Methods for Causal Inference
Lecture 9: D-separation and intro to Pearl's framework

Ava Khamseh

School of Informatics
2024-2025

Various casual quantities of interest

AverageTreatment Effect (ATE) $\mathbb{E}[Y_1 - Y_0]$ is a (common) causal quantity of interest, but it's not the only one …

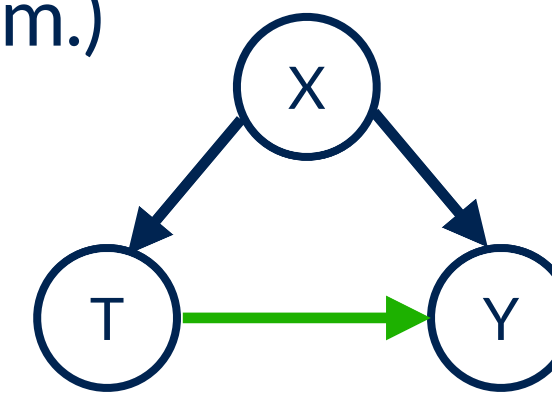
We have talked about Conditional Average Treatment Effect (CATE) $\mathbb{E}[Y_1 - Y_0|X = x]$ which is the average treatment effect for individuals with a certain feature X=x.

Other causal quantities of interest: Causal interaction of two treatments on outcome.

For example: Two drugs for cancer (chemotherapy and radiotherapy)
Is this interaction positive, negative or neutral?

Causal effect of interactions on outcome

Key: function that can be computed for ***any*** **statistical model**. (Function of the distribution without needing to specify its parametric form.)

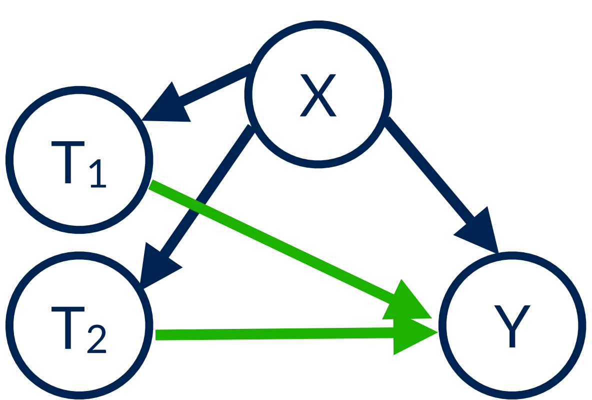


Average Treatment Effect (ATE):

$$\mathrm{ATE}_T(Y) = \mathbb{E}_X\left[\mathbb{E}(Y \mid T=1, X) - \mathbb{E}(Y \mid T=0, X)\right]$$

Interactions between genes i and j leading to outcome Y:

$$I^a_{i,j} = \boxed{\left[\mathbb{E}(Y \mid (T_1, T_2) = (1,1)) - \mathbb{E}(Y \mid (T_1, T_2) = (0,1), X)\right]}$$
$$-\left[\mathbb{E}(Y \mid (T_1, T_2) = (1,0)) - \mathbb{E}(Y \mid (T_1, T_2) = (0,0), X)\right].$$



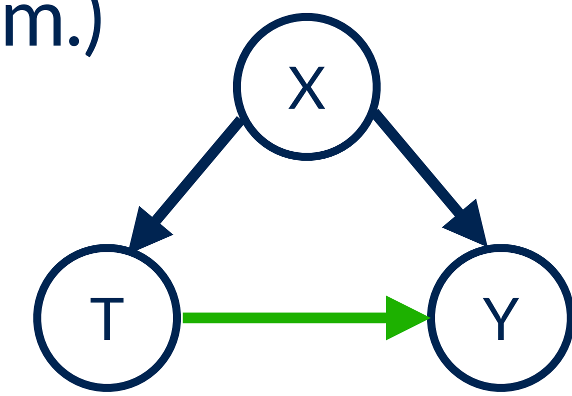
Treatment i given or not

Treatment j <u>given</u>

Beentjes & Khamseh, Physical Review E (2020)

Causal effect of interactions on outcome

Key: function that can be computed for ***any*** **statistical model**. (Function of the distribution without needing to specify its parametric form.)

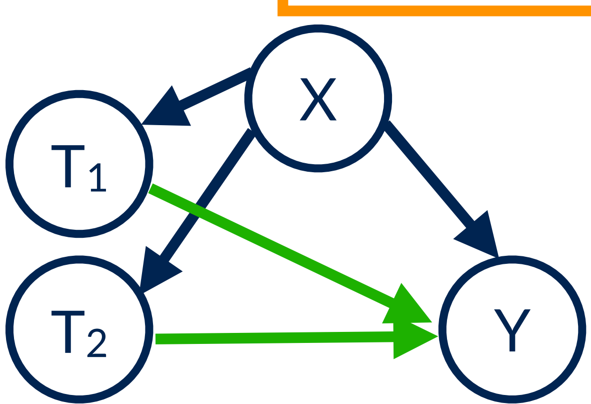


Average Treatment Effect (ATE):

$$\mathrm{ATE}_T(Y) = \mathbb{E}_X \left[\mathbb{E}(Y \mid T = 1, X) - \mathbb{E}(Y \mid T = 0, X)\right]$$

Interactions between genes i and j leading to outcome Y:

$$\begin{aligned} I^a_{i,j} = & \Big[\mathbb{E}(Y \mid (T_1, T_2) = (1,1)) - \mathbb{E}(Y \mid (T_1, T_2) = (0,1), X)\Big] \\ & - \boxed{\Big[\mathbb{E}(Y \mid (T_1, T_2) = (1,0)) - \mathbb{E}(Y \mid (T_1, T_2) = (0,0), X)\Big].} \end{aligned}$$



Treatment i given or not

Treatment j <u>not given</u>

Beentjes & Khamseh, Physical Review E (2020)

Example: Linear regression

Suppose a linear ground truth:

$$Y = \alpha_0 + \alpha_1 T_1 + \alpha_2 T_2 + \gamma T_1 T_2$$

Example: Linear regression

Suppose a linear ground truth:

$$Y = \alpha_0 + \alpha_1 T_1 + \alpha_2 T_2 + \gamma T_1 T_2$$

$$\begin{aligned}
\mathbb{E}(Y \mid T_1 = 1, T_2 = 1) &= \alpha_0 + \alpha_1 + \alpha_2 + \gamma \\
\mathbb{E}(Y \mid T_1 = 1, T_2 = 0) &= \alpha_0 + \alpha_1 \\
\mathbb{E}(Y \mid T_1 = 0, T_2 = 1) &= \alpha_0 + \alpha_2 \\
\mathbb{E}(Y \mid T_1 = 0, T_2 = 0) &= \alpha_0
\end{aligned}$$

$$\begin{aligned}
\mathrm{ATE}_{T_1}(Y \mid T_2 = 1) &= \alpha_1 + \gamma, \qquad & \mathrm{ATE}_{T_2}(Y \mid T_1 = 1) &= \alpha_2 + \gamma, \\
\mathrm{ATE}_{T_1}(Y \mid T_2 = 0) &= \alpha_1 & \mathrm{ATE}_{T_2}(Y \mid T_1 = 0) &= \alpha_2
\end{aligned}$$

$$\boxed{I^a_{1,2} = \gamma = I^a_{2,1}}$$

Beyond effect sizes [non-examinable]

Importantly, it allows us to **target** very precise questions:

1. Effect on health if all people **were** treated, i.e.,

$$\mathbb{E}[Y_1 - Y] \stackrel{!}{=} \mathbb{E}_X\left[\mathbb{E}[Y|T=1, X]\right] - \mathbb{E}[Y] > 0?$$

Beyond effect sizes [non-examinable]

Importantly, it allows us to **target** very precise questions:

1. Effect on health if all people **were** treated, i.e.,

$$\mathbb{E}[Y_1 - Y] \stackrel{!}{=} \mathbb{E}_X\left[\mathbb{E}[Y|T=1, X]\right] - \mathbb{E}[Y] > 0?$$

2. Effect on health if people **were** treated based on confounders,

$$\mathbb{E}[Y_{d(X)} - Y] \stackrel{!}{=} \mathbb{E}_X\left[\mathbb{E}[Y|T=d(X), X]\right] - \mathbb{E}[Y] > 0?$$

Beyond effect sizes [non-examinable]

Importantly, it allows us to **target** very precise questions:

1. Effect on health if all people **were** treated, i.e.,

$$\mathbb{E}[Y_1 - Y] \stackrel{!}{=} \mathbb{E}_X\left[\mathbb{E}[Y|T=1, X]\right] - \mathbb{E}[Y] > 0?$$

2. Effect on health if people **were** treated based on confounders,

$$\mathbb{E}[Y_{d(X)} - Y] \stackrel{!}{=} \mathbb{E}_X\left[\mathbb{E}[Y|T=d(X), X]\right] - \mathbb{E}[Y] > 0?$$

3. Flip it around: What is the ***optimal*** **treatment rule**, i.e.

$$d_{\mathrm{opt}}(X) = \arg\max_{d(X)}\left\{\mathbb{E}_X\left[\mathbb{E}[Y|T=d(X), X]\right] - \mathbb{E}[Y]\right\}?$$

Slide from Sjoerd Beentjes

Time to event causal estimates [non-examinable]

What is the data structure? Example: $O = (Y, T, X, \tilde{\tau}, \Delta) \sim P_0$

Here we have

Y = outcome (i.e. event occurs or not (1 or 0))

T = treatment

X = confounders / covariates

Tau = time of event (if it occurs)

C = time of censoring (if patient drops out before event)

Only one of tau or C is observed, namely $\tilde{\tau} = \min(\tau, C)$

= censoring occurs or not (1 or 0), i.e., is $\tau \leq C?$

Causal question:

"Is survival time greater under treatment or not?", i.e

$$\mathbb{P}(\tau_1 > \tau^*) - \mathbb{P}(\tau_0 > \tau^*) > 0?$$

Time to event causal estimates [non-examinable]

- **Model** a causal inference problem with assumptions manifest in Causal Graphical Models [**Pearl]**

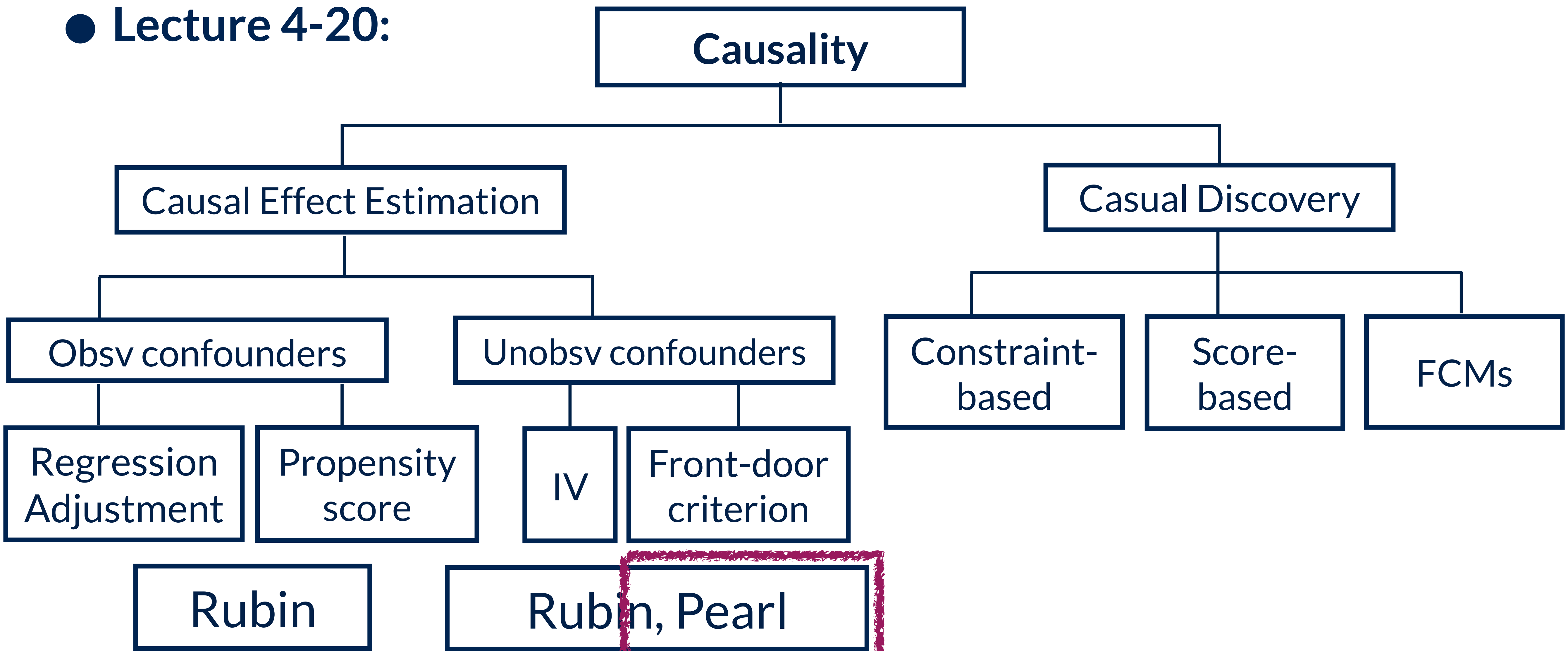
- **Identify** an expression for the causal effect under these assumptions ("causal estimand"), [**Pearl**]

- **Estimate** the expression using statistical methods such as matching or instrumental variables,

- **Verify** the validity of the estimate using a variety of robustness checks.

Overview of the course

- **Lecture 1**: Introduction & Motivation, why do we care about causality? Why deriving causality from observational data is non-trivial.
- **Lecture 2:** Recap of probability theory, variables, events, conditional probabilities, independence, law of total probability, Bayes' rule
- **Lecture 3**: Recap of regression, multiple regression, graphs, SCM
- **Lecture 4-20:**



Pearl's Model of Causality

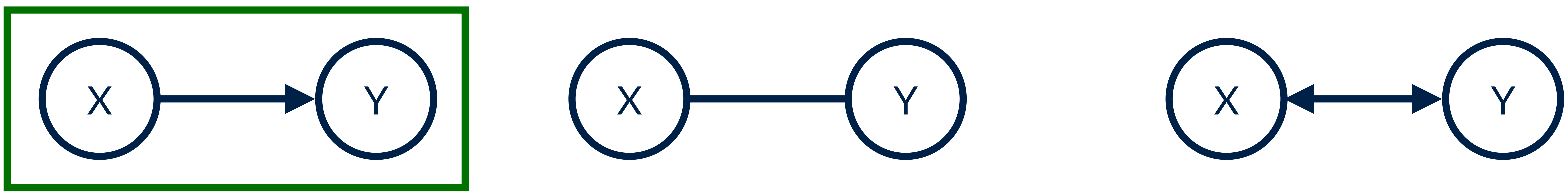
- Ladder of causation:
  - **Association:** What does a symptom tell me about a disease?
  - **Intervention (perturbation):** If I take aspirin will my headache be cured?
  - **Counterfactual:** Was it the aspirin that stopped the headache? (alternative versions of past events, strongest causal statements e.g. **physical laws**)

- Aim: To **model** and **identify** the causal estimand

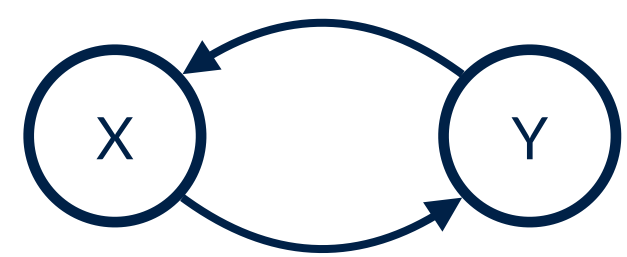
- Causal graphical models + structural equations

Causal Graphical Models

- Diagrammatic representation of probability distributions + **causal info**
- **Graph:** Consists of a set of **vertices** V (nodes), **edges** E
- V are the variables and E contains information between the variables
- Graphs can be directed, undirected and bidirectional (confounder?)

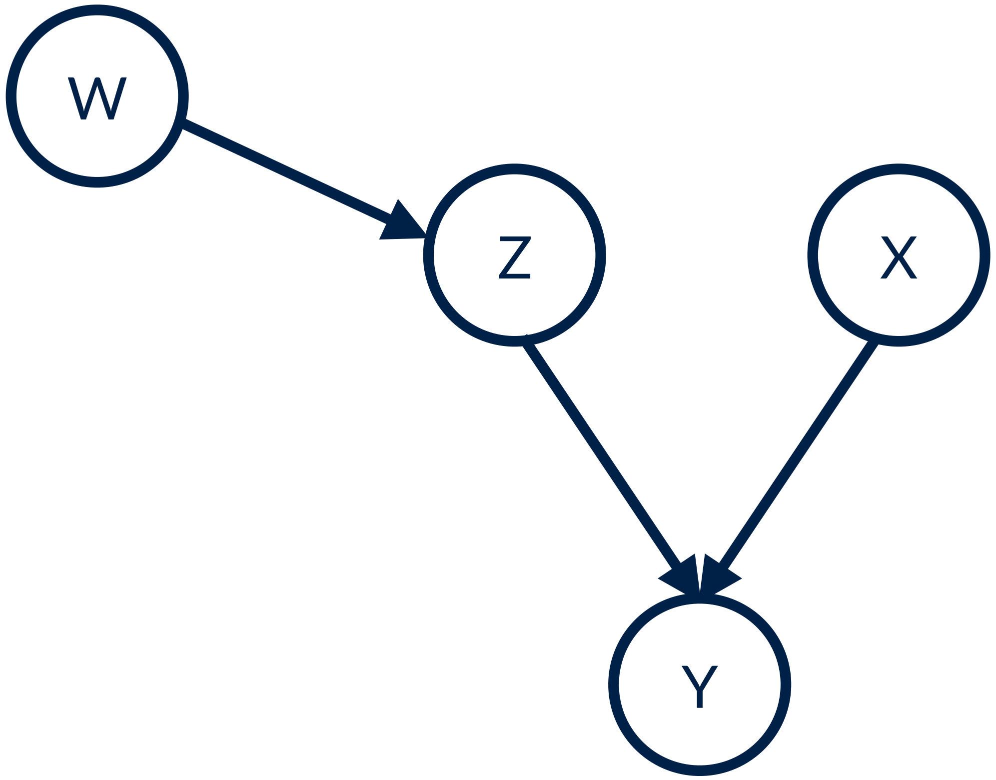


- Directed graphs may include directed cycles, i.e., mutual causation/feedback process.
- A graph with no directed cycles is an **acyclic** graph.



Causality, Pearl (2009)

Directed Acyclic Graphs (DAGs)



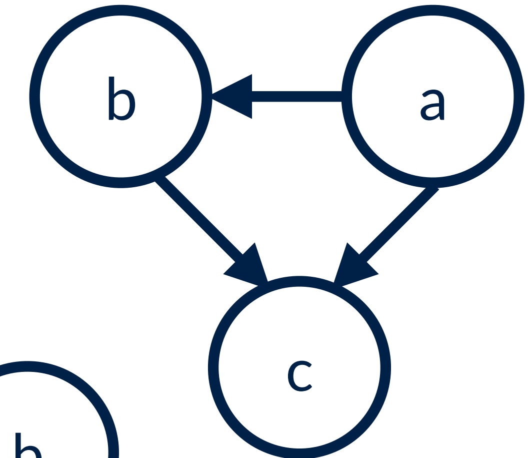
Z, X are parents of Y
Z, X, W are ancestors of Y
Y has no children
X has no parents

- DAG in which every node has at most one parent is a **tree**
- A tree in which every node has at most one child is a **chain**
- **DAG:**
  - Expresses **model assumptions** explicitly
  - Represents **joint probability** functions
  - Provides **efficient inference** of observations

Causality, Pearl (2009)

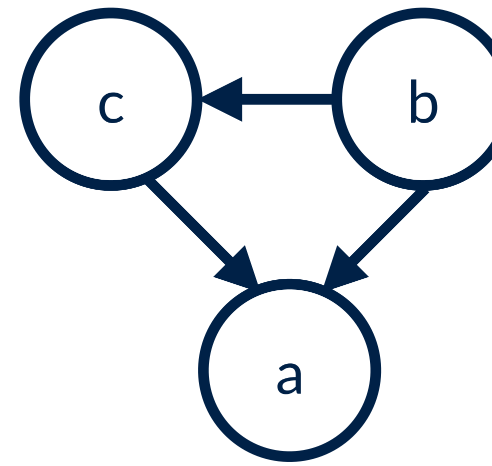
DAG contains more info than joint probability

$$p(a,b,c) = p(c|a,b)p(a,b) = p(c|a,b)p(b|a)p(a)$$



$$p(a,b,c) = p(a|b,c)p(b,c) = p(a|b,c)p(c|b)p(b)$$

Symmetric
in a, b, c



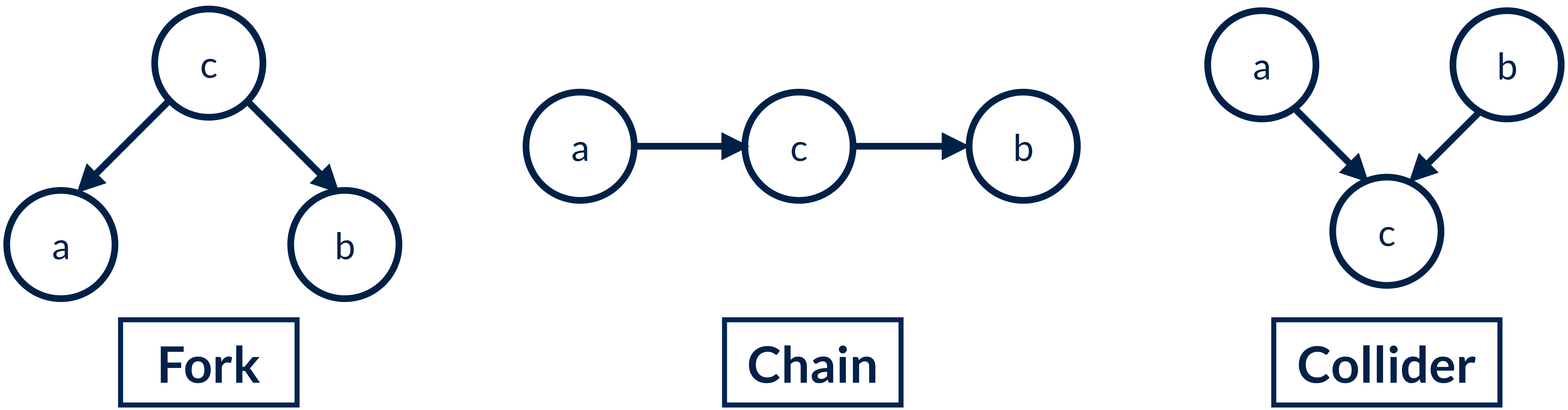
NOT
symmetric

- Probabilistic notations are <u>**not enough**</u> to describe causal aspects
- Using repeated application of product/Bayes' rule, one can write any joint probability distribution in terms of its marginals and conditionals
- A graph is **fully connected** if there is a link between every pair of nodes
- The interest lies in the **absence** of a link and link **direction**.

Bishop, Pattern Recognition and Machine Learning, Chapter 8

Basic DAG structures:

- Conditional independence via graphs and **D-separation**
- 3 main graph structures:



- Next Lecture: **Do-calculus** and **causal identification**

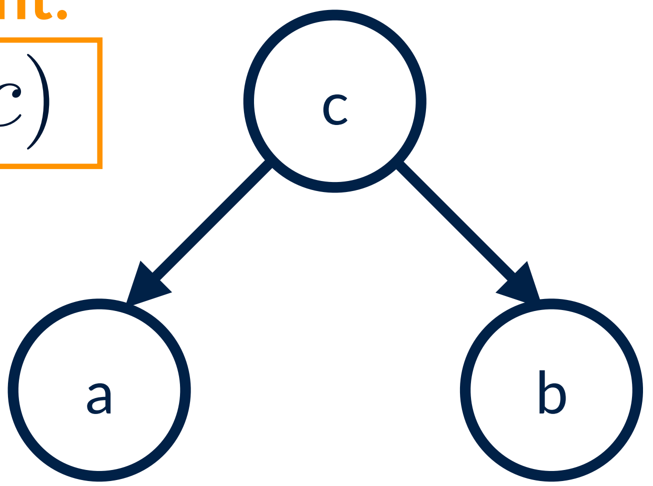
Causality, Pearl (2009)

Fork

$$p(a,b,c) = p(a|c)p(b|c)p(c)$$

In contrast to the full joint:

$$\boxed{p(a|b,c)p(b|c)p(c)}$$



Fork

Case 1: No conditioning

$$p(a,b) = \sum_c p(a,b,c) = \sum_c p(a|c)p(b|c)p(c) \neq p(a)p(b) \text{ in general}$$

$$\Rightarrow a \not\perp\!\!\!\perp b|\emptyset$$

Case 2: Conditioning on c

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a|c)p(b|c)p(c)}{p(c)} = p(a|c)p(b|c)$$

$$\Rightarrow a \perp\!\!\!\perp b|c$$

c **blocks (d-separates)** the path from a to b

Chain

$$p(a,b,c) = p(a)p(c|a)p(b|c)$$

$$\begin{tikzcd} a \arrow[r] & c \arrow[r] & b \end{tikzcd}$$

Chain

Case 1: No conditioning

$$p(a,b) = \sum_c p(a)p(c|a)p(b|c) = p(a)\sum_c p(b|c)p(c|a) = p(a)p(b|a) \neq p(a)p(b)$$

$$\Rightarrow a \not\perp\!\!\!\perp b|\emptyset$$

Using:

$$\sum_c p(b|c)p(c|a) = \sum_c p(b|c,a)p(c|a) = \sum_c p(b,c|a) = p(b|a)$$

Case 2: Conditioning on c

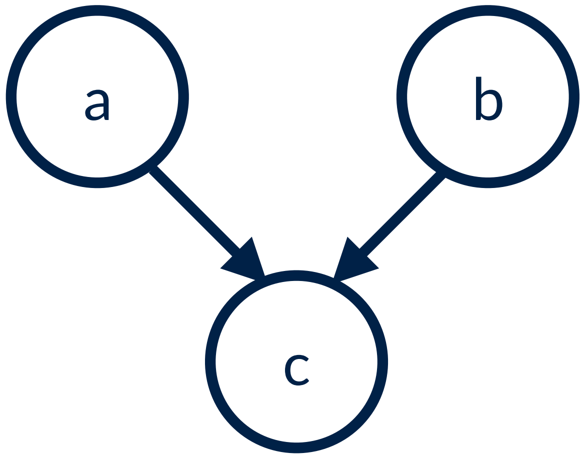
$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(c|a)p(b|c)}{p(c)} = \frac{p(a)p(b|c)}{p(c)}\frac{p(a|c)p(c)}{p(a)} = p(a|c)p(b|c)$$

$$\Rightarrow a \perp\!\!\!\perp b|c$$

c **blocks (d-separates)** the path from a to b

Collider

$$p(a,b,c) = p(a)p(b)p(c|a,b)$$



Case 1: No conditioning

$$p(a,b) = \sum_c p(a)p(b)p(c|a,b) = p(a)p(b)\sum_c p(c|a,b) = p(a)p(b)$$

Collider

$\Rightarrow a \perp\!\!\!\perp b|\emptyset$ with no conditioning, a and b are independent

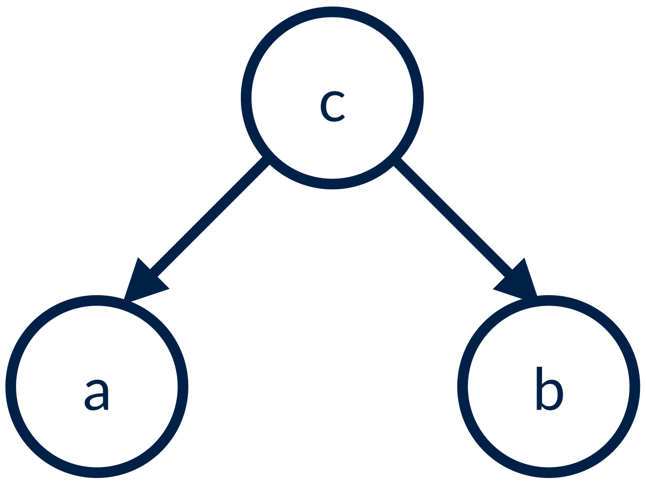
Case 2: Conditioning on c

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(b)p(c|a,b)}{p(c)} \neq p(a|c)p(b|c) \text{ in general}$$

$\Rightarrow a \not\perp\!\!\!\perp b|c$ c **unblocks** the path from a to b

Summary

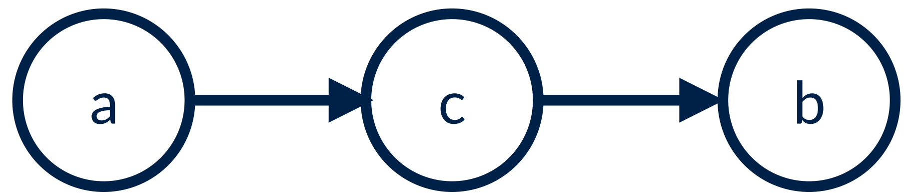
- Conditional independence via graphs and **D-separation**
- 3 main graph structures:



Fork

$$a \not\perp\!\!\!\perp b | \emptyset$$

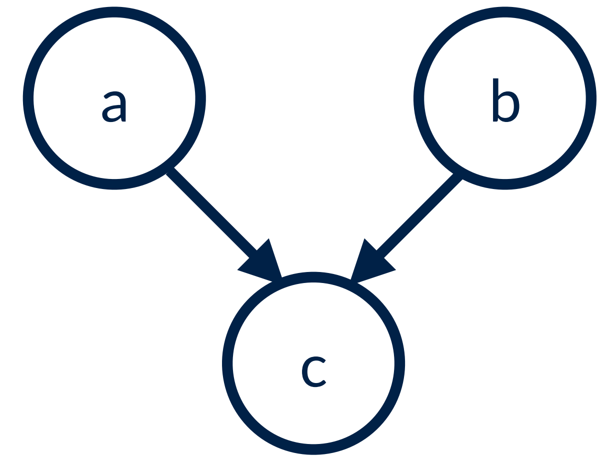
$$a \perp\!\!\!\perp b | c$$



Chain

$$a \not\perp\!\!\!\perp b | \emptyset$$

$$a \perp\!\!\!\perp b | c$$



Collider

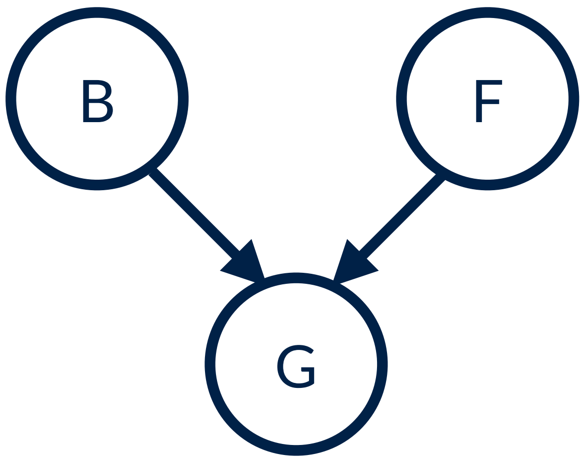
$$a \perp\!\!\!\perp b | \emptyset$$

$$a \not\perp\!\!\!\perp b | c$$

Pearl, Causality (2009)

Collider example

B: State of battery, B=1 charged, B=0 flat
F: State of fuel tank, F=1 full, F=0 empty
G: State of electric fuel gauge, G=1 full, G=0 empty



Collider

Given Info:

$$p(B = 1) = 0.9$$

$$p(F = 1) = 0.9$$

$$p(G = 1 | B = 1, F = 1) = 0.8$$

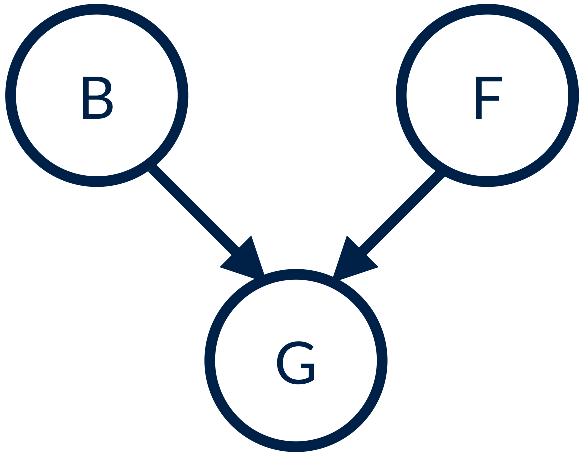
$$p(G = 1 | B = 1, F = 0) = 0.2$$

$$p(G = 1 | B = 0, F = 1) = 0.2$$

$$p(G = 1 | B = 0, F = 0) = 0.1$$

 Bishop, Pattern Recognition and Machine Learning, Chapter 8

Collider example

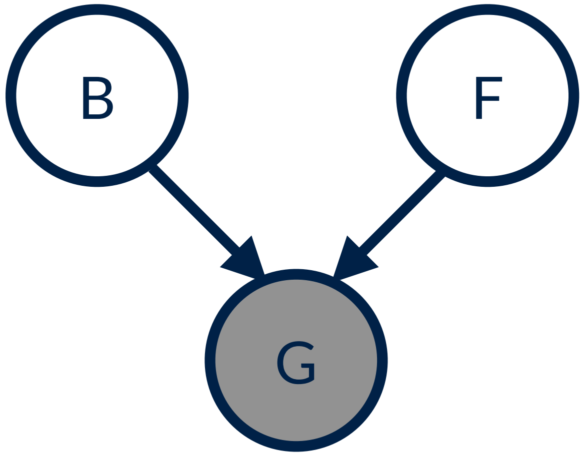


B: State of battery, B=1 charged, B=0 flat
F: State of fuel tank, F=1 full, F=0 empty
G: State of electric fuel gauge, G=1 full, G=0 empty

1 Before any conditioning (before observing):

$$p(F = 0) = 0.1$$

Collider example



B: State of battery, B=1 charged, B=0 flat
F: State of fuel tank, F=1 full, F=0 empty
G: State of electric fuel gauge, G=1 full, G=0 empty

1 Before any conditioning (before observing):

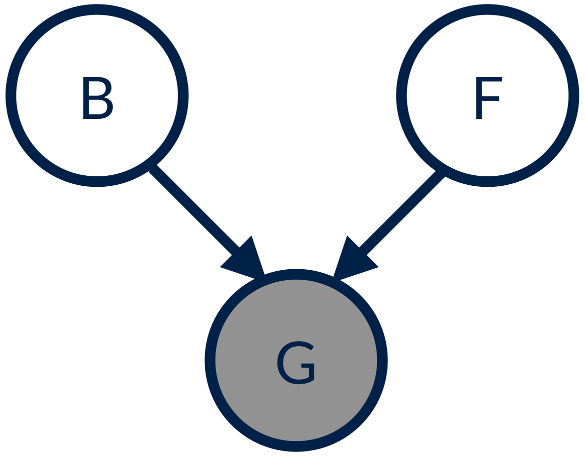
$$p(F=0)=0.1$$

2 Now suppose we observe G=0

$$p(F=0|G=0)=\frac{p(G=0|F=0)p(F=0)}{p(G=0)}$$

 Bishop, Pattern Recognition and Machine Learning, Chapter 8

Collider example



B: State of battery, B=1 charged, B=0 flat
F: State of fuel tank, F=1 full, F=0 empty
G: State of electric fuel gauge, G=1 full, G=0 empty

1 Before any conditioning (before observing):

$$p(F=0) = 0.1$$

2 Now suppose we observe G=0

$$p(F=0|G=0) = \frac{p(G=0|F=0)p(F=0)}{p(G=0)}$$

$$\sum_{B \in \{0,1\}} p(G=0|F=0,B)p(B) = 0.81$$

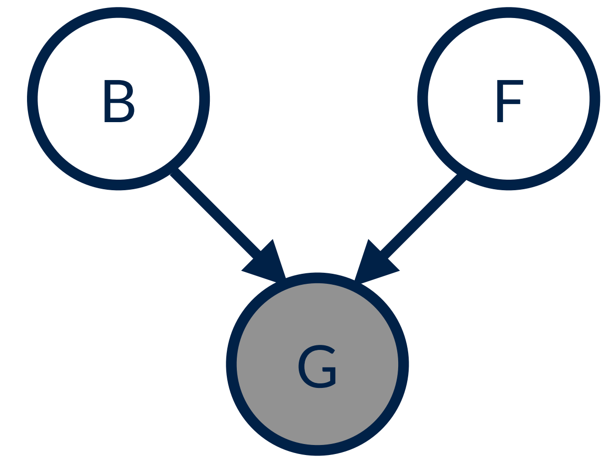
$$\sum_{B,F \in \{0,1\}} p(G=0,B,F)$$

$$= \sum_{B,F \in \{0,1\}} p(G=0|B,F)p(B|F)p(F)$$

$$= \sum_{B,F \in \{0,1\}} p(G=0|B,F)p(B)p(F) = 0.315$$

Since B and F are independent

Collider example



B: State of battery, B=1 charged, B=0 flat
F: State of fuel tank, F=1 full, F=0 empty
G: State of electric fuel gauge, G=1 full, G=0 empty

1 Before any conditioning (before observing):

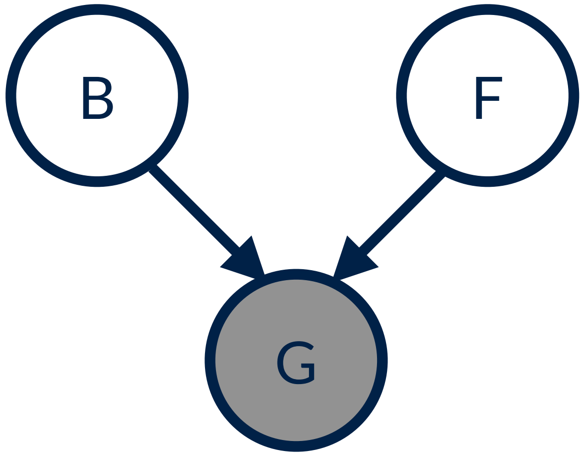
$$p(F=0)=0.1$$

2 $$p(F=0|G=0)=0.257$$

$$p(F=0)<p(F=0|G=0)$$

Observing that gauge reads empty makes it more likely that the tank is indeed empty.

Collider example



B: State of battery, B=1 charged, B=0 flat
F: State of fuel tank, F=1 full, F=0 empty
G: State of electric fuel gauge, G=1 full, G=0 empty

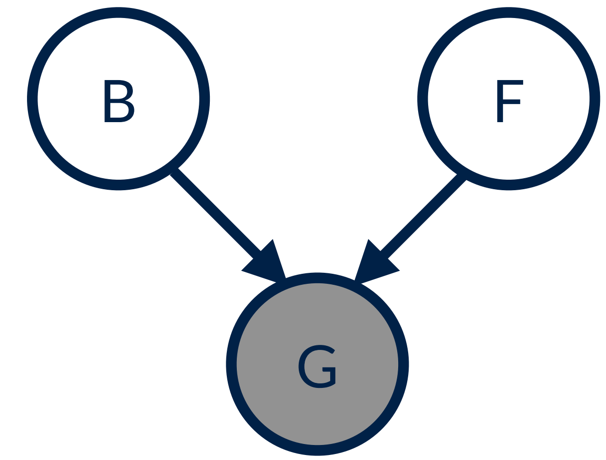
2 $p(F=0|G=0)=0.257$

3 Now we **also** observe B=0

$$p(F=0|G=0,B=0)=\frac{p(F=0,G=0,B=0)}{p(G=0,B=0)}$$

$$=\frac{p(G=0|B=0,F=0)p(F=0)p(B=0|F=0)}{\sum_{F\in\{0,1\}}p(G=0|B=0,F)p(F)p(B=0|F)}=0.111$$

Collider example



B: State of battery, B=1 charged, B=0 flat
F: State of fuel tank, F=1 full, F=0 empty
G: State of electric fuel gauge, G=1 full, G=0 empty

2 $p(F=0|G=0) = 0.257$

3 Now we **also** observe B=0

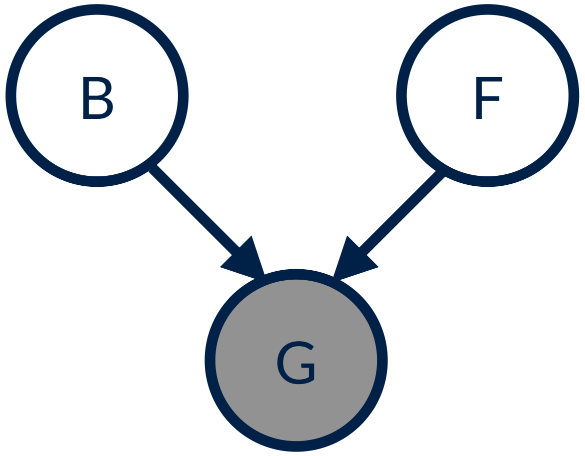
$$p(F=0|G=0,B=0) = \frac{p(F=0,G=0,B=0)}{p(G=0,B=0)}$$

$$= \frac{p(G=0|B=0,F=0)p(F=0)p(B=0|F=0)}{\sum_{F\in\{0,1\}} p(G=0|B=0,F)p(F)p(B=0|F)} = 0.111$$

$$p(F=0|G=0) > p(F=0|G=0,B=0)$$

Probability that tank is empty F=0 has decreased with extra information on the state of the battery

Collider example



B: State of battery, B=1 charged, B=0 flat
F: State of fuel tank, F=1 full, F=0 empty
G: State of electric fuel gauge, G=1 full, G=0 empty

1 $p(F=0)=0.1$

2 $p(F=0|G=0)=0.257$

3 $p(F=0|G=0,B=0)=0.111$

Conditioning on G, then finding out the battery is flat, 'explains away' the observation that the fuel gauge reads empty. The state of the fuel tank and the battery have become dependent:

$$p(F=0|G=0) \neq p(F=0|G=0,B=0)$$

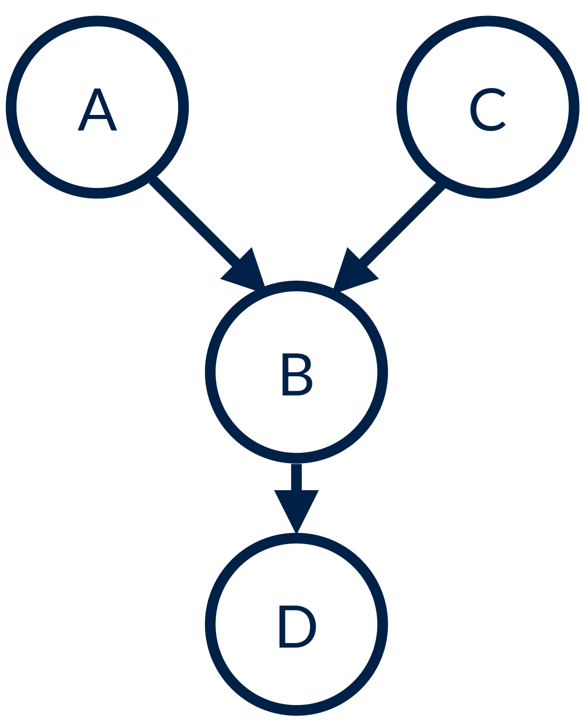
(Even though: $p(F)=p(F|B)$)

D-separation

A path p is **blocked** by a set of nodes Z if and only if:

1) p contains a **chain** of nodes A -> B -> C or a **fork** A <- B -> C such that the middle node B is in Z (i.e. B is conditioned on), or

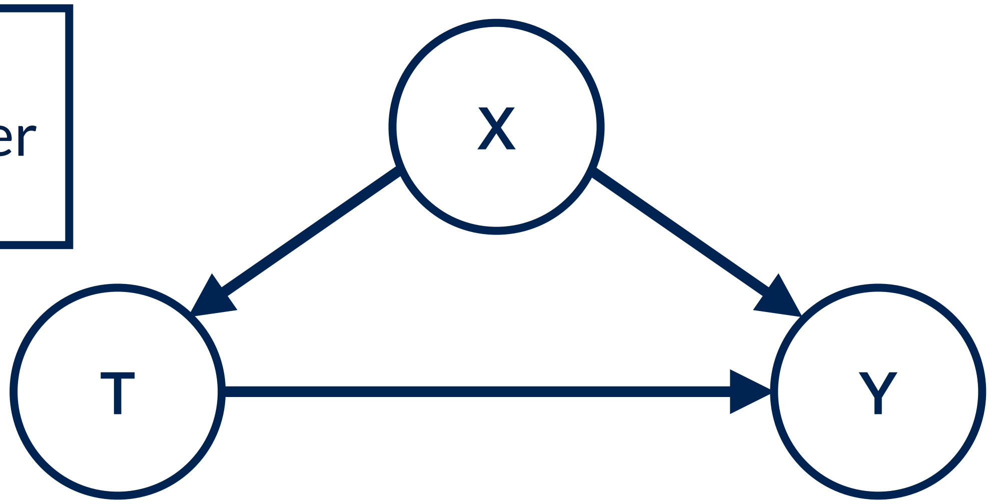
2) p contains a **collider** A -> B <- C such that the collision node B is not in Z, and no descendant of B is in Z.

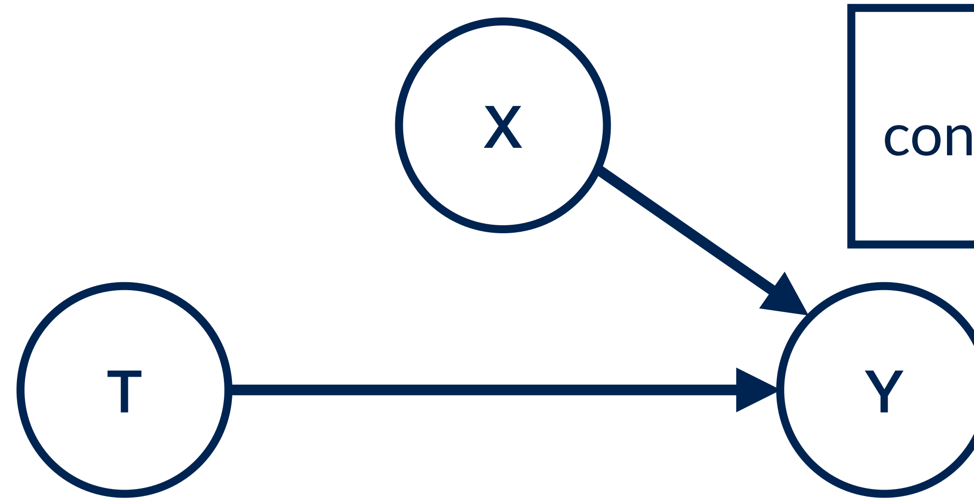


Causality, Pearl (2009)

Confounder vs not a confounder

Classical confounder case





No confounder case

$$\begin{aligned}
\mathbb{E}_X\left[\mathbb{E}_Y\left[Y|X,T\right]\right] &= \int dx\ p(x) \int dy\ y\ p(y|x,t) \\
&= \int dx\ p(x) \int dy\ y\ \frac{p(y,x|t)}{p(x|t)} \\
&= \int dx\ p(x) \int dy\ y\ \frac{p(y,x|t)}{p(x)} \\
&= \int dy\ y\ p(y|t) = \mathbb{E}_Y\left[Y|T\right],
\end{aligned}$$

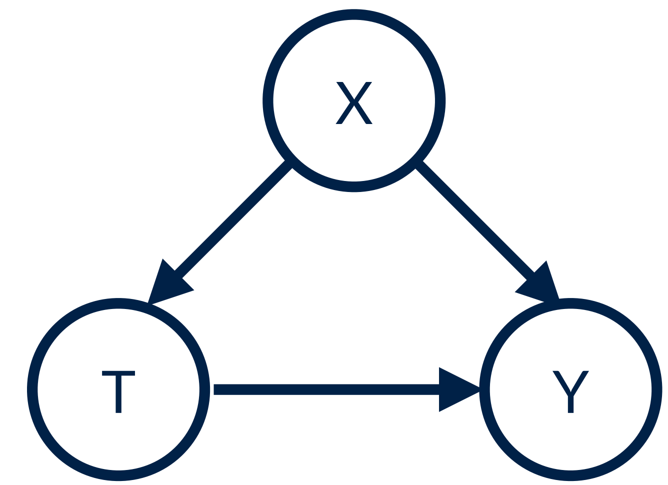
Independence of X and W on the RHS graph

Pearl’s framework

Graphical models & Do-calculus

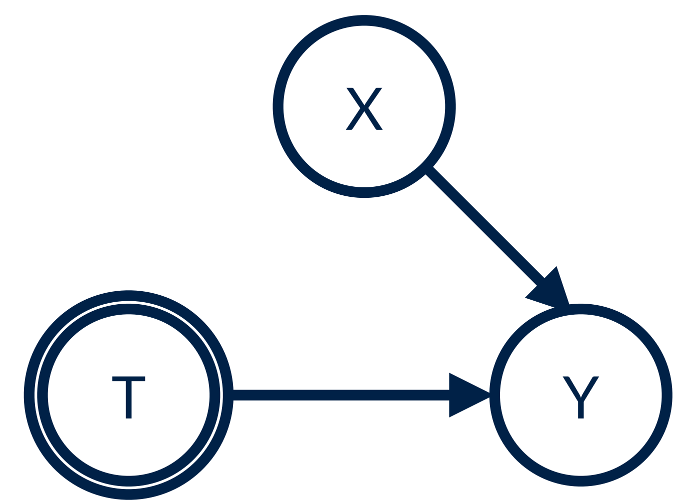
Observation (conditioning) vs intervention

Distinguish between: a variable T takes a value t naturally and cases where we **fix** T=t by denoting the latter do(T=t)



$$p(Y = y|T = t)$$

Probability that Y=y **conditional** on finding T=t
i.e., population distribution of Y among individuals whose T value is t (subset)



$$p(Y = y|do(T = t))$$

Probability that Y=y when we **intervene** to make T=t
i.e., population distribution of Y if **everyone in the population** had their T value fixed at t.

Graph surgery

Pearl, Causal Inference in Statistics (2016)

Structural Causal Models (SCM)

An SCM consists of d structural assignments

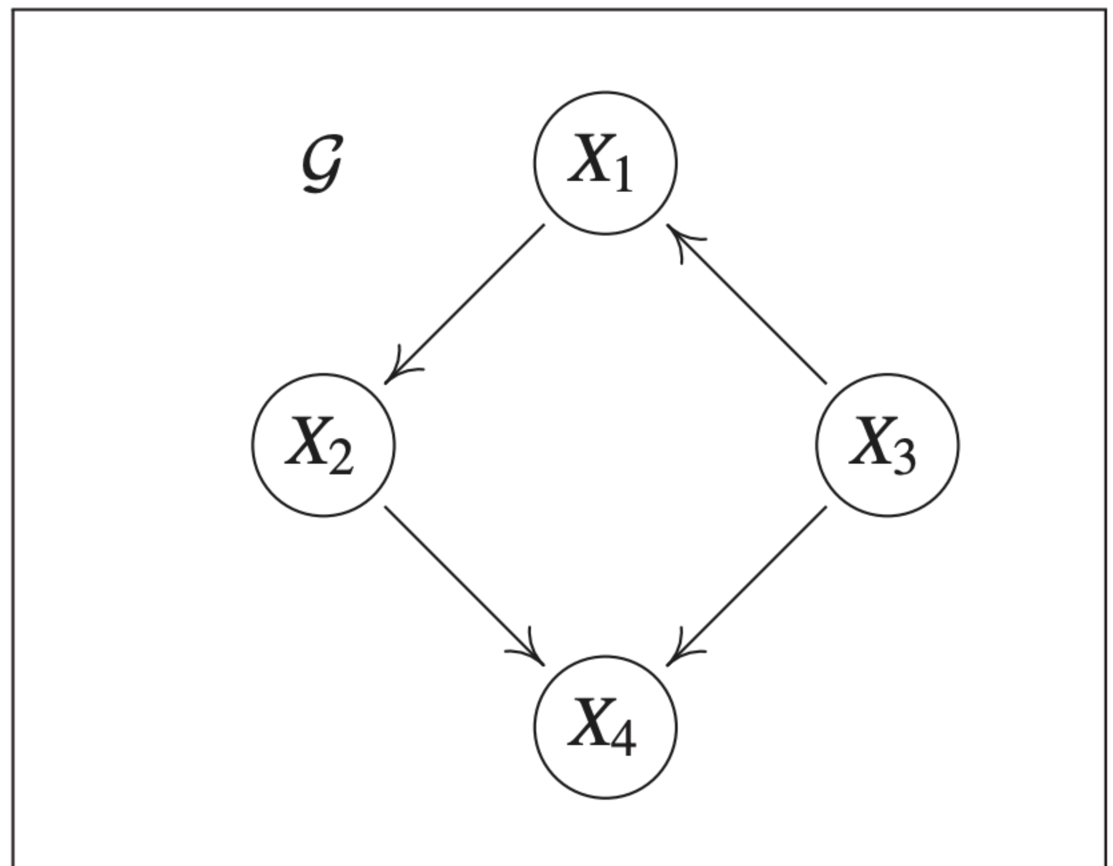
$$X_j := f_j(PA_j, N_j) \quad , \quad j = 1, \cdots, d$$

Parents of X_j, i.e., direct causes of X_j

Jointly independent noise variables

$$X_1 := f_1(X_3, N_1)$$
$$X_2 := f_2(X_1, N_2)$$
$$X_3 := f_3(N_3)$$
$$X_4 := f_4(X_2, X_3, N_4)$$

- $N_1, \ldots, N_4$ jointly independent
- $\mathcal{G}$ is acyclic



Jonas Peters et al, Elements of Causal Inference (2017)

THE UNIVERSITY of EDINBURGH

Methods for Causal Inference
Lecture 9: D-separation and intro to Pearl's framework

Ava Khamseh

School of Informatics

2024-2025