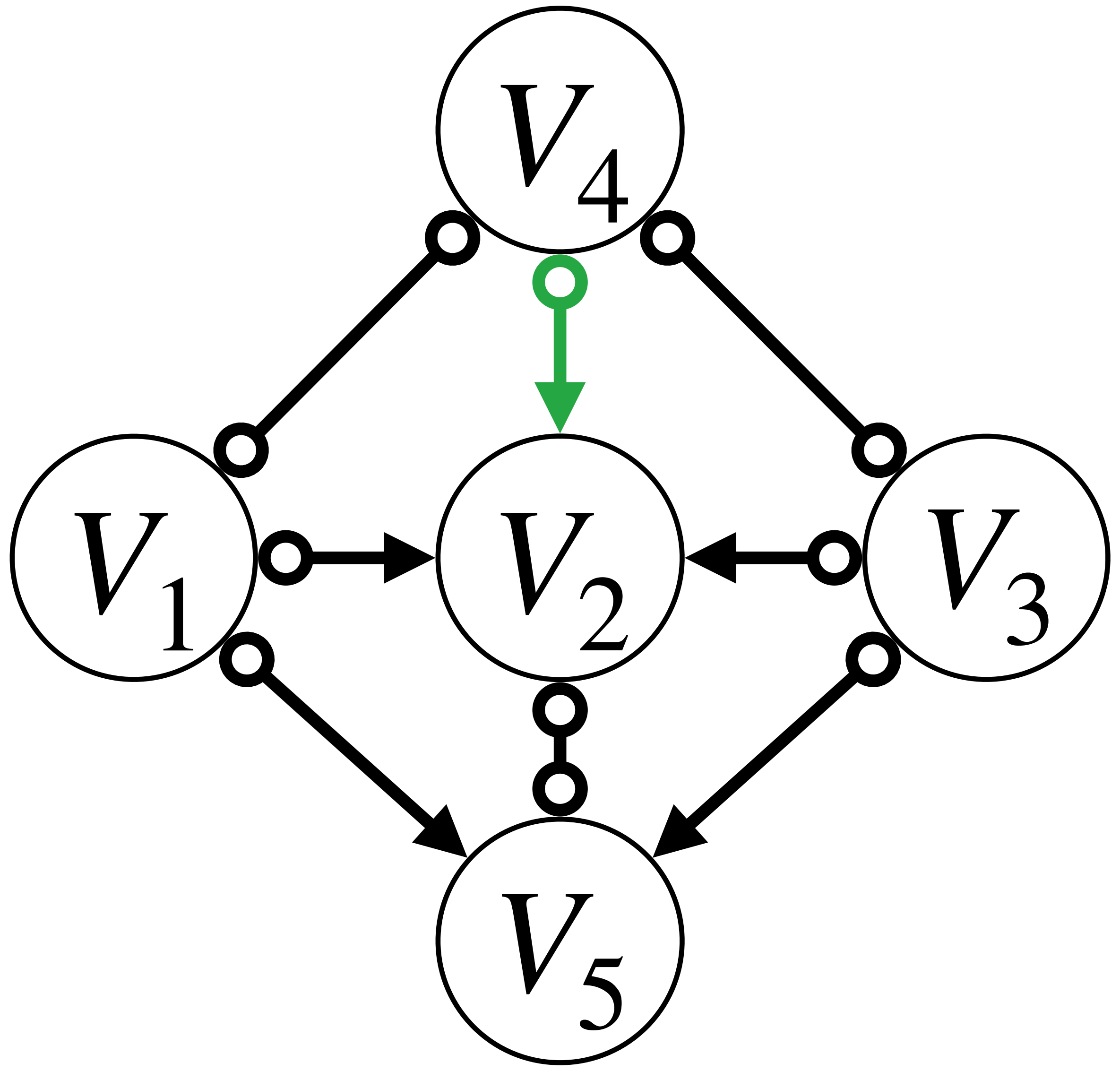
where V_1 and V_3 are not adjacent



True, unknown ADMG



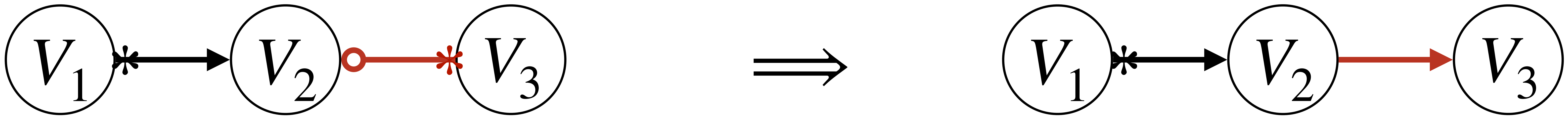
After Skeleton + R0



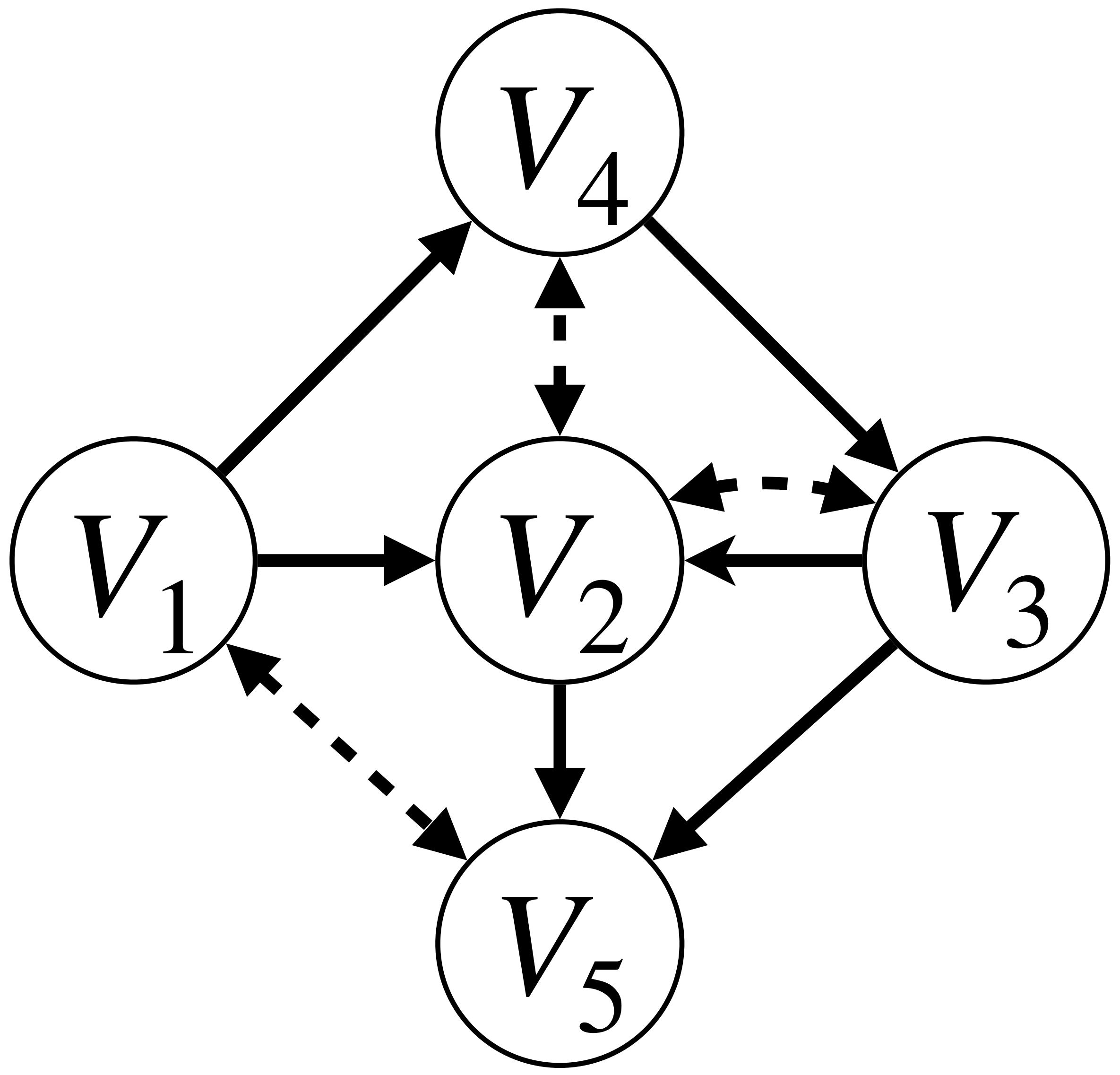
Applying R3

Applying Mark Inference Rules

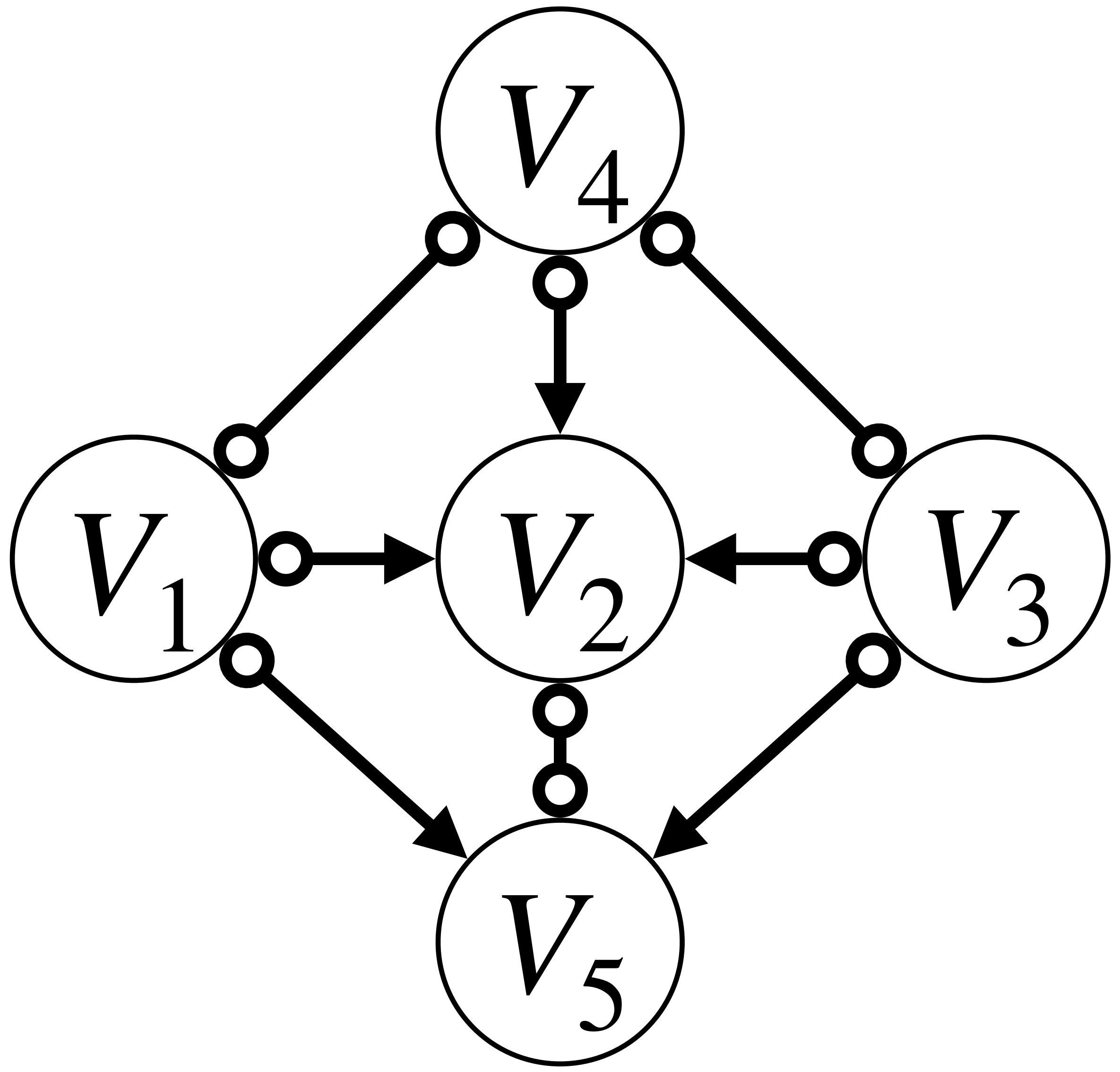
R1:



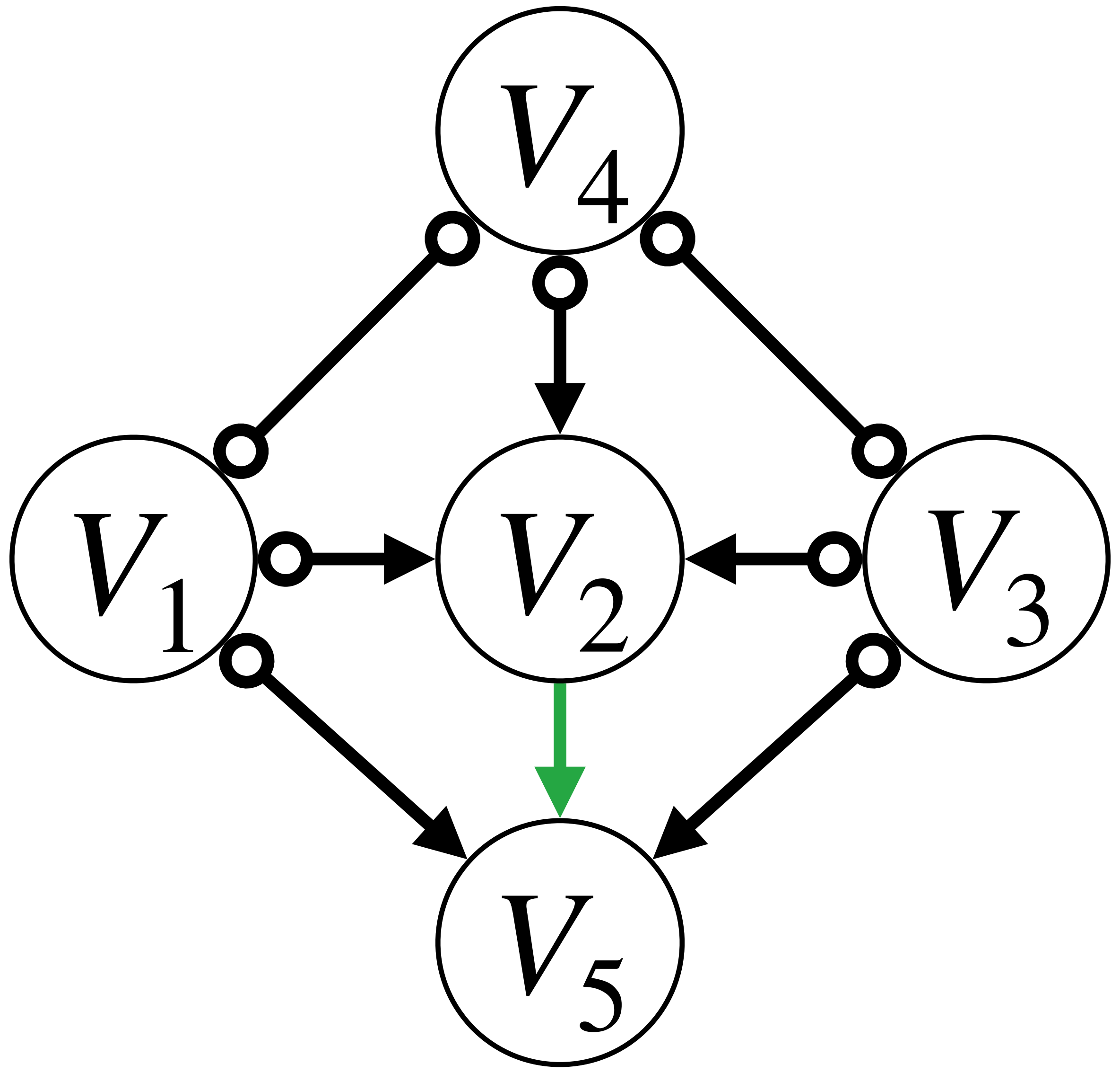
where V_1 and V_3 are not adjacent



True, unknown ADMG



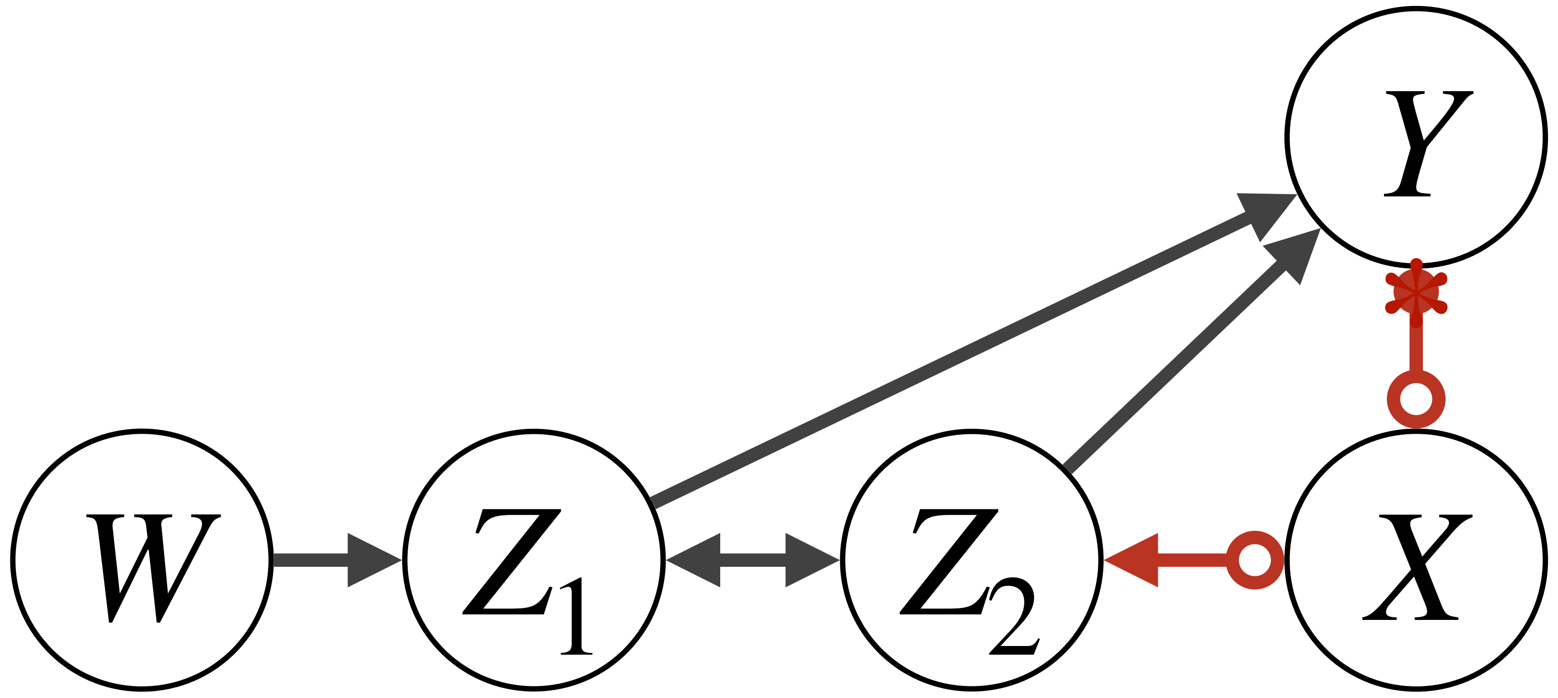
After Skel + R0 + R3



Applying R1

Applying Mark Inference Rules

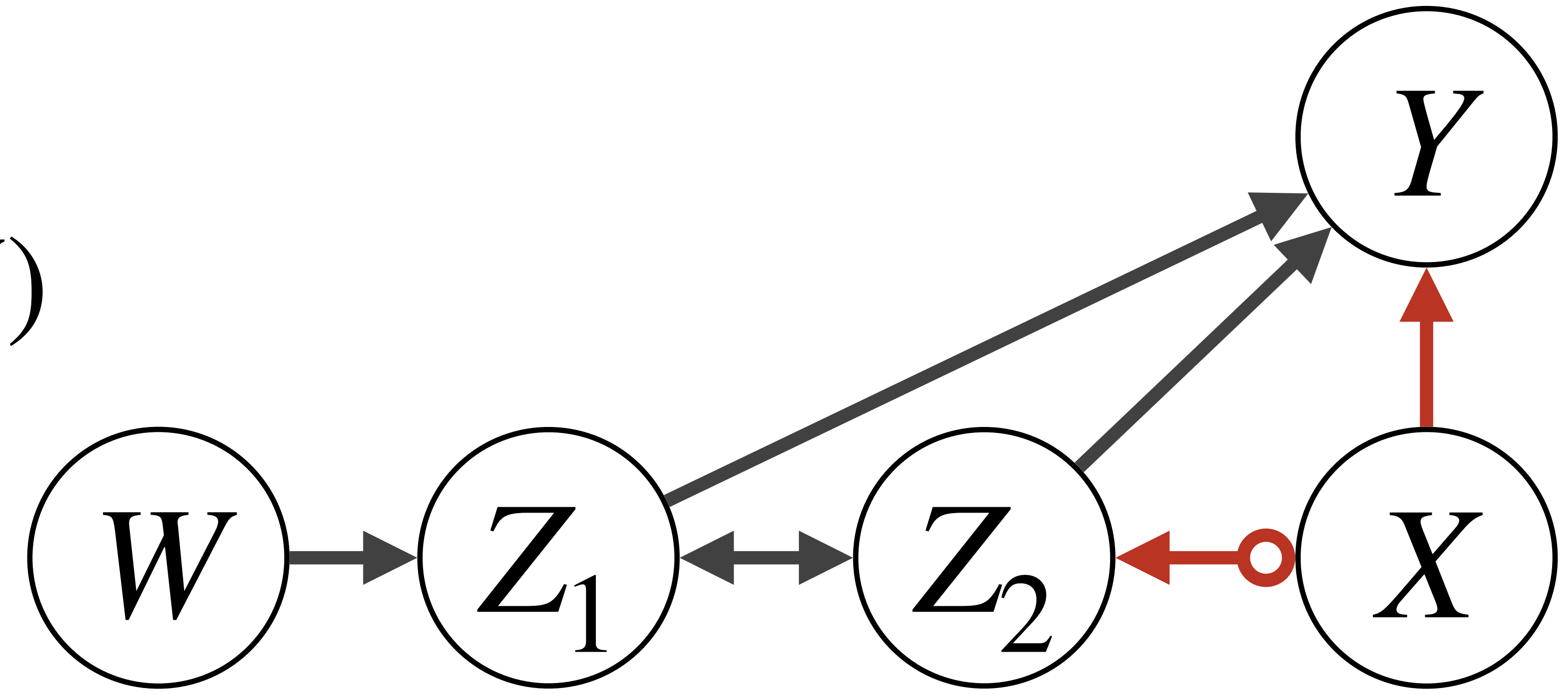
R4:

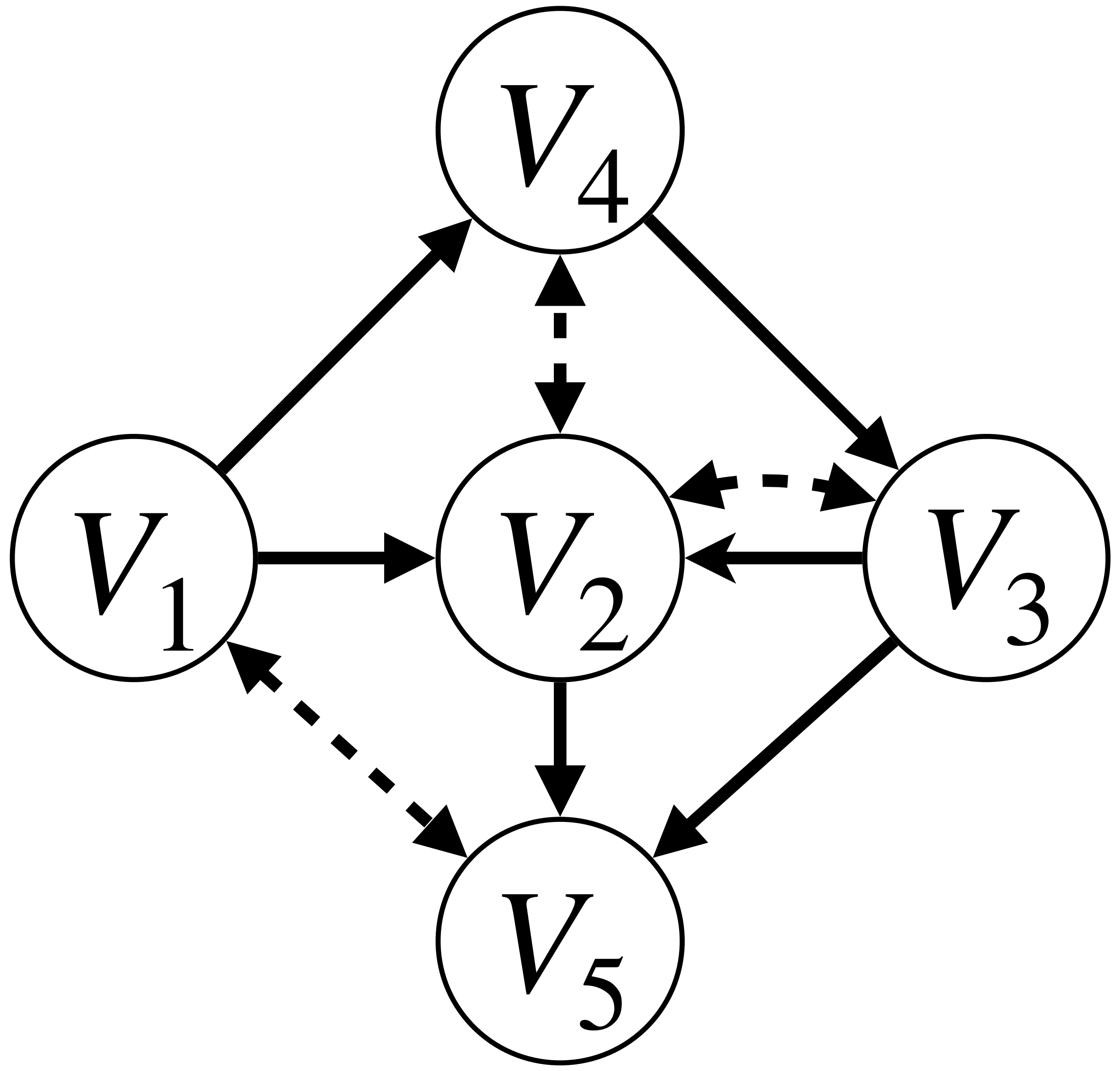


$\langle W, Z_1, Z_2, X, Y \rangle$ is a discriminating path for X

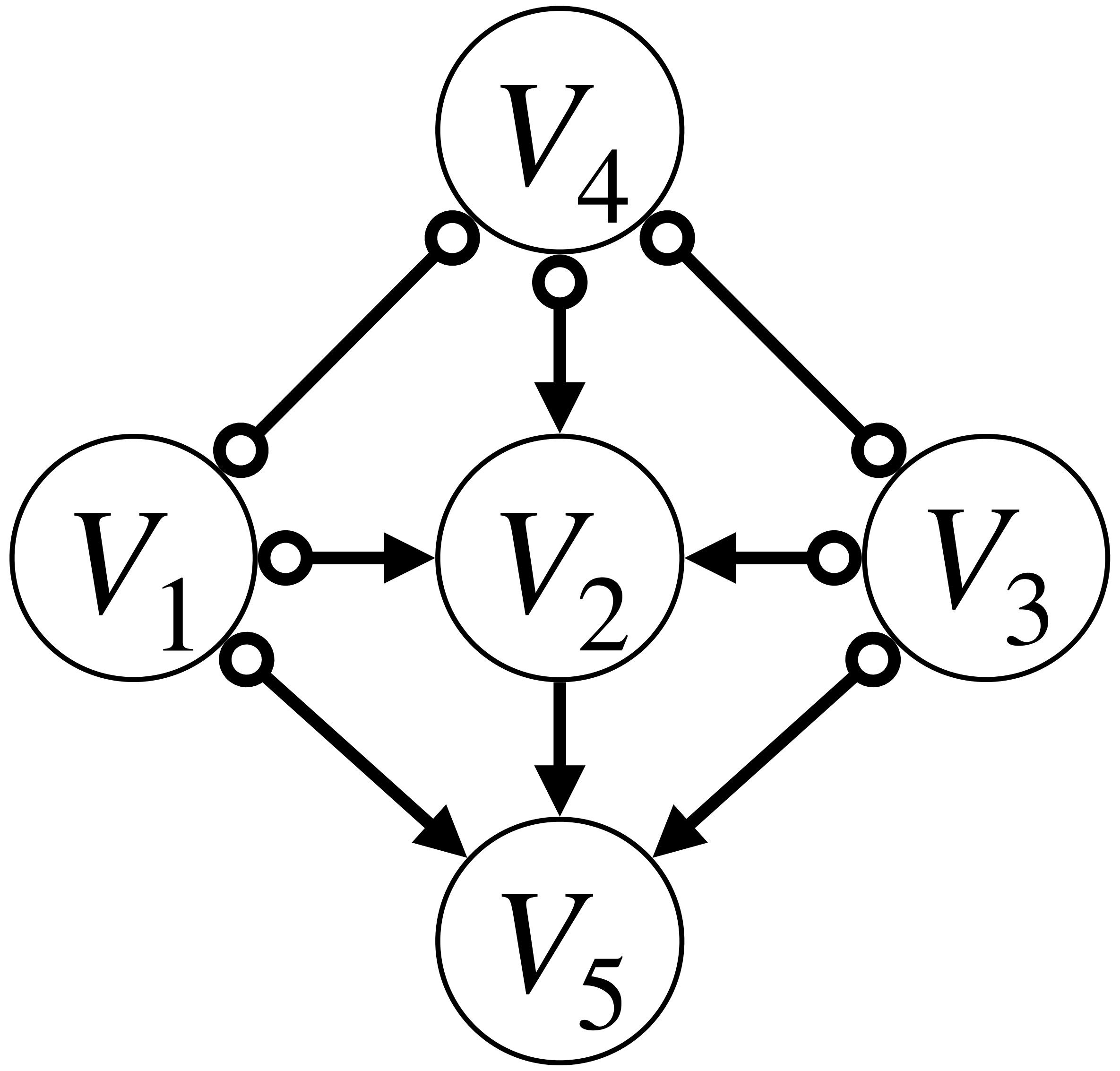
$X \in Sepset(W, Y)$

$\Longrightarrow$





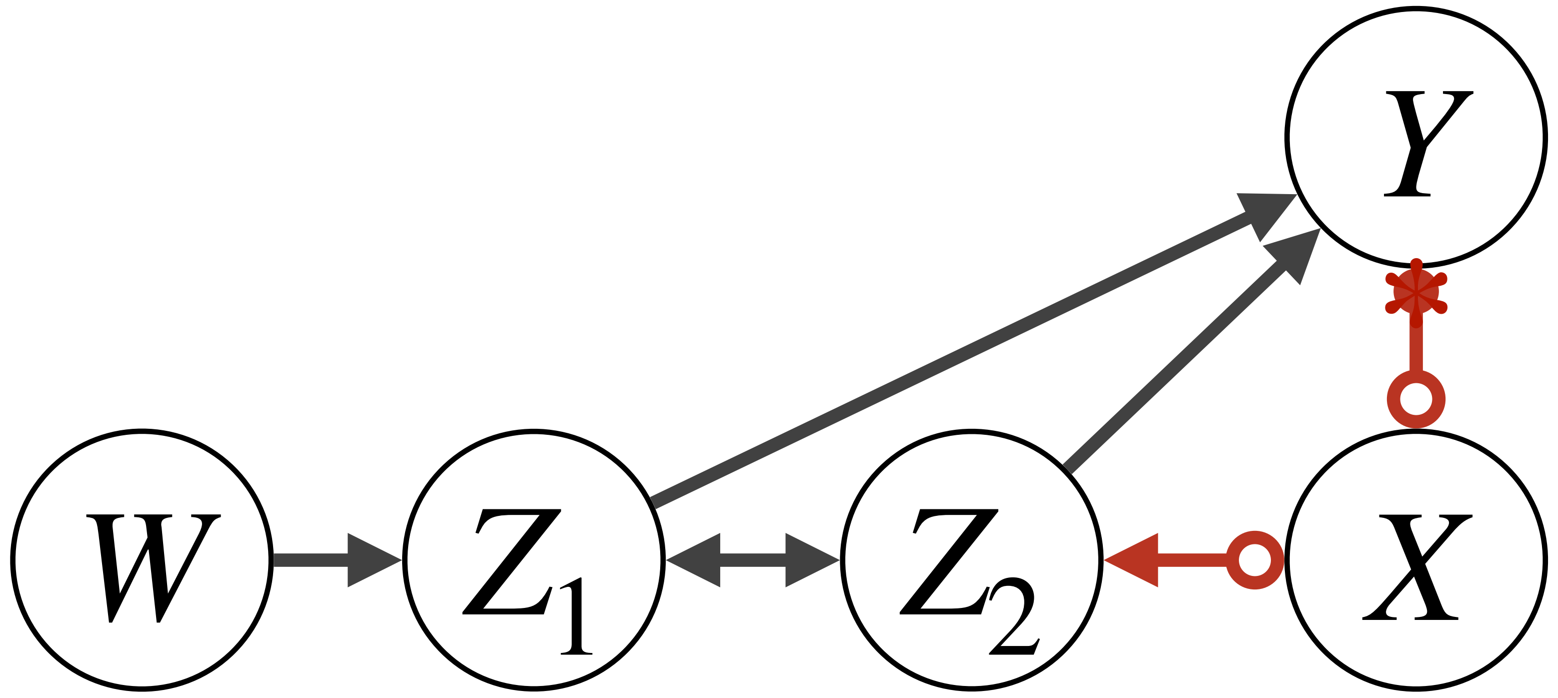
True, unknown ADMG



After Skel + R0 + R3 + R1

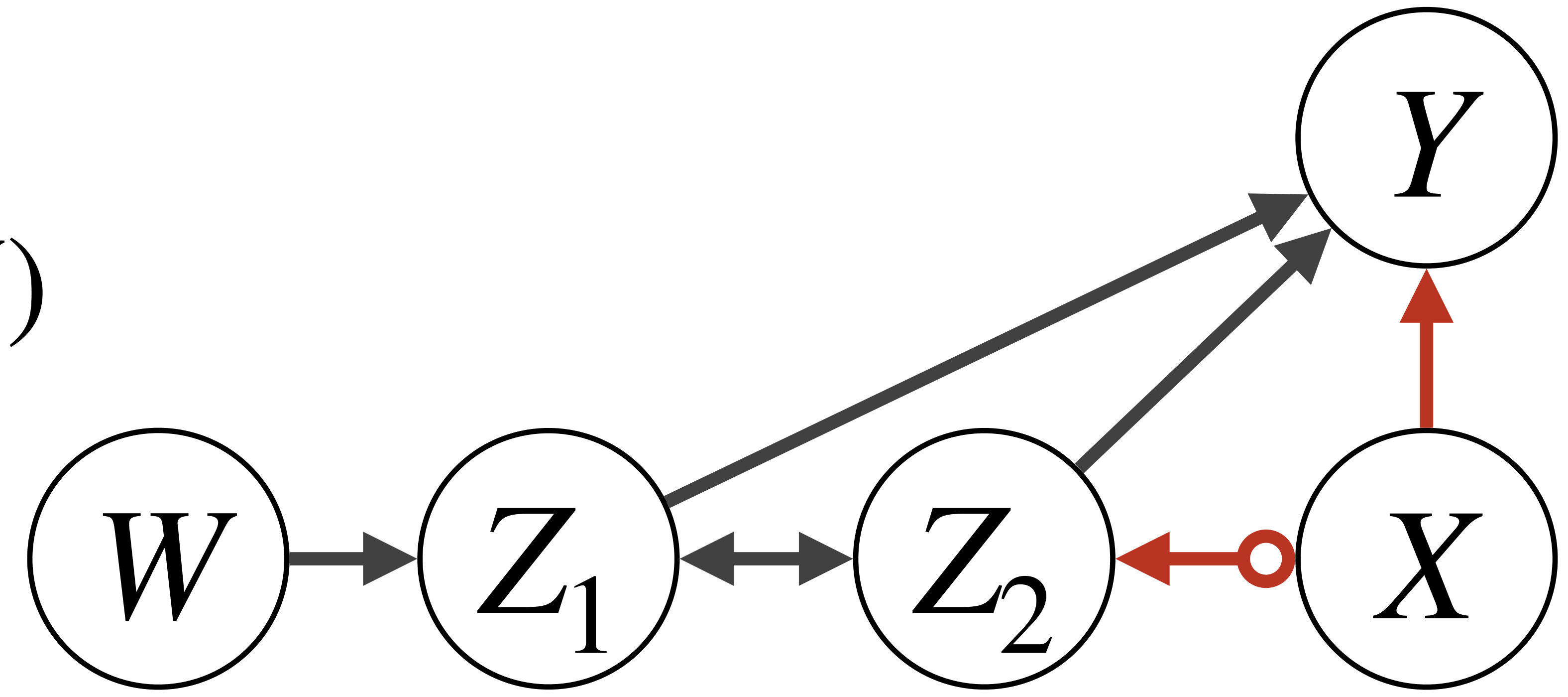
Applying Mark Inference Rules

R4:

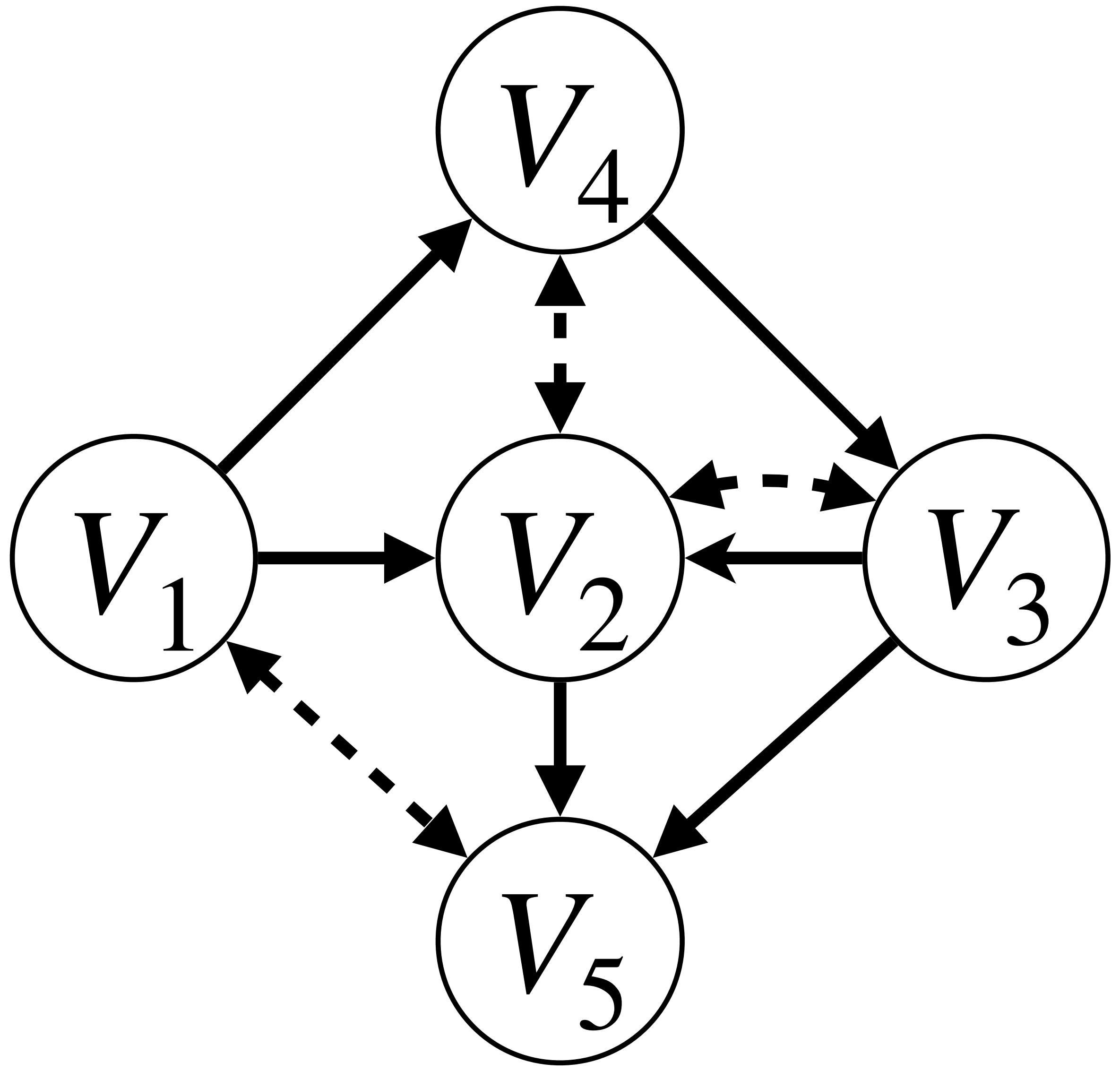


$X \in Sepset(W, Y)$

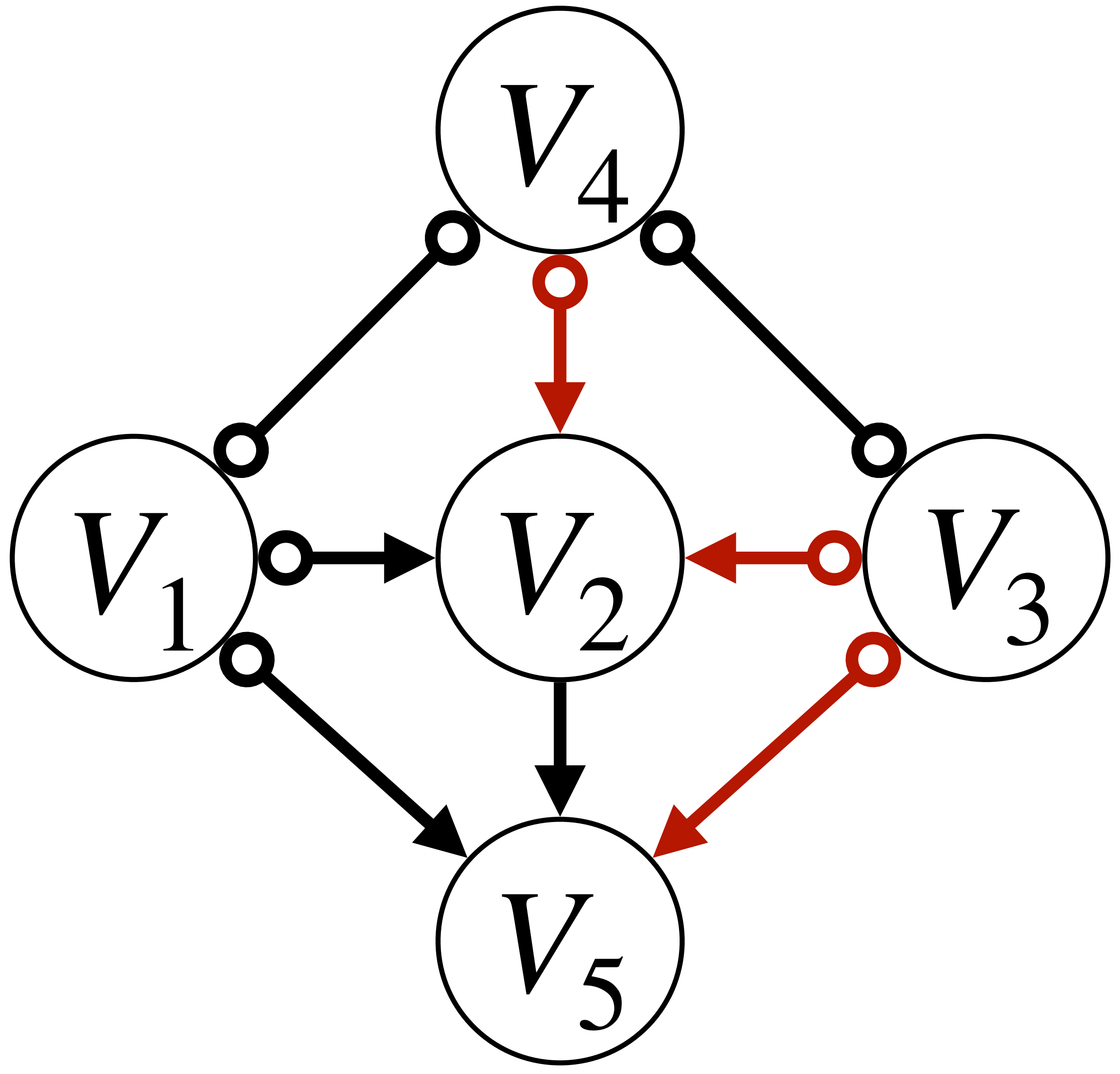
$\Longrightarrow$

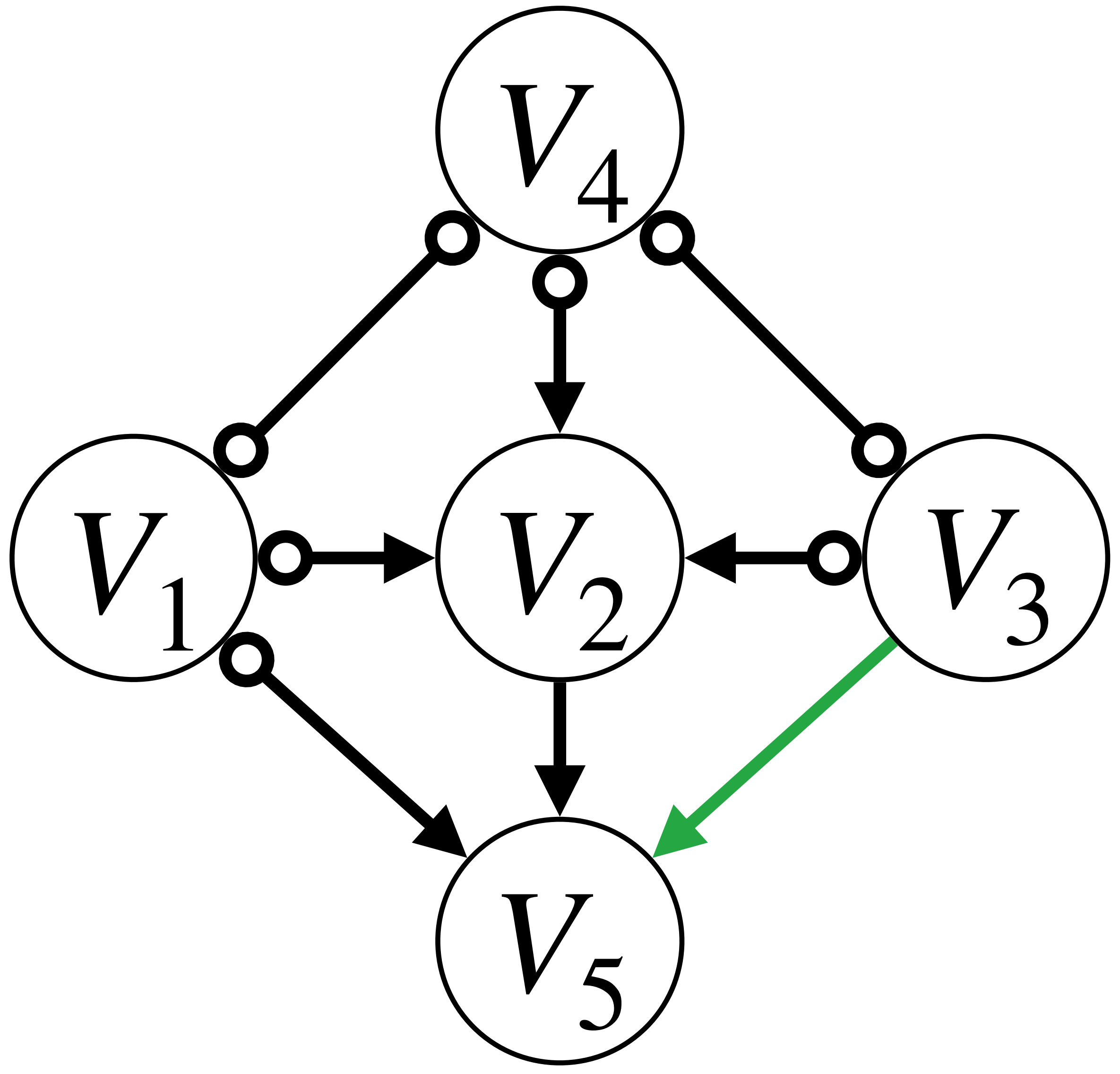


$\langle W, Z_1, Z_2, X, Y \rangle$ is a discriminating path for X



True, unknown ADMG



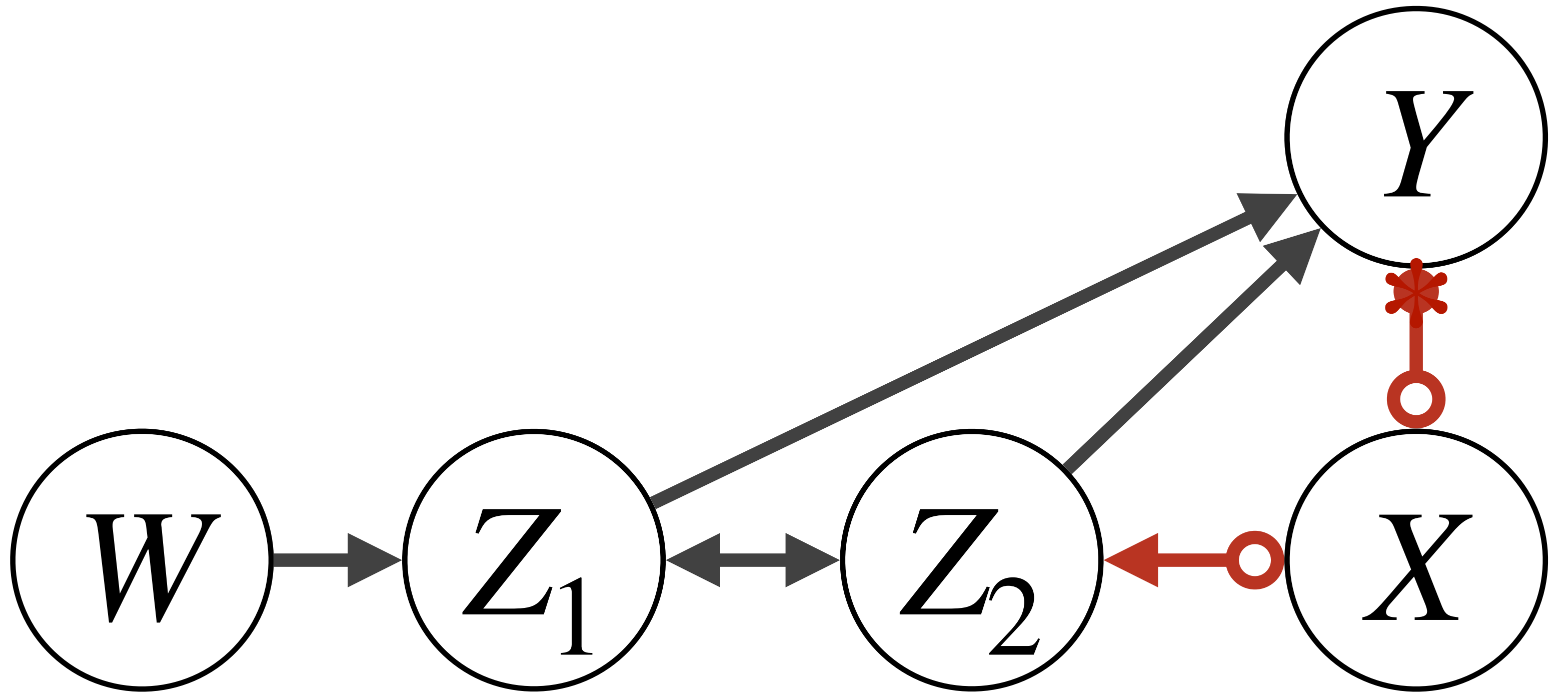


$\langle V_4, V_2, V_3, V_5 \rangle$ is a discriminating path for V_3 and

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

Applying Mark Inference Rules

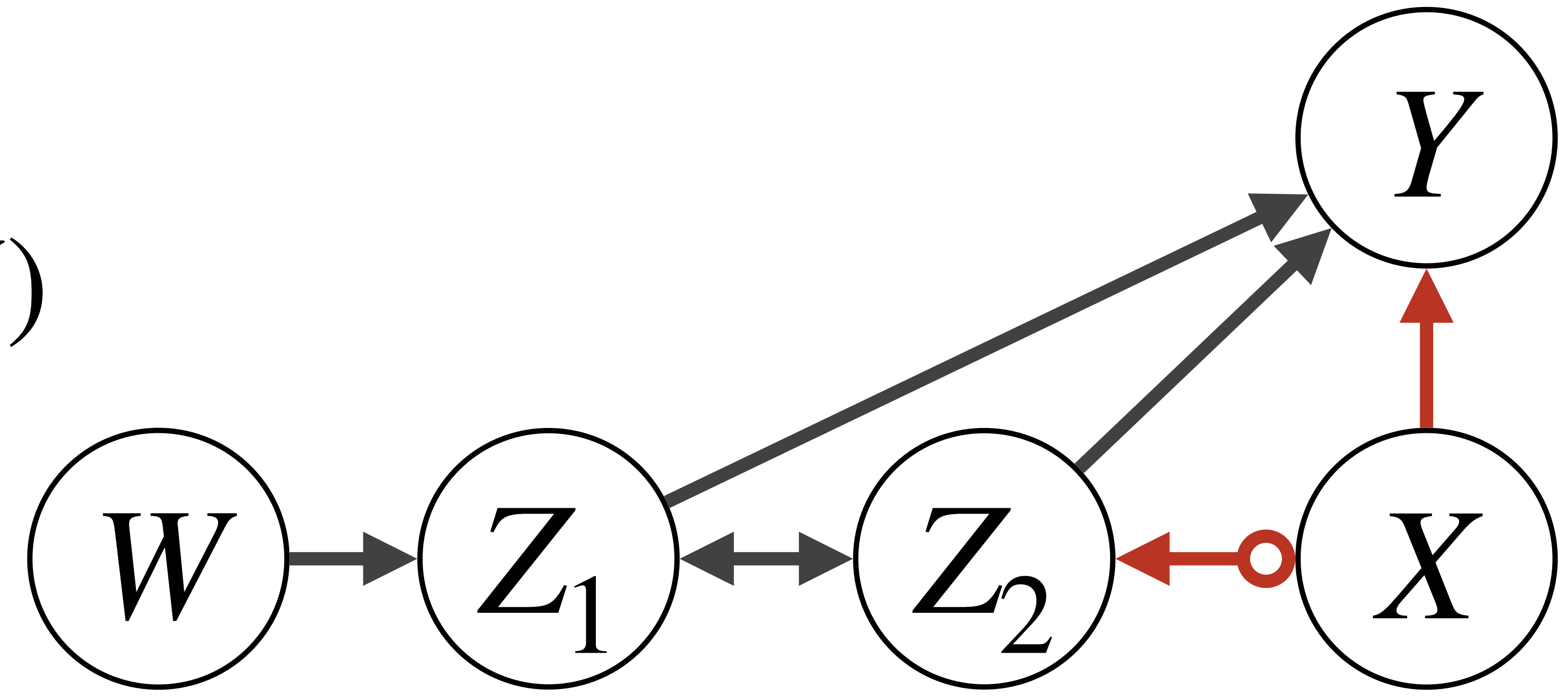
R4:

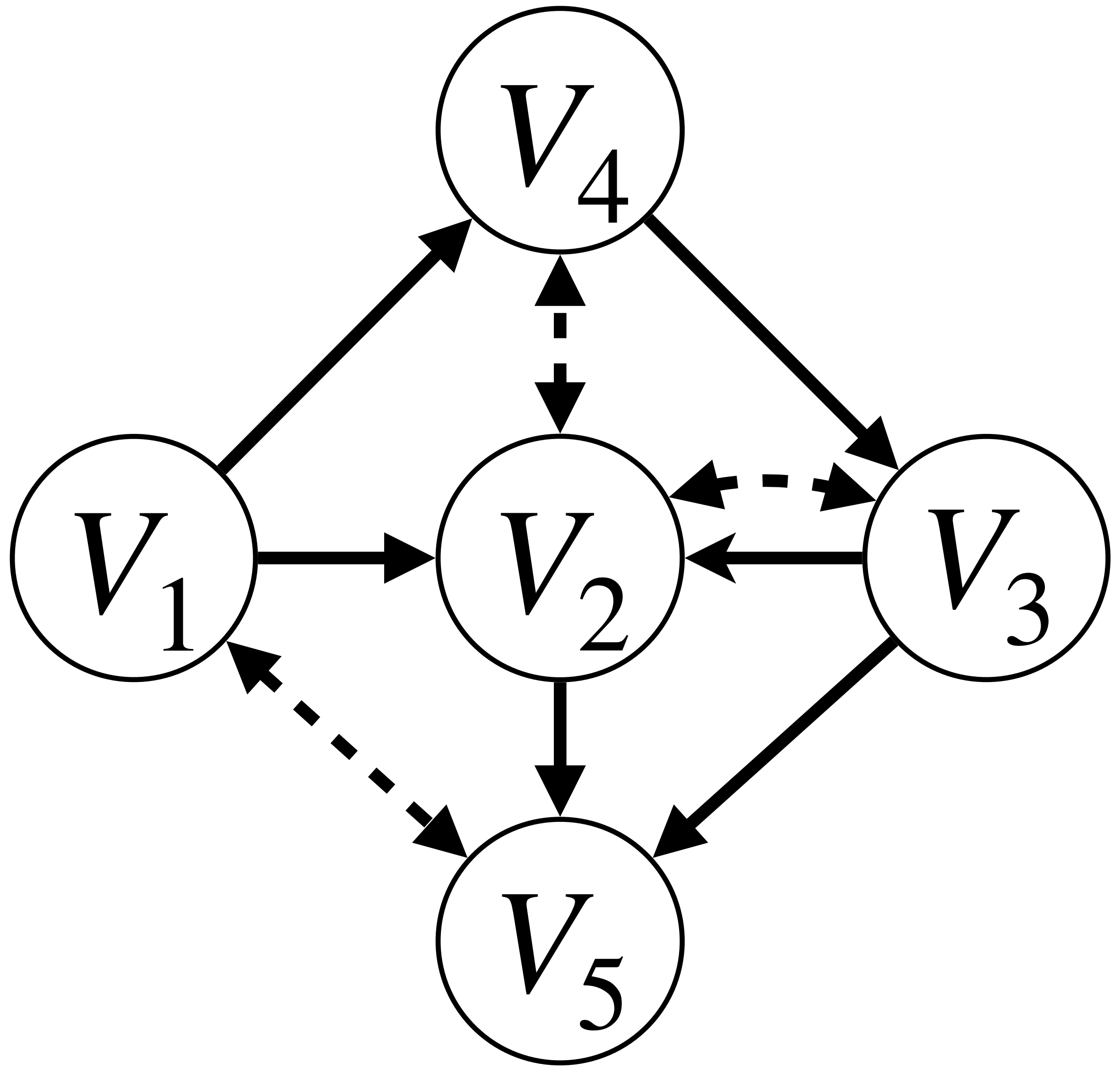


$\langle W, Z_1, Z_2, X, Y\rangle$ is a discriminating path for X

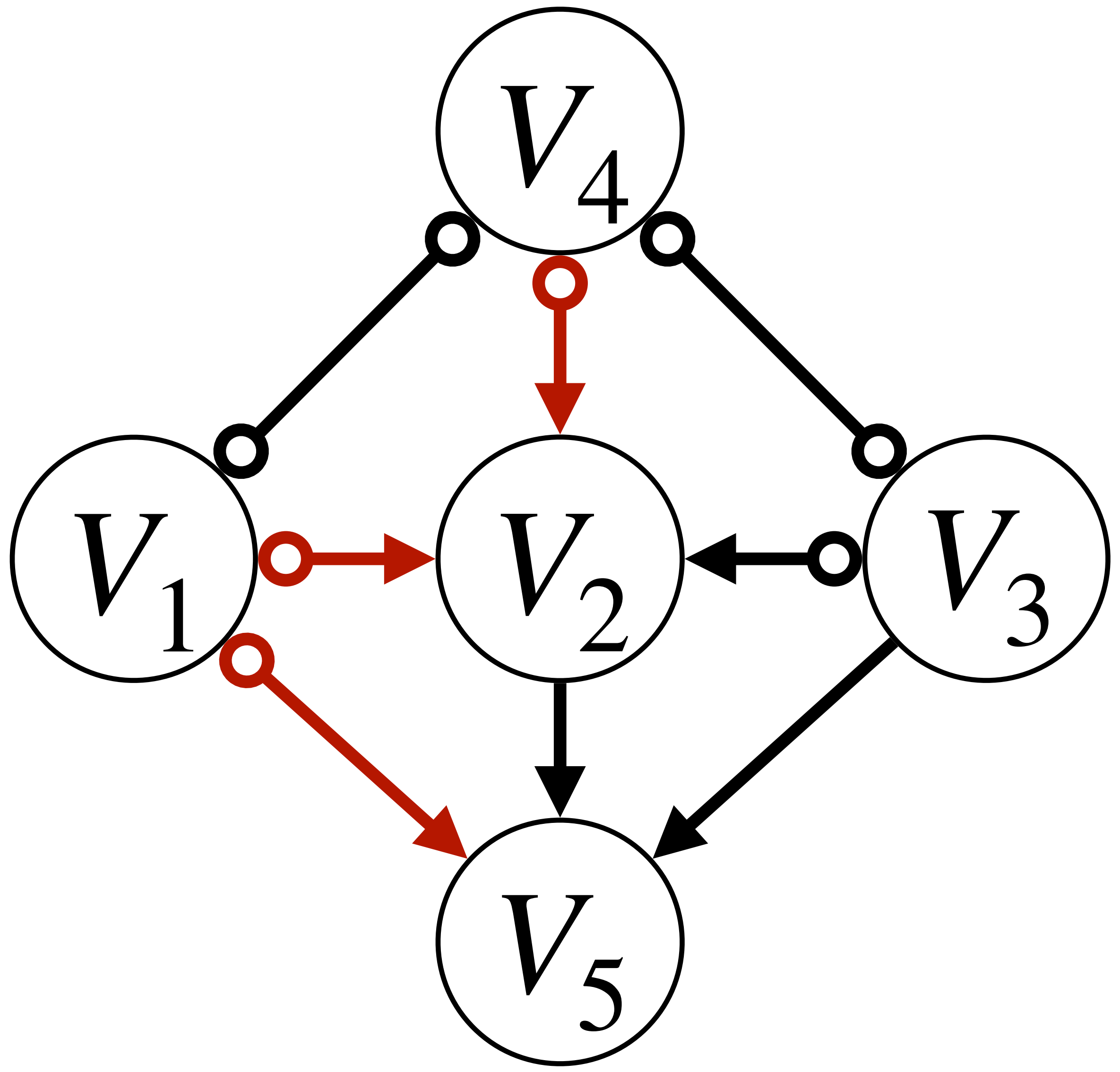
$X \in Sepset(W, Y)$

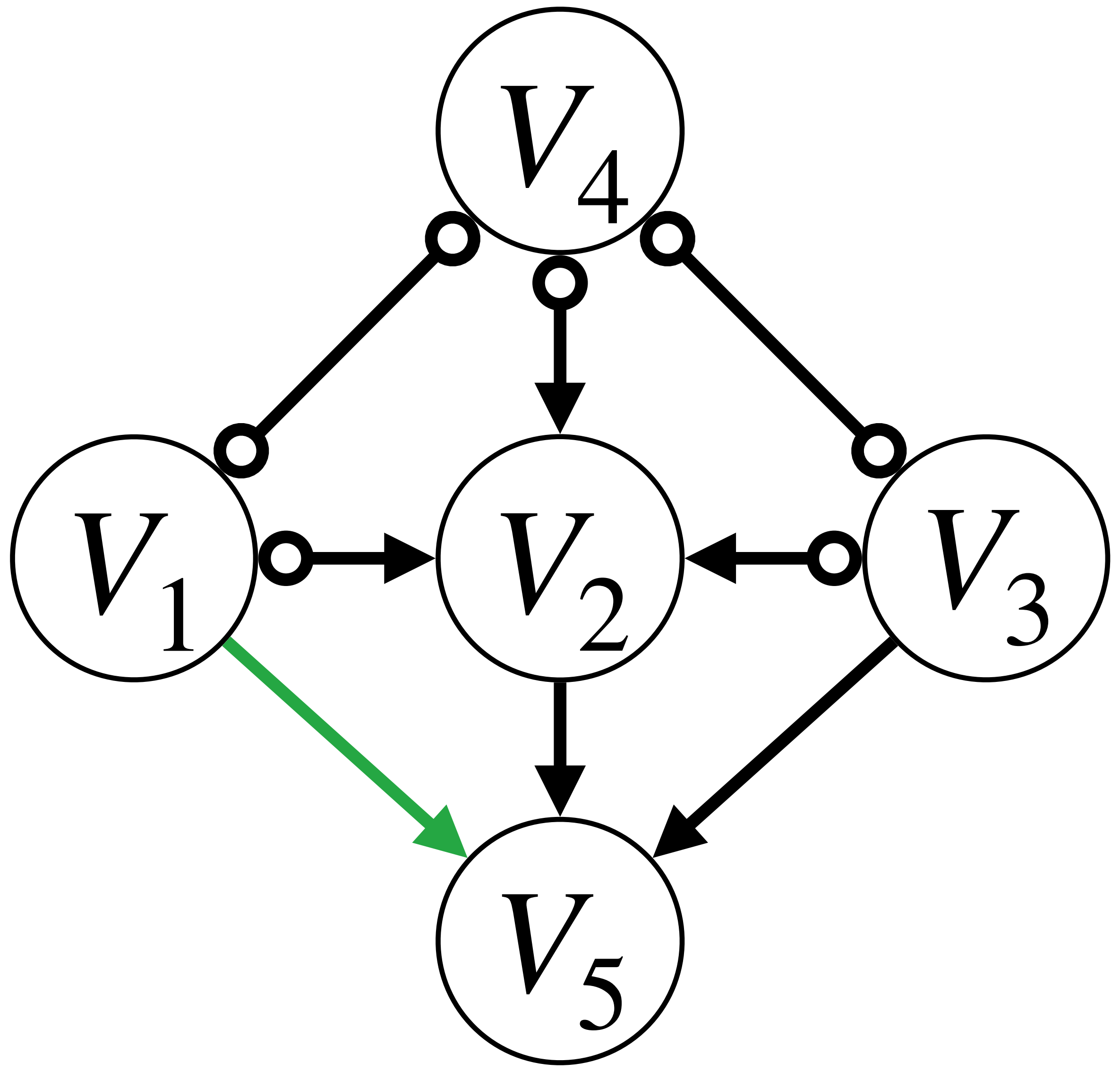
$\Longrightarrow$





True, unknown ADMG

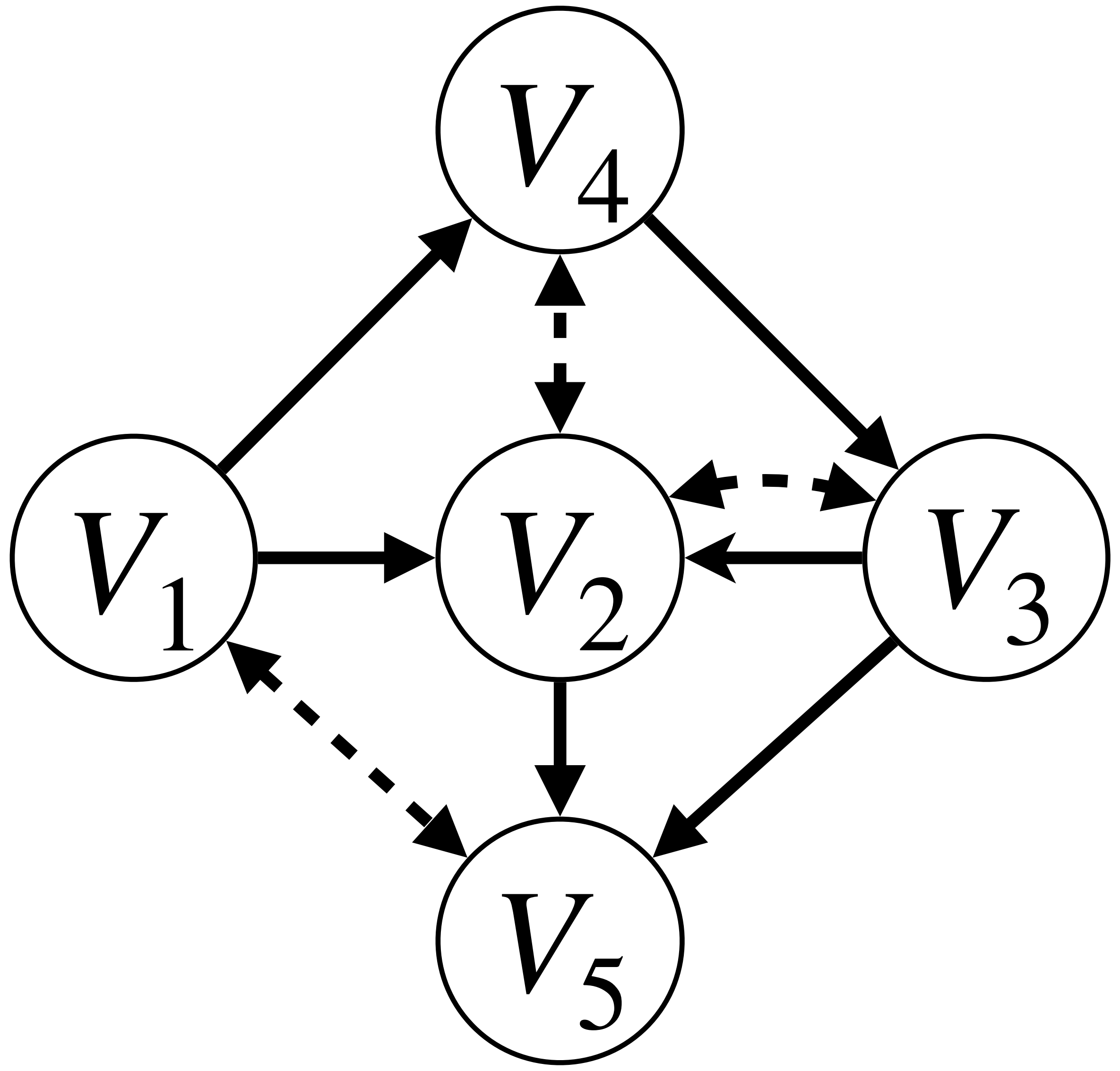




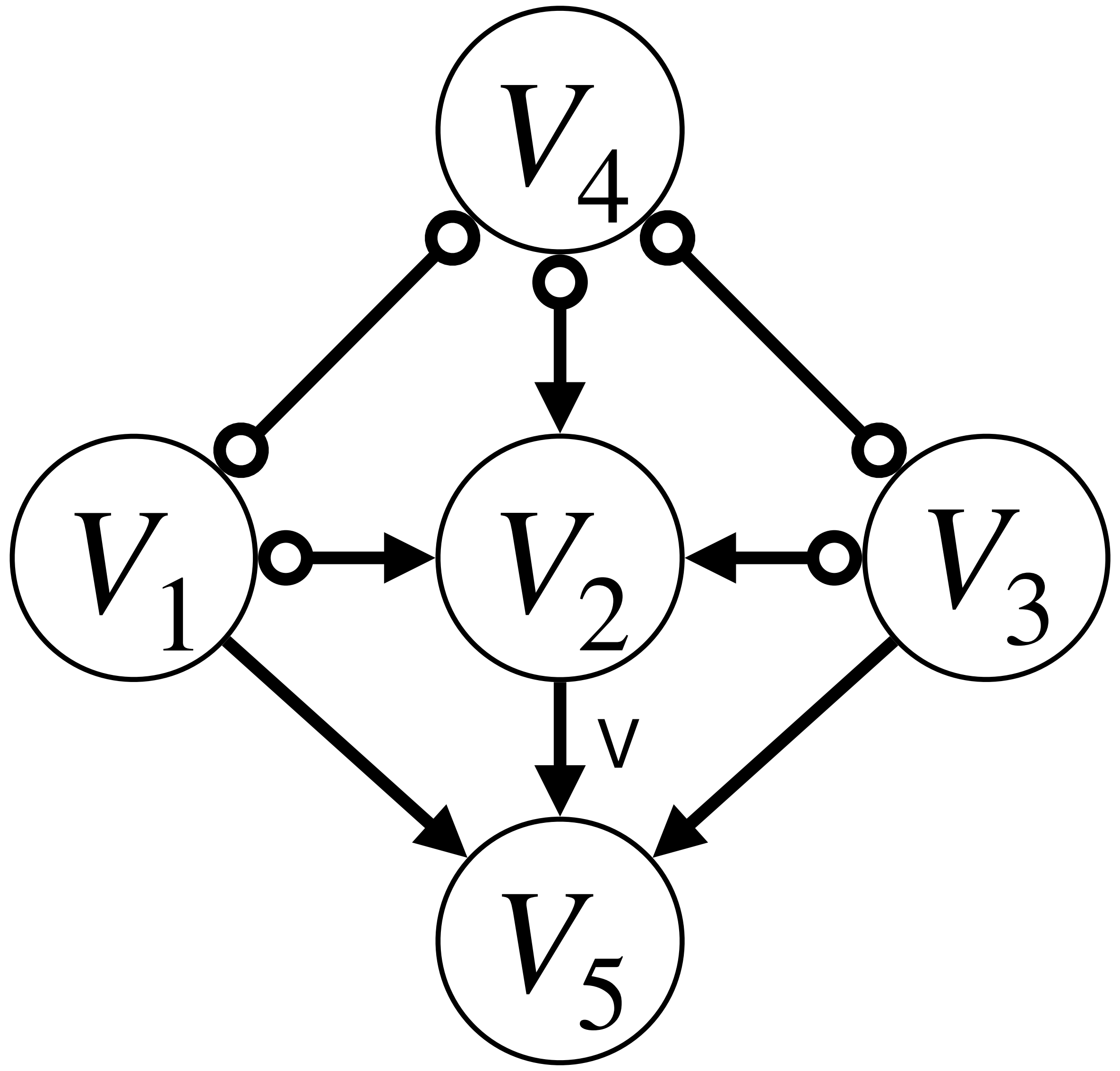
$\langle V_4, V_2, V_1, V_5\rangle$ is a discriminating path for V_1 and

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

Final PAG



True, unknown ADMG



Final PAG

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

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

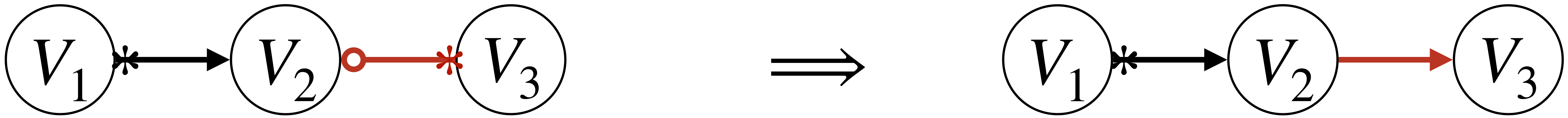
$V_4 \perp\!\!\!\perp V_5 \mid V_1, V_2, V_3$

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

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

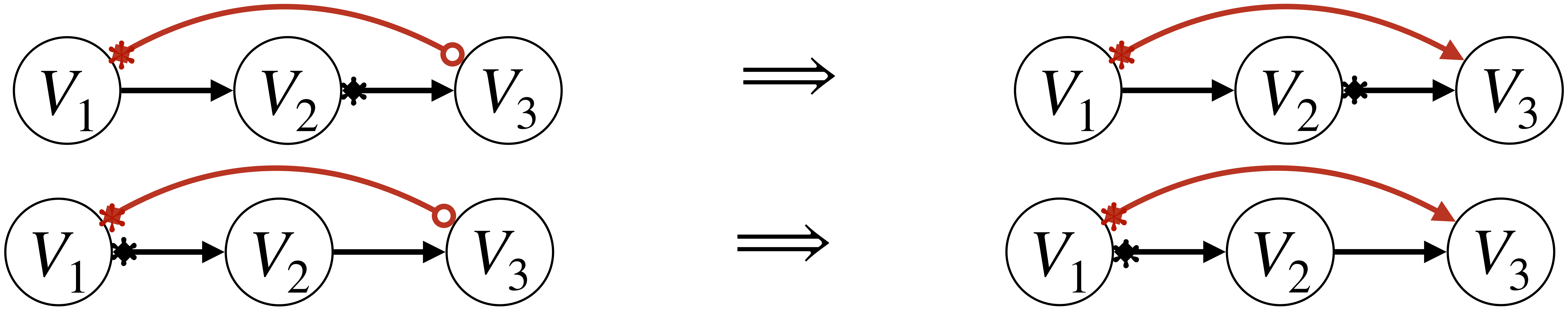
FCI - Complete Set of Mark Inference Rules

R1:

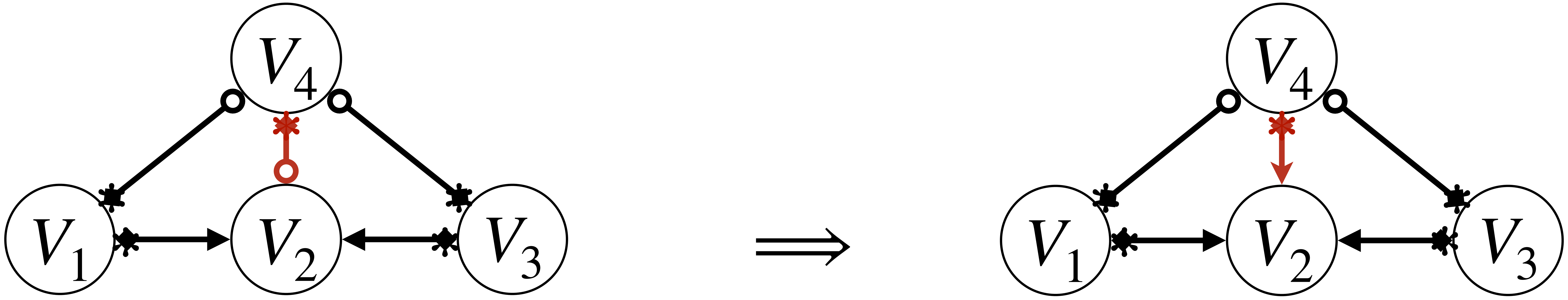


where V_1 and V_3 are not adjacent

R2:



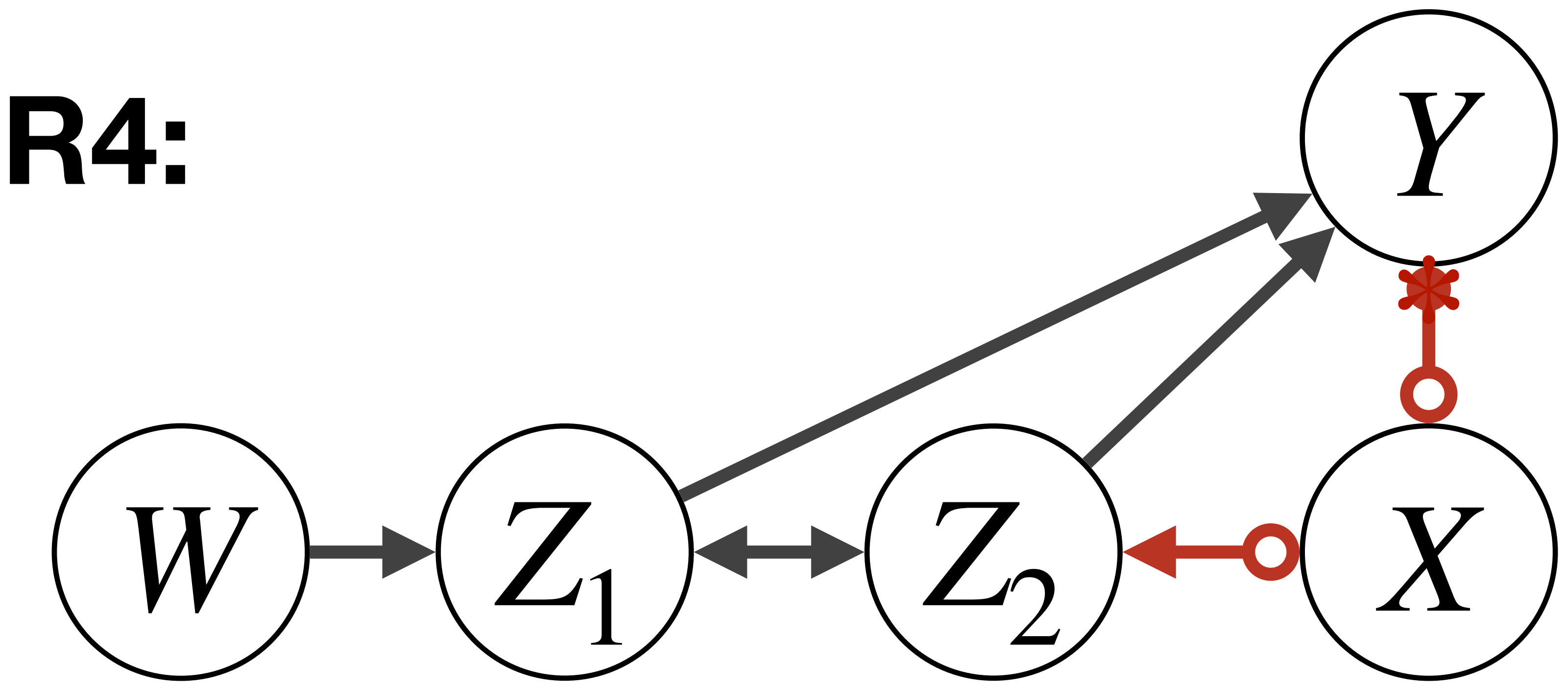
R3:



where V_1 and V_3 are not adjacent

FCI - Complete Set of Mark Inference Rules

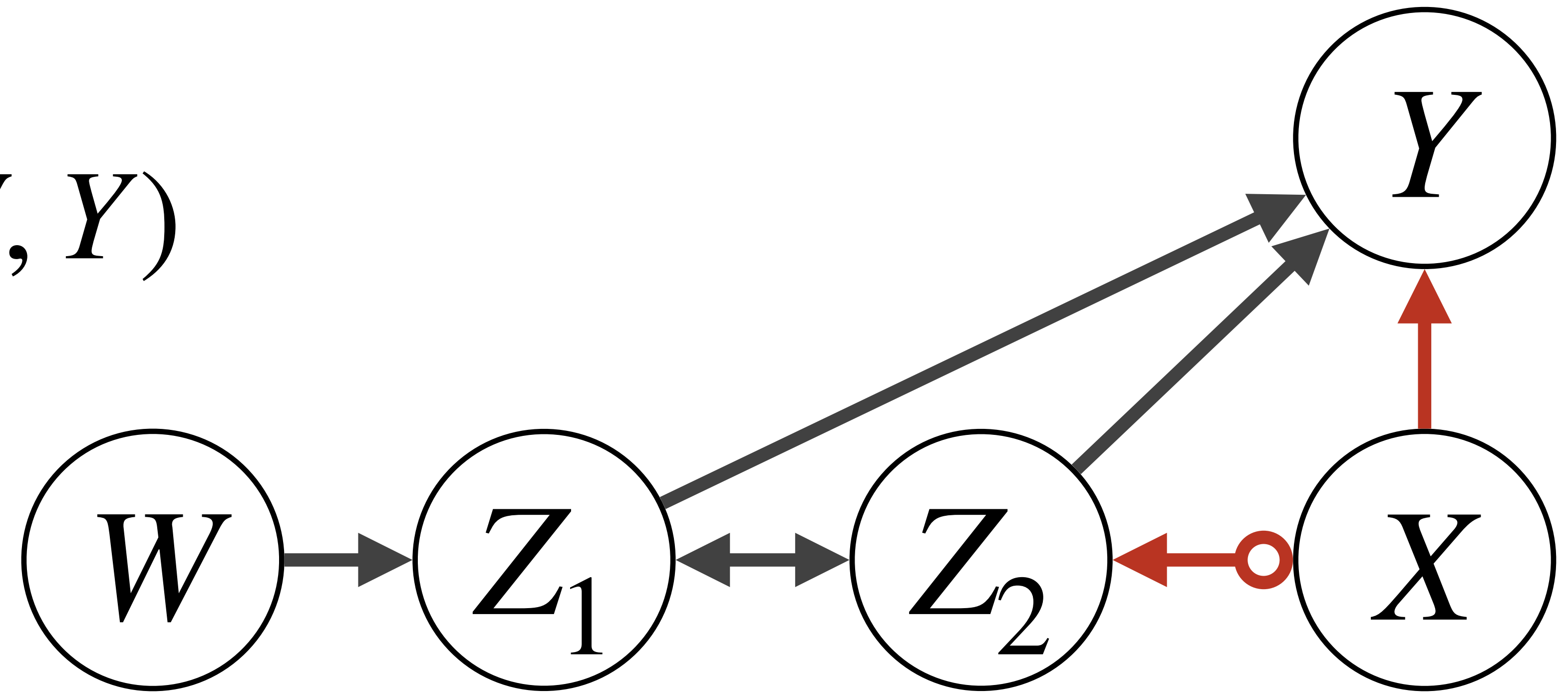
R4:



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

$X \in Sepset(W, Y)$

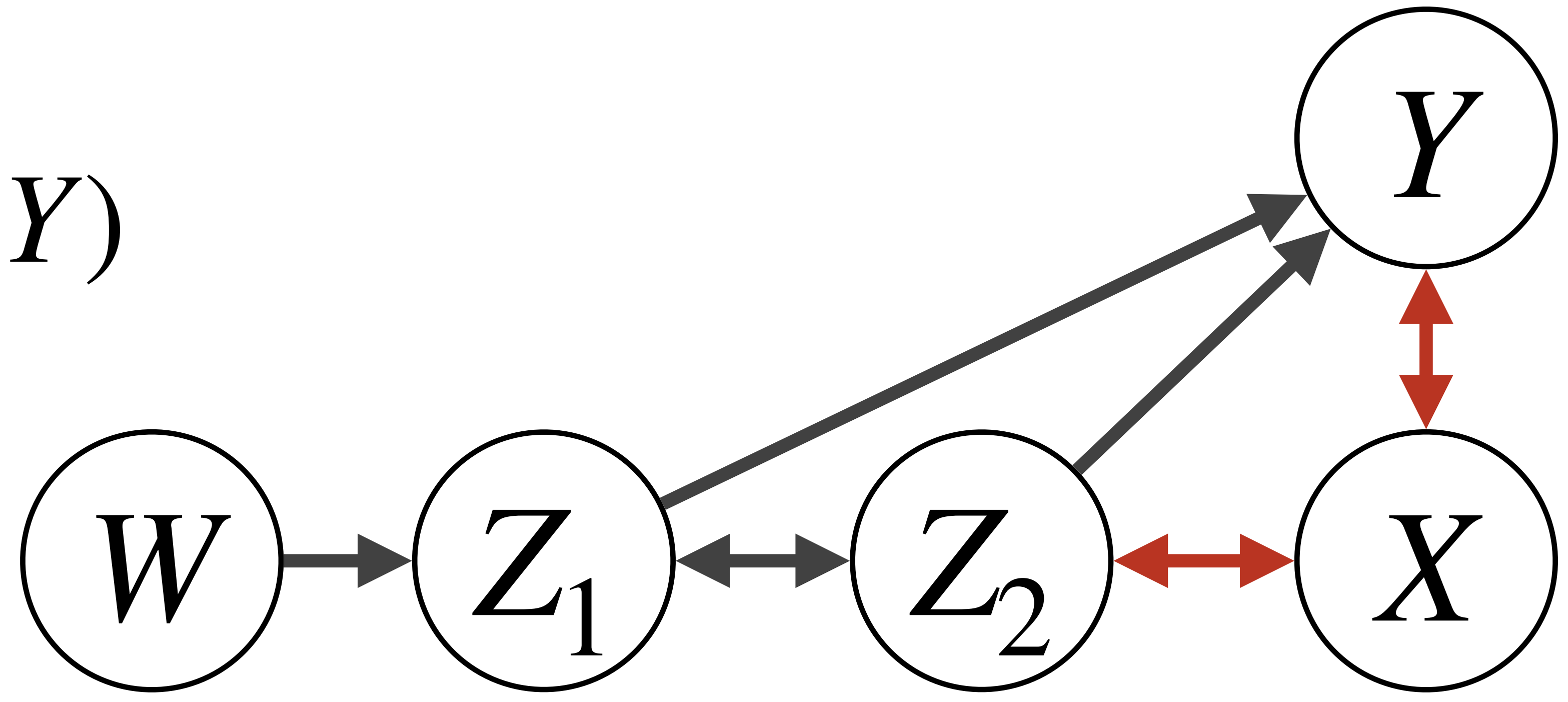
$\Longrightarrow$



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

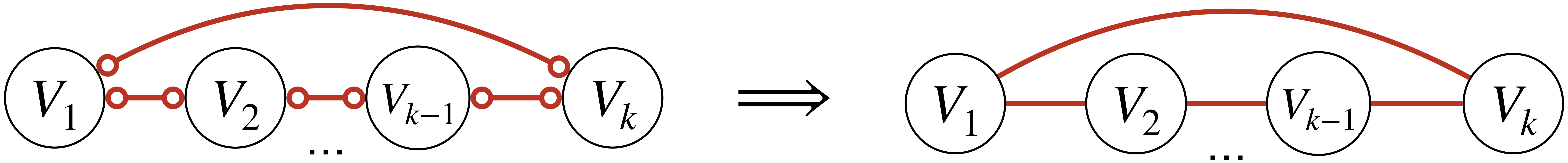
(i) p includes at least three edges;

(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:

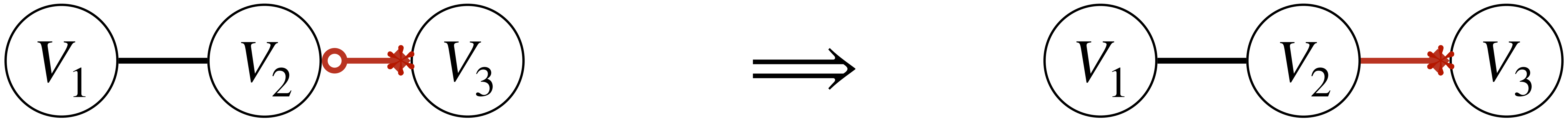


$\langle V_1, V_2, \dots, V_{k-1}, V_k \rangle$ is an uncovered circle path

V_1 and V_{k-1} are not adjacent

V_2 and V_k are not adjacent

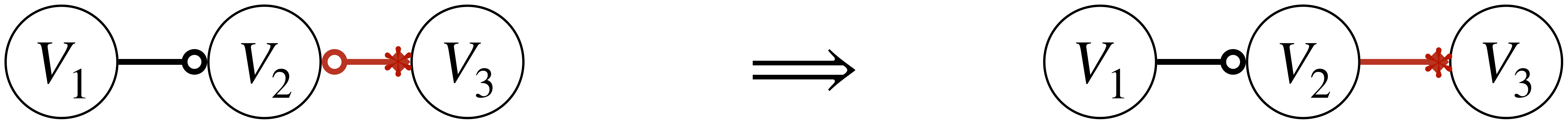
R6:



V_1 and V_3 may or may not be adjacent

FCI - Complete Set of Mark Inference Rules

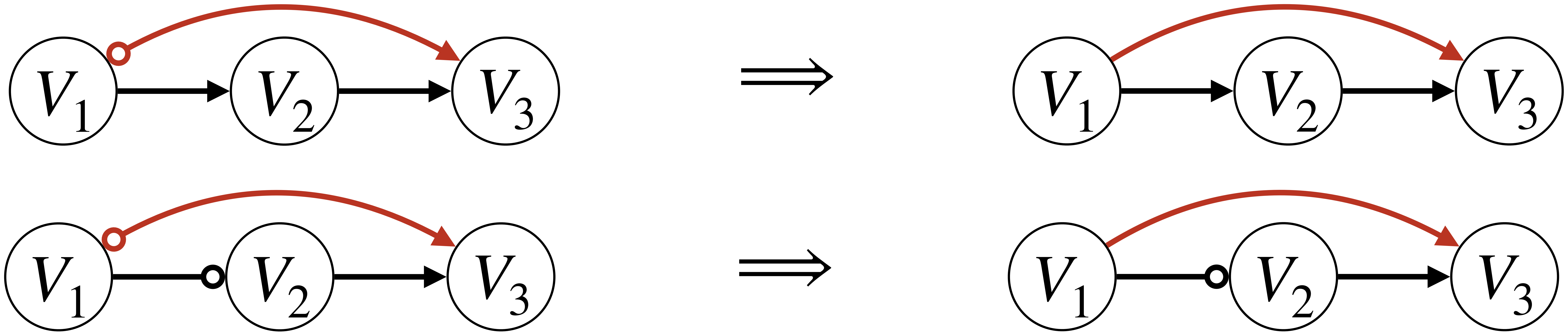
R7:



V_1 and V_3 are not adjacent

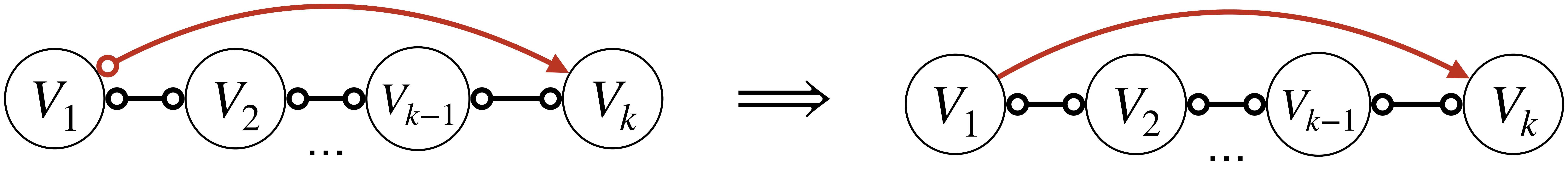
R8:

or



FCI - Complete Set of Mark Inference Rules

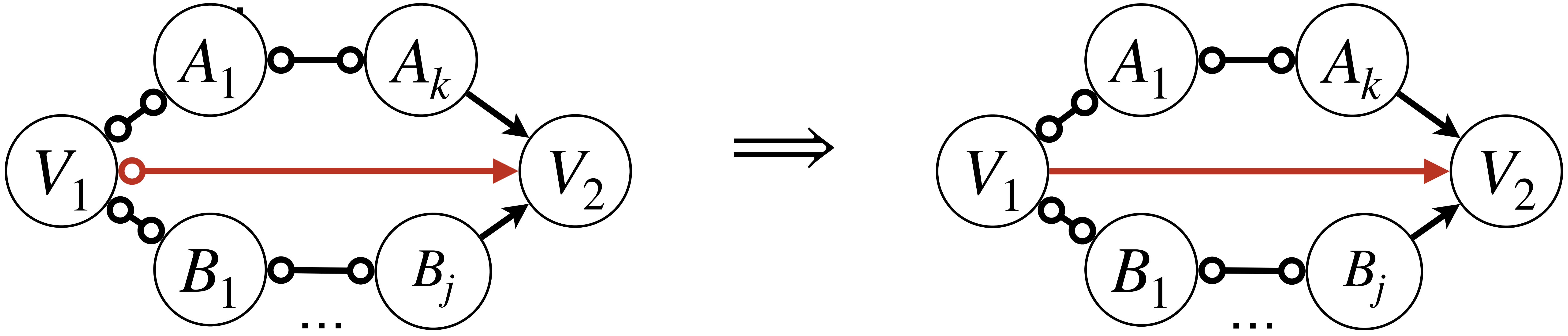
R9:



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

V_2 and V_k are not adjacent

R10:

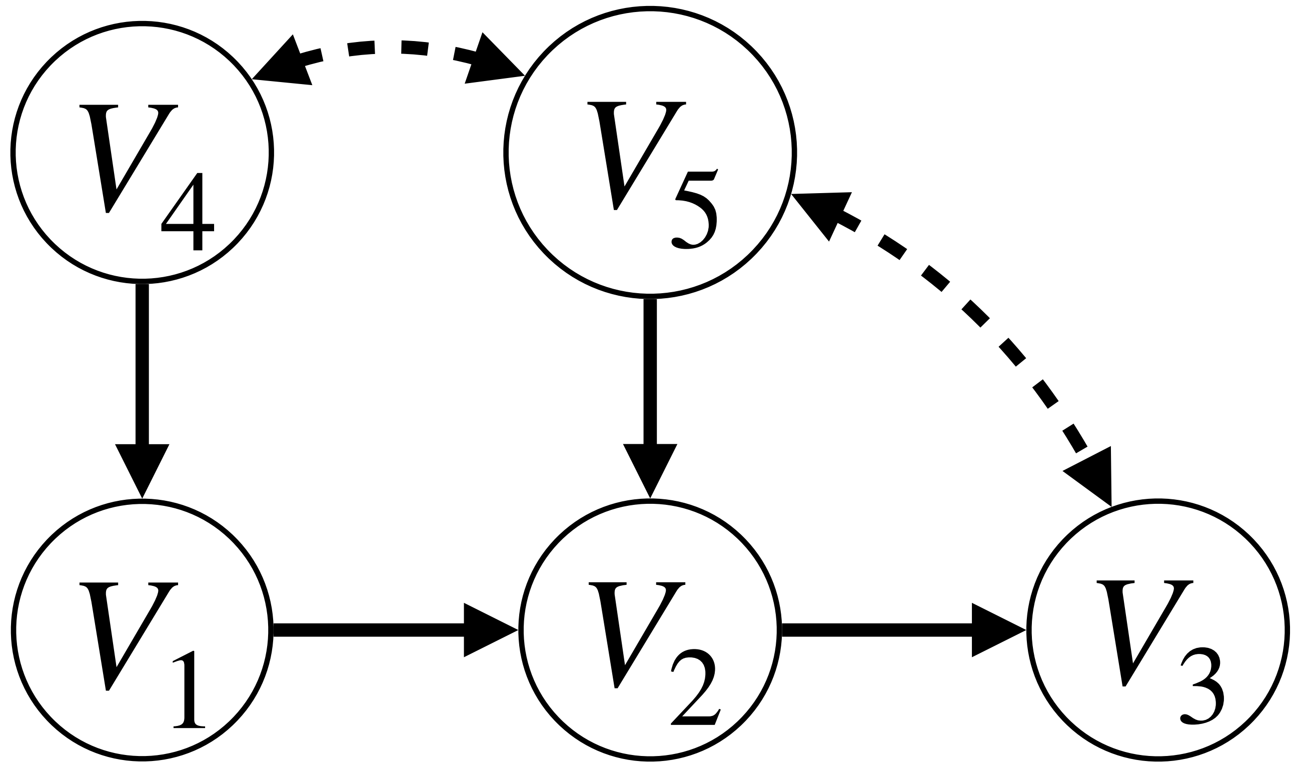


$\langle V_1, A_1, \ldots, A_k \rangle$ is an uncovered potentially directed path from V_1 to A_k (A_1 may be A_k)

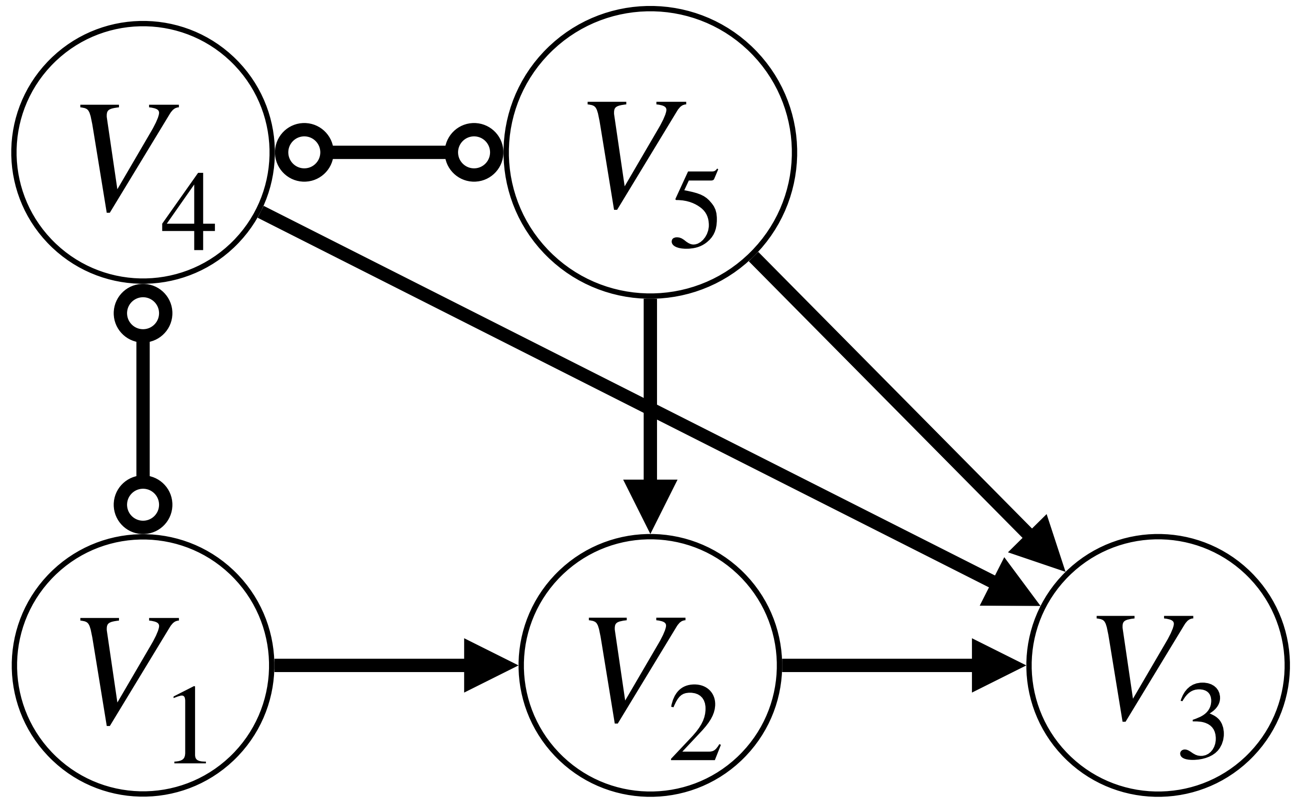
$\langle V_1, B_1, \ldots, B_k \rangle$ is an uncovered potentially directed path from V_1 to B_k (B_1 may be B_k)

$A_1 \neq B_1$ 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.

More on Causal Discovery and PAGs

Causal discovery from observational and experimental data:

- Kocaoglu, M., Jaber, A., Shanmugam, K., Bareinboim, E. Characterization and Learning of Causal Graphs with Latent Variables from Soft Interventions. In Proceedings of the 33rd Annual Conference on Neural Information Processing Systems. 2019. (Link)

- Jaber, A., Kocaoglu, M., Shanmugam, K., Bareinboim, E. Causal Discovery from Soft Interventions with Unknown Targets: Characterization & Learning. In Advances in Neural Information Processing Systems 2020. (Link)

Causal effect identification from PAGs:

- Jaber A., Ribeiro A. H., Zhang, J., Bareinboim, E. Causal Identification under Markov Equivalence - Calculus, Algorithm, and Completeness. In Proceedings of the 36th Annual Conference on Neural Information Processing Systems, NeurIPS 2022. (Link)

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

Thank you for your attention!

adele.ribeiro@uni-marburg.de