Causal Inference
Theory and Applications in Enterprise Computing

Christopher Hagedorn, Johannes Huegle, Dr. Michael Perscheid

May 05, 2020

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Agenda
May 05, 2020

- **Lecture Organization**
- **Embedding: Causal Inference in a Nutshell**
- **Introduction to Causal Graphical Models**

Lecture Organization

Lecture Organization
Topics to be Discussed

Q&A

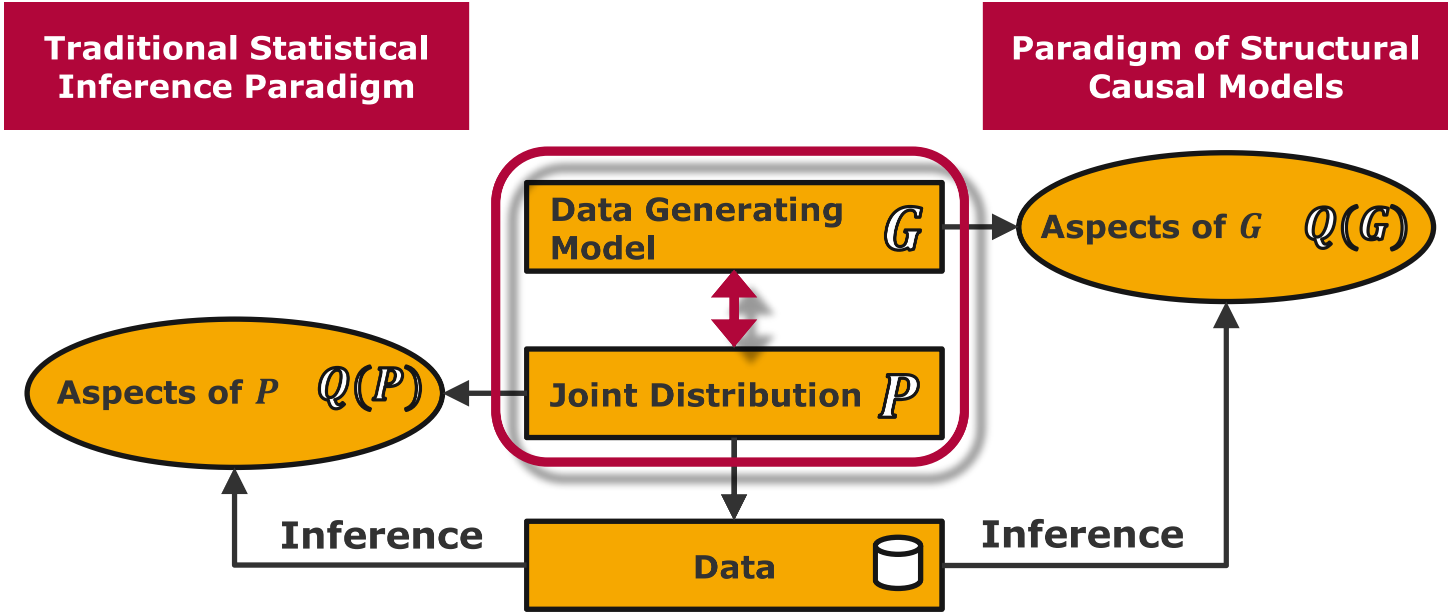
- Questions concerning Jupyter lab or R exercises?
- Open Questions concerning last week's lecture topics?

Dies Academicus (6th of May, postponed)

- Exercise is happening as intended

Embedding: Causal Inference in a Nutshell

Embedding: Causal Inference in a Nutshell
Concept



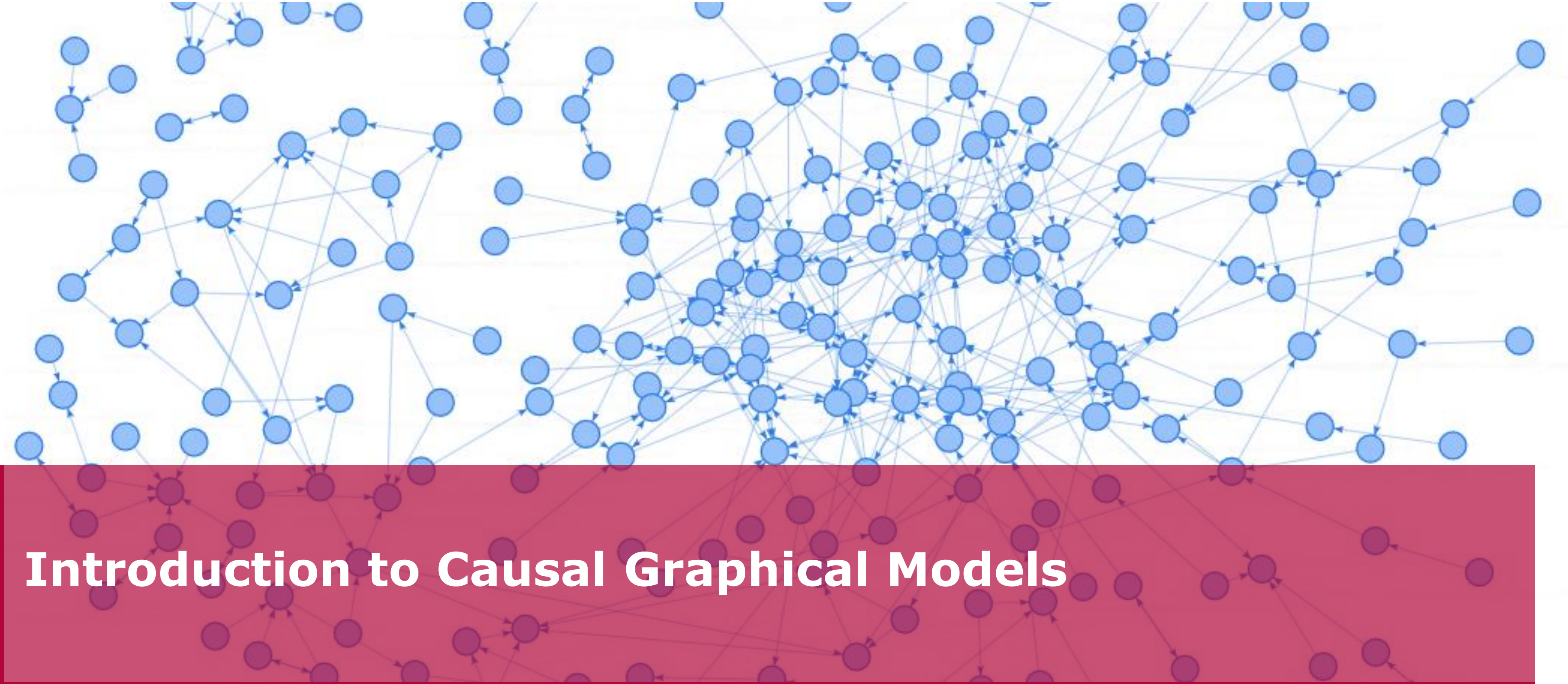
E.g., what is the sailors' probability of recovery when **we see** a treatment with lemons?

$$Q(P) = P(recovery|lemons)$$

E.g., what is the sailors' probability of recovery if **we do** treat them with lemons?

$$Q(G) = P(recovery|do(lemons))$$

HPI Hasso Plattner Institut
Digital Engineering • Universität Potsdam



Introduction to Causal Graphical Models

Introduction to Causal Graphical Models
Content

1. Preliminaries
Notation

- A, B, A_i events
- X, Y, Z, W, S, V_i sets of random variables
- x value of random variable

- $\Pr$ probability measure
- P_X probability distribution of X
- p density
- $p(X)$ density of P_X (always assume the existence of joint density, w.r.t. a product measure)
- $p(x)$ density of P_X evaluated at the point x

- $X \perp\!\!\!\perp Y$ independence of X and Y
- $X \perp\!\!\!\perp Y \mid Z$ conditional independence of X and Y given Z

- f, g, f_i functions of a function class $\mathcal{F}$

Independence of Events

- Two events A and B are called *independent*, if

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)\,,$$

or - rewritten in *conditional probabilities* - *if*

$$\Pr(\mathrm{A}) = \frac{A \cap B}{B} = \Pr(A|B)\,,$$
$$\Pr(\mathrm{B}) = \frac{A \cap B}{A} = \Pr(B|A).$$

- $A_1, \dots, A_N$ are called *(mutually) independent* if for every subset $S \subset \{1, \dots, N\}$ we have

$$\Pr\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} \Pr(A_i)\,.$$

- **Note:**
for $N \geq 3$, pairwise independence $\Pr\left(A_i \cap A_j\right) = \Pr(A_i) \cdot \Pr\left(A_j\right)$ for all i, j where $i, j = 1, \dots, N$, and $i \neq j$ does not imply (mutual) independence.

1. Preliminaries
Independence of Random Variables

- Two real-valued random variables X and Y are called *independent*,
$$X \perp\!\!\!\perp Y,$$
if for every $x, y \in \mathbb{R}$, the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent,

  Or, in terms of densities: for all x, y,
$$p(x, y) = p(x)p(y).$$

- **Note:**
If $X \perp\!\!\!\perp Y$, then $\mathrm{E}[\mathrm{X\ Y}] = \mathrm{E}[\mathrm{X}]\ \mathrm{E}[\mathrm{Y}]$, and $cov(X, Y) = E[X\ Y] - E[X]\ E[Y] = 0$, i.e.,
$$X \perp\!\!\!\perp Y \ \Rightarrow\ cov(X, Y) = 0.$$
But: $cov(X, Y) = 0 \nRightarrow X \perp\!\!\!\perp Y$.

No correlation does not imply independence

However, we have, for large $\mathcal{F}$: $\left(\forall f, g \in \mathcal{F}: cov\big(f(X), g(Y)\big) = 0\right)$, then $X \perp\!\!\!\perp Y$.

- Two real-valued random variables X and Y are called *conditionally independent* given Z,

$$X \perp\!\!\!\perp Y \mid Z \quad \text{or} \quad (X \perp\!\!\!\perp Y \mid Z)_P$$

  if

$$p(x, y|z) = p(x|z)\, p(y|z)$$

  for all x, y and for all z s.t. $p(z) > 0$.

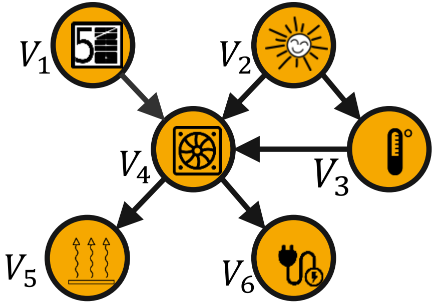
- **Note:**
  It is possible to find X, Y which are conditionally independent given a variable Z but unconditionally dependent, and vice versa.

2. Causal Graphical Models
Definition (Pearl)

- Directed Acyclic Graph (DAG) $G = (V, E)$
  - *Vertices* V_i, $i = 1, \dots, N$
  - *Directed edges* $E = (V_i, V_j)$, i.e., $V_i \rightarrow V_j$
  - *No cycles*
- Use kinship terminology, e.g., for path $V_i \rightarrow V_j \rightarrow V_k$
  - $V_i = Pa(V_j)$ *parent* of V_j
  - $\{V_i, V_j\} = Ang(V_k)$ *ancestors* of V_k
  - $\{V_j, V_k\} = Des(V_i)$ *descendants* of V_i
- Directed Edges encode *direct causes* via
  - $V_i = f_i(Pa(V_i), N_i)$ with independent noise N_i

➡ This forms the Causal Graphical Model

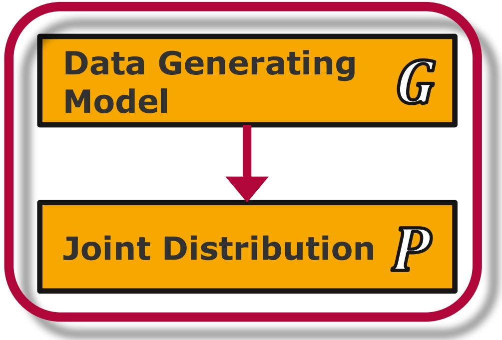
Cooling House Example:



- $V_1 = \mathcal{N}(0,1)$
- $V_2 = \mathcal{N}(0,1)$
- $V_3 = 3\,V_2 + \mathcal{N}(0,1)$
- $V_4 = 4\,V_1 + 5\,V_2 + 0.7\,V_3 + \mathcal{N}(0,1)$
- $V_5 = V_4 + \mathcal{N}(0,1)$
- $V_6 = 1.2\,V_4 + \mathcal{N}(0,1)$

2. Causal Graphical Models
Connecting G and P

- Basic Assumption: *Causal Sufficiency*
  - All relevant variables are included in the DAG G



$$(X \perp\!\!\!\perp Y|Z)_G \Rightarrow (X \perp\!\!\!\perp Y|Z)_P$$

- Key Postulate: *(Local) Markov Condition*
- Essential mathematical concept: *d-Separation*

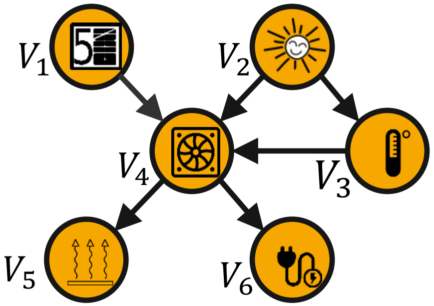
  (describes the conditional independences required by a causal DAG)

3. (Local) Markov Condition
Theorem

(Local) Markov Condition:
V_j independent of nondescendants $ND(V_j)$, given parents $Pa(V_j)$, i.e.,

$$\boldsymbol{V_j \perp\!\!\!\perp V_{V/(Des(V_j)\cup Pa(V_j))} | Pa(V_j)}.$$

- I.e., every information exchange with its nondescendants involves its parents
- Example:



- $V_6 \perp\!\!\!\perp \{V_1, V_2, V_3, V_5\} | V_4$
- $V_5 \perp\!\!\!\perp \{V_1, V_2, V_3, V_6\} | V_4$

3. (Local) Markov Condition
Supplement (Lauritzen 1996)

- Assume V_N has no descendants, then $ND(V_N) = \{V_1, \dots, V_{N-1}\}$.

- Thus the local Markov condition implies
$$V_N \perp\!\!\!\perp \{V_1, \dots, V_{N-1}\}/Pa(V_N) \mid Pa(V_N).$$

- Hence, the general decomposition
$$p(v_1, \dots, v_N) = p(v_N|v_1, \dots, v_{N-1})\, p(v_1, \dots, v_{N-1})$$
becomes
$$p(v_1, \dots, v_N) = p(v_N|Pa(v_N))\, p\big(\{v_1, \dots, v_{N-1}\}/Pa(v_N)\big).$$

- Induction over N yields to
$$p(v_1, \dots, v_N) = \prod_{i=1}^{N} p(v_i|Pa(v_i)).$$

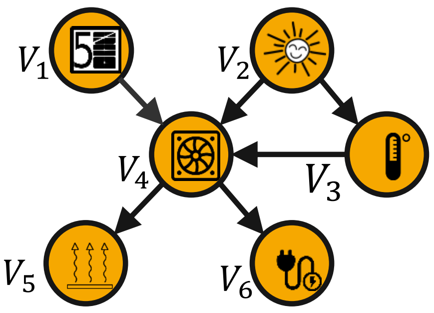
- I.e., the graph shows us how to factor the joint distribution P_V.

4. Factorization
Definition

Factorization:

$$p(v_1, \dots, v_N) = \prod_{i=1}^{N} p(v_i | Pa(v_i)).$$

- I.e., conditionals as causal mechanisms generating statistical dependence
- Example:



$$\begin{aligned} & p(v) \\ & = p(v_1, \dots, v_6) \\ & = p(v_1) \cdot p(v_2) \\ & \quad \cdot p(v_3|v_2) \cdot p(v_4|v_1, v_2, v_3) \\ & \quad \cdot p(v_5|v_4) \cdot p(v_6|v_4) \\ & = \prod_{i=1}^{6} p(v_i|Pa(v_i)) \end{aligned}$$

5. Global Markov Condition
D-Separation (Pearl 1988)

- *Path* = sequence of pairwise distinct vertices where consecutive ones are adjacent

- A path q is said to be *blocked* by a set S if
  - q contains a *chain* $V_i \rightarrow V_j \rightarrow V_k$ or a *fork* $V_i \leftarrow V_j \rightarrow V_k$ such that the middle node is in S, or
  - q contains a *collider* $V_i \rightarrow V_j \leftarrow V_k$ such that the middle node is not in S and such that no descendant of V_j is in S.

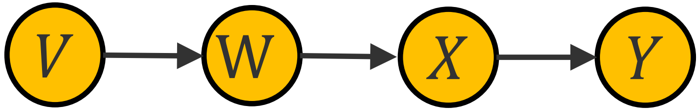
D-Separation:
S is said to **d-separate X and Y** in the DAG G, i.e.,

$$(X \perp\!\!\!\perp Y \mid S)_G,$$

if S blocks every path from a vertex in X to a vertex in Y.

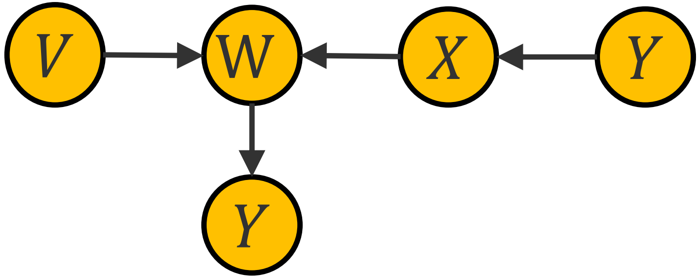
5. Global Markov Condition
Blocking of Paths (I/II)

- **Example:** Blocking of paths



  - Path from V to Y is blocked by conditioning on W, X, or $\{W, X\}$.

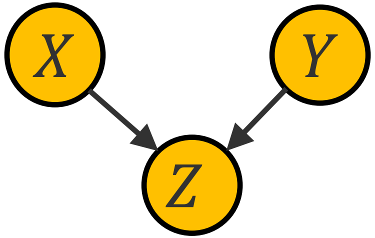
- **Example:** Unblocking of paths



  - Path from V to Y is blocked by $\emptyset$.
  - Path from V to Y is unblocked by conditioning on W, Y, or $\{W, Y\}$.

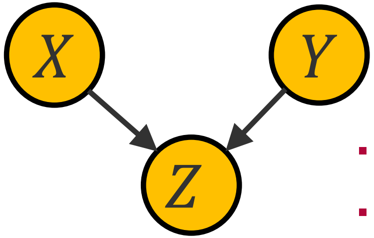
5. Global Markov Condition
Blocking of Paths (II/II)

- **Example (Berkson's Paradox 1946):** Unblocking by conditioning on common effects

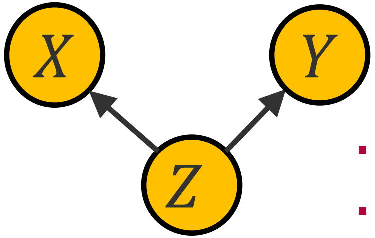


- The path from X to Y is unblocked by conditioning on Z, i.e.,
  - $X \perp\!\!\!\perp Y$
  - but: $X \not\perp\!\!\!\perp Y \mid Z$
- *E.g., the false observation of a negative correlation between two unrelated – or even positive correlated - traits*.

- **Asymmetry under Inverting Arrows (Reichenbach 1956):**



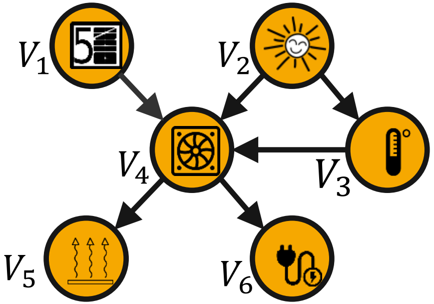
- $X \perp\!\!\!\perp Y$
- $X \not\perp\!\!\!\perp Y \mid Z$



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

5. Global Markov Condition
D-Separation

- **Example (Cooling House Scenario):**



- The path from V_1 to V_6 is blocked by V_4.
- V_1 and V_6 are d-separated by V_4.

- The path $V_2 \rightarrow V_3 \rightarrow V_4 \rightarrow V_6$ is blocked by V_3, V_4, or $\{V_3, V_4\}$.
- But: V_2 and V_6 are d-separated only by V_4, or $\{V_3, V_4\}$.

- The paths $V_1 \rightarrow V_4 \leftarrow V_2$ is blocked by $\emptyset$
- …but unblocked by conditioning on V_4 or $\{V_3, V_4\}$. .
- Note: V_1 and V_2 are d-separated by $\emptyset$ or V_3.

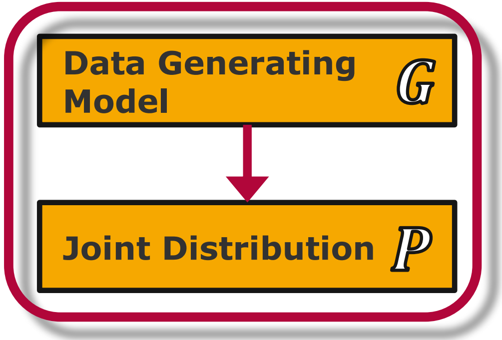
- V_4 is a fork in $V_5 \leftarrow V_4 \rightarrow V_6$.
- V_5 and V_6 are d-separated by V_4.

5. Global Markov Condition
Theorem

Global Markov Condition:
For all disjoint subsets of vertices X, Y and Z we have that

$$X, Y \text{ d-separated by } Z \Rightarrow (X \perp\!\!\!\perp Y \mid Z)_P \,.$$

- I.e., we have $(X \perp\!\!\!\perp Y | \, Z)_G \Rightarrow (X \perp\!\!\!\perp Y | \, Z)_P$



6. Functional Model and Markov Conditions
Theorem (Lauritzen 1996, Pearl 2000)

Theorem:

The following are equivalent:

- Existence of a *functional causal model* G;
- *(Local) Markov condition*: statistical independence of nondescendants given parents (i.e.: every information exchange with its nondescendants involves its parents)
- *Global Markov condition*: d-separation (characterizes the set of independences implied by local Markov condition)
- *Factorization:* $p(v_1, \dots, v_N) = \prod_{i=1}^{N} p(v_i | Pa(v_i))$.

(subject to technical conditions)

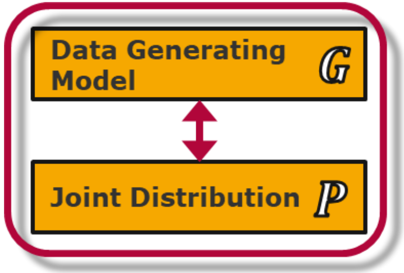
I.e., $(X \perp\!\!\!\perp Y | Z)_G \Rightarrow (X \perp\!\!\!\perp Y | Z)_P$

7. Causal Faithfulness
The Key-Postulate

Causal Faithfulness:
p is called faithful relative to G if only those independencies hold true that are implied by the Markov condition, i.e.,

$$(X \perp Y | Z)_G \Leftarrow (X \perp Y | Z)_P$$

- I.e., we assume that any population P produced by this causal graph G has the independence relations obtained by applying d-separation to it
- Seems like a hefty assumption, but it really isn't:
  It assumes that whatever independencies occur in it arise not from incredible coincidence but rather from structure, i.e., data generating model G.
- Hence:



8. Outlook Causal Structure Learning
Concept (Spirtes, Glymor, Scheines and Pearl)

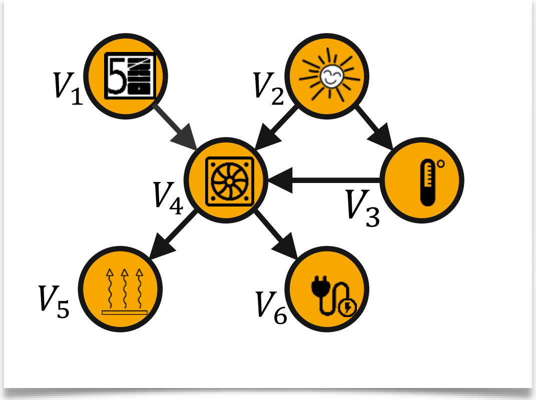
- **Assumptions:**
  - Causal Sufficiency
  - Global Markov Condition
  - Causal Faithfulness

- **Causal Structure Learning:**
  - Accept only those DAG’s G as causal hypothesis for which
    $$(X \perp Y \mid Z)_G \Leftrightarrow (X \perp Y \mid Z)_P.$$
  - Defines the basis of constraint-based causal structure learning
  - Identifies causal DAG up to Markov equivalence class (DAGs that imply the same conditional independencies)

9. Markov Equivalence Class
Theorem (Verma and Pearl)

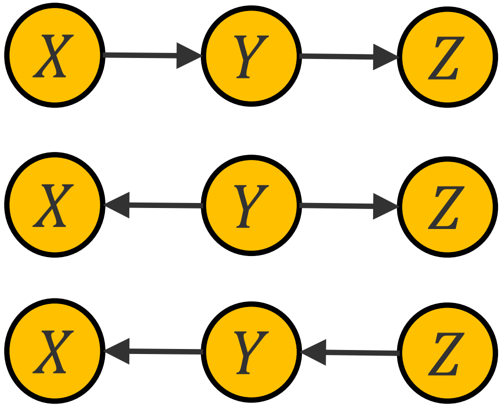
Theorem:
Two DAGs are Markov equivalent if and only if they have the same skeleton and the same v-structures

- *Skeleton:*
  corresponding undirected graph
- *V-Structure:*
  substructure $X \rightarrow Y \leftarrow Z$ with no edges between X and Z.



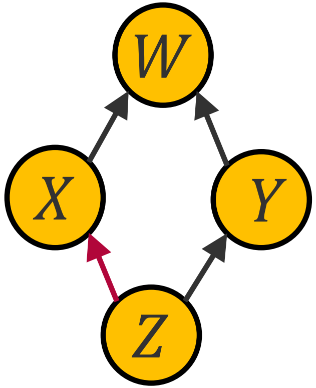
9. Markov Equivalence Class
Examples

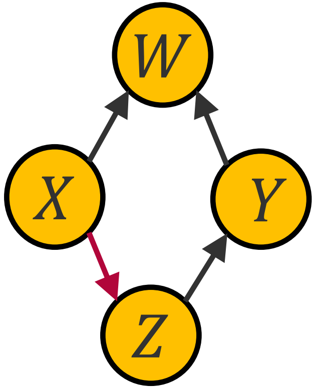
- Same skeleton, no v-structure



$$X \perp Z \mid Y$$

- Same skeleton, same v-structure at W





10. Summary
Causal Graphical Models

- *Causal Graphical Models* formalized by DAG (directed acyclic graph) G with random variables V_i, $i = 1, \dots, N$, as vertices.

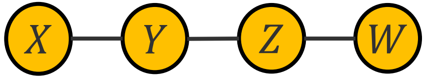
- *Causal Sufficiency*, *Causal Faithfulness* and *(Local) Markov Condition* imply
$$(X \perp Y \mid Z)_G \Leftrightarrow (X \perp Y \mid Z)_P.$$

- *(Local) Markov Condition* states that the density $p(\mathrm{v}_1, \dots, v_N)$ then factorizes into
$$p(v_1, \dots, v_N) = \prod_{i=1}^{N} p(v_i | Pa(v_i)).$$

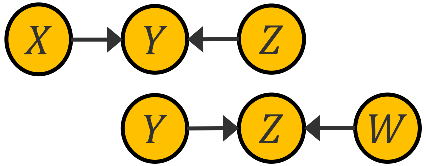
- Causal conditional $p(v_i | Pa(v_i))$ represent *causal mechanisms*.

11. Excursion: Maximal Ancestral Graphs
Motivating Example

- Suppose, we are given the following list of dependency properties among X, Y, Z and W:
  - $X \perp\!\!\!\perp Z$
  - $X \not\perp\!\!\!\perp Y$
  - $Y \perp\!\!\!\perp W$
  - $Y \not\perp\!\!\!\perp Z$
  - $X \perp\!\!\!\perp W$
  - $Z \not\perp\!\!\!\perp W$
- Which DAG could have generated these, and only these, pattern of dependencies?
- The skeleton representing the pattern of dependencies must be:



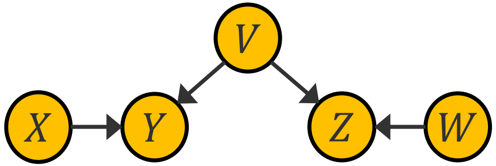
- And there must be the following colliders:



- There is no orientation of $Y - Z$ that is consistent with the independencies.

11. Excursion: Maximal Ancestral Graphs
DAG Models and Marginalization

- Let's include an additional variable V:

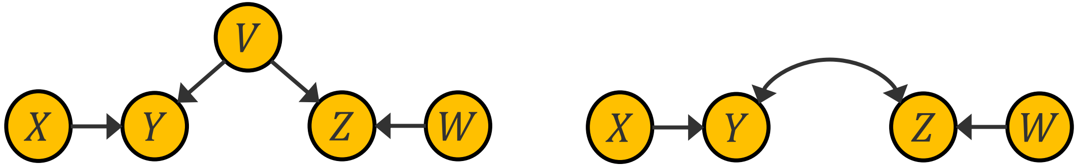


- This DAG model generates a probability distribution $P_{\{V,W,X,Y,Z\}}$ in which:
  - $X \perp\!\!\!\perp Z$
  - $X \not\perp\!\!\!\perp Y$
  - $\mathrm{Y} \perp\!\!\!\perp W$
  - $\mathrm{Y} \not\perp\!\!\!\perp Z$
  - $X \perp\!\!\!\perp W$
  - $\mathrm{Z} \not\perp\!\!\!\perp W$
- The marginal distribution $P_{\{W,X,Y,Z\}} = P_{\{V,WX,Y,Z\}} dv$ must adhere the same dependencies.
- But: this marginal distribution cannot be faithfully generated by any DAG.

➡ **DAG models are not closed under marginalization!**

- **Ancestral Graph (AG)**

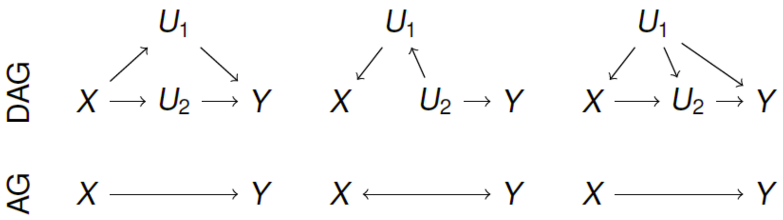
  is a graph containing both directed and bi-directed edges, where the bi-directed edges stand for *latent variables, e.g.,*



- **m-Separation**

  If S m-separates X and Y in an ancestral graph M, then $X \perp\!\!\!\perp Y \mid S$ in every density p that factorizes according to any DAG G that is represented by the AG M.

- **Example**



11. Excursion: Maximal Ancestral Graphs
DAGs vs. AGs

- **Advantages of AGs**
  - AGs can faithfully represent more probability distributions than DAGs.
  - AG models are closed under marginalization.
  - AGs can (implicitly) represent unobserved variables, which exist in many (possibly almost all) applications.

- **Disadvantages of AGs**
  - Parameterization is difficult in the general case.
  - Markov equivalence is difficult.

References

Literature

- Pearl, J. (2009). *Causal inference in statistics: An overview*. Statistics Surveys, 3:96-146.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference.* Cambridge University Press.
- Spirtes, P., Glymour, C., and Scheines, R. (2000). Causation, Prediction, and Search. The MIT Press.

Thank you
for your attention!