Single World Intervention Graphs

Thomas Richardson

TUM Short Course Lecture III

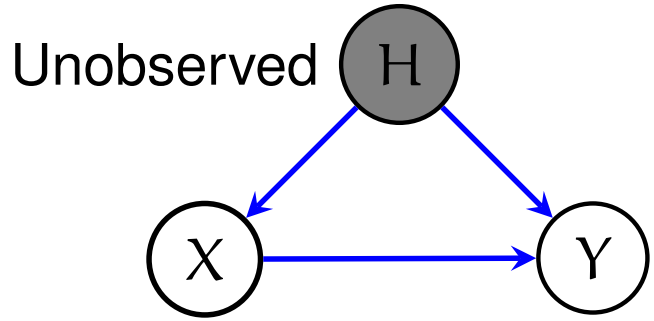
Outline

- Relating graphs and counterfactuals via node-splitting
- Simple examples
- General procedure
- Local Property
- Adjustment for Confounding
- Potential Outcomes (po) Calculus
- Sequentially Randomized Experiments / Time Dependent Confounding

Joint work with James M. Robins (Harvard) and Ilya Shpitser (JHU)

Graphical Approach to Causality

No Confounding



Confounding

- Graph intended to represent direct causal relations.
- Convention that confounding variables (e.g. H) are always included on the graph.
- Approach originates in the path diagrams introduced by Sewall Wright in the 1920s.
- If $X \rightarrow Y$ then X is said to be a *parent* of Y; Y is *child* of X.

Graphical Approach to Causality

No Confounding

- Associated factorization:

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

- In the absence of confounding the *causal* model asserts:

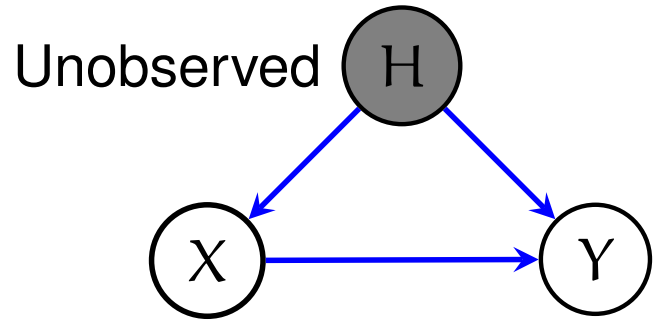
$$P(Y(x) = y) = P(Y = y \mid \mathrm{do}(X = x)) = P(Y = y \mid X = x).$$

  Thus $\mathrm{ACE}(X \to Y)$ is identified under this model.

- Q: *How does this relate to the non-graphical approach?*

Linking the two approaches

$X \perp\!\!\!\perp Y(x_0)$ & $X \perp\!\!\!\perp Y(x_1)$

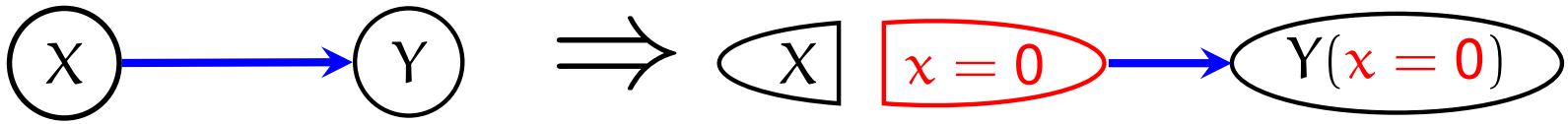


$X \not\perp\!\!\!\perp Y(x_0)$ *or* $X \not\perp\!\!\!\perp Y(x_1)$

- Elephant in the room:
  The variables $Y(x_0)$ *and* $Y(x_1)$ *do not appear on these graphs!!*

Node splitting: Setting X to 0

$$P(X=\tilde{x}, Y=\tilde{y}) = P(X=\tilde{x})P(Y=\tilde{y} \mid X=\tilde{x})$$



Can now 'read' the independence: $X \perp\!\!\!\perp Y(x=0)$.
Also associate a new factorization:

$$P\left(X=\tilde{x}, Y(x=0)=\tilde{y}\right) = P(X=\tilde{x})P\left(Y(x=0)=\tilde{y}\right)$$

where:

$$P\left(Y(x=0)=\tilde{y}\right) = P(Y=\tilde{y} \mid X=0).$$

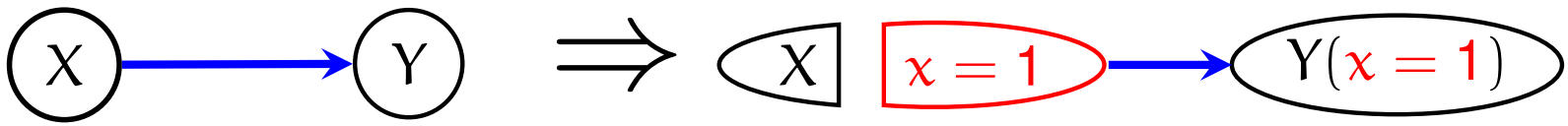
This last equation links a term in the original factorization to the new factorization. We term this the 'modularity assumption'.

From counterfactual perspective modularity follows from factorization + consistency:

$$P\left(Y(x=0)=\tilde{y}\right) = P\left(Y(x=0)=\tilde{y} \mid X=0\right) = P(Y=\tilde{y} \mid X=0)$$

Node splitting: Setting X to 1

$$P(X=\tilde{x}, Y=\tilde{y}) = P(X=\tilde{x})P(Y=\tilde{y} \mid X=\tilde{x})$$



Can now 'read' the independence: $X \perp\!\!\!\perp Y(x=1)$.
Also associate a new factorization:

$$P\left(X=\tilde{x}, Y(x=1)=\tilde{y}\right) = P(X=\tilde{x})P\left(Y(x=1)=\tilde{y}\right)$$

where:

$$P\left(Y(x=1)=y\right) = P(Y=y \mid X=1).$$

Marginals represented by SWIGs are identified

The SWIG $\mathcal{G}(x_0)$ represents $P(X, Y(x_0))$.
The SWIG $\mathcal{G}(x_1)$ represents $P(X, Y(x_1))$.
Under no confounding these marginals are identified from $P(X, Y)$.
In contrast the distribution $P(X, Y(x_0), Y(x_1))$ is not identified.
$Y(x=0)$ and $Y(x=1)$ are never on the same SWIG.
Although we have:

$$X \perp\!\!\!\perp Y(x=0) \quad \text{and} \quad X \perp\!\!\!\perp Y(x=1)$$

we do not assume

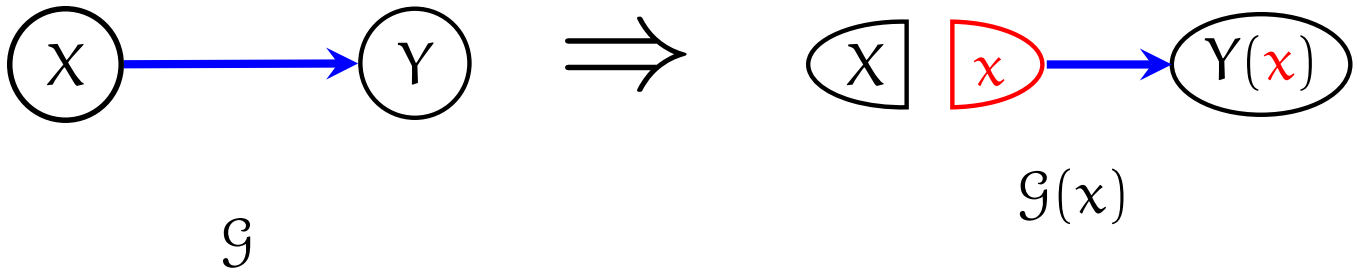
$$X \perp\!\!\!\perp Y(x=0), Y(x=1)$$

Had we tried to construct a single graph containing both $Y(x=0)$ and $Y(x=1)$ this would have been impossible.

$\Rightarrow$ *Single-World Intervention Graphs* (SWIGs).

Representing both graphs via a 'template'

$$P(X=\tilde{x}, Y=\tilde{y}) = P(X=\tilde{x})P(Y=\tilde{y} \mid X=\tilde{x})$$



$\mathcal{G}$ $\mathcal{G}(x)$

Represent both graphs via a *template*:

Formally the template is a 'graph valued function' (not a graph!):

- Takes as input a specific value x^*
- Returns as output a SWIG $\mathcal{G}(x^*)$.

Each *instantiation* of the template represents a different margin:
SWIG $\mathcal{G}(x_0)$ represents $P(X, Y(x_0))$;
SWIG $\mathcal{G}(x_1)$ represents $P(X, Y(x_1))$.

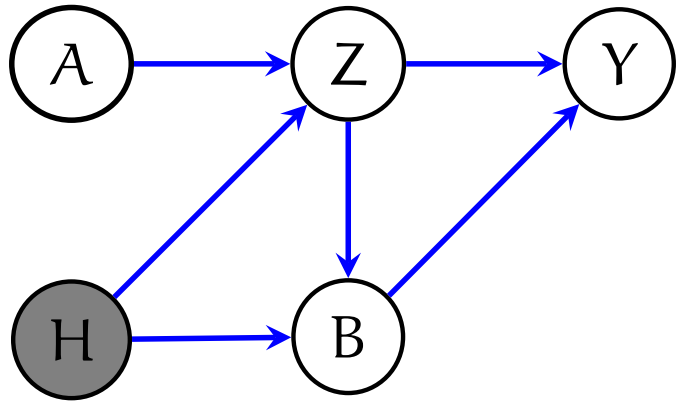
Intuition behind node splitting:

(Robins, VanderWeele, Richardson 2007)

Q: *How could we identify whether someone would choose to take treatment, i.e. have* $X = 1$*, and at the same time find out what happens to such a person if they don't take treatment* $Y(x = 0)$*?*

A: Consider an experiment in which, whenever a patient is observed to swallow the drug have $X = 1$, we instantly intervene by administering a safe 'emetic' that causes the pill to be regurgitated before any drug can enter the bloodstream. Since we assume the emetic has no side effects, the patient's recorded outcome is then $Y(x = 0)$.

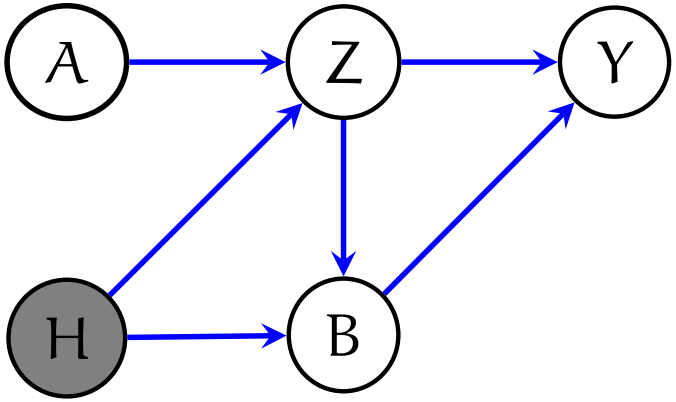
Harder Inferential problem

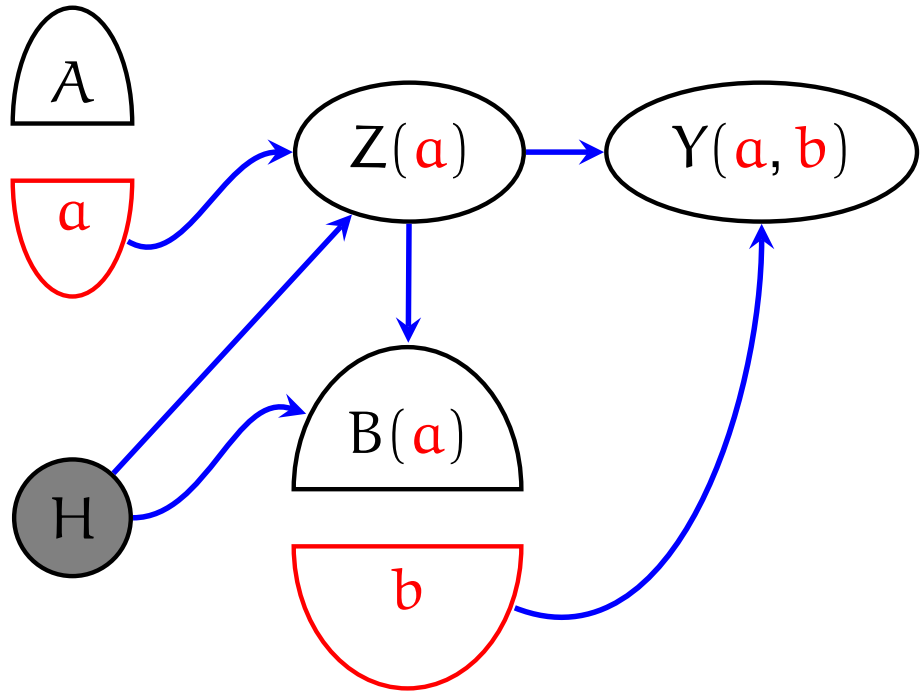


Query: does this causal graph imply:

$$Y(a,b) \perp\!\!\!\perp B(a) \mid Z(a), A \qquad ?$$

Simple solution





Query does this graph imply:

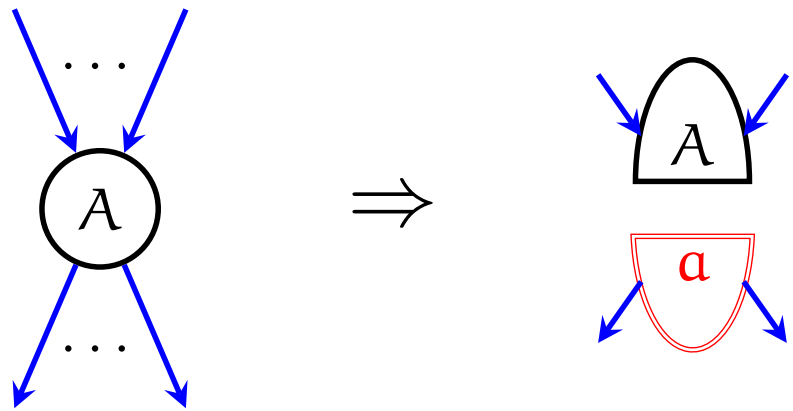
$$Y(a, b) \perp\!\!\!\perp B(a) \mid Z(a), A \qquad ?$$

Answer: Yes – applying d-separation to the SWIG on the right we see that there is no d-connecting path from $Y(a, b)$ given $Z(a)$.

Single World Intervention Template Construction (1)

Given a graph G, a subset of vertices $\mathbf{A} = \{A_1, \ldots, A_k\}$ to be intervened on, we form $\mathrm{G}(\mathbf{a})$ in two steps:

(1) (Node splitting): For every $A \in \mathbf{A}$ split the node into a *random* node A and a *fixed* node a:

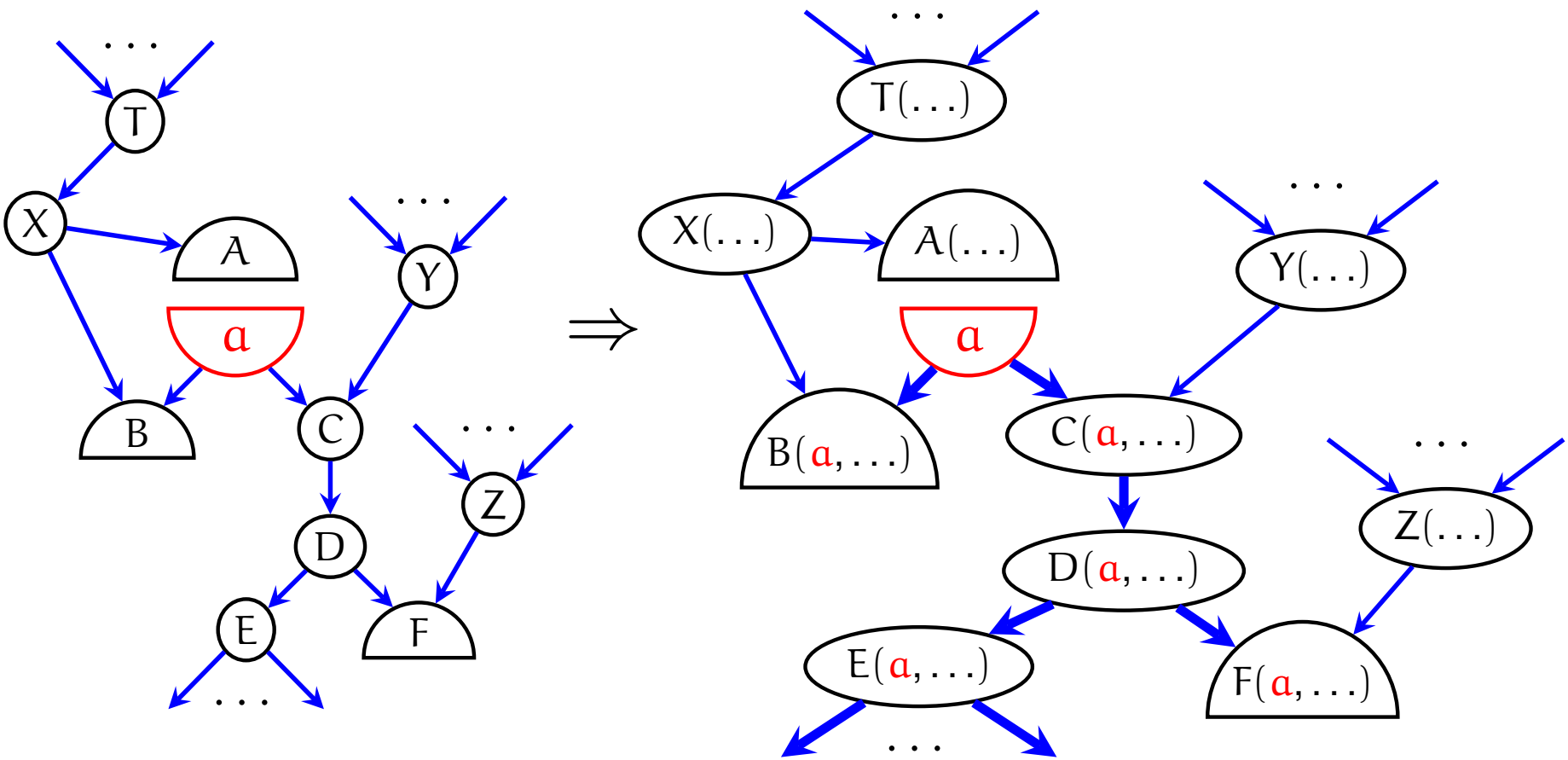


Splitting: Schematic Illustrating the Splitting of Node A

- The random half inherits all edges directed into A in $\mathcal{G}$;
- The fixed half inherits all edges directed out of A in $\mathcal{G}$.

Single World Intervention Template Construction (2)

(2) Relabel descendants of fixed nodes:



Relating Observed and Potential Outcome Distributions

Original graph $\mathcal{G}$: observed distribution $\mathrm{P}(\mathbf{V})$
SWIG $\mathcal{G}(\tilde{\mathbf{a}})$: counterfactual distribution $\mathrm{P}(\mathbf{V}(\tilde{\mathbf{a}}))$

Note that under minimal labeling variables in $\mathbf{V}(\tilde{\mathbf{a}})$ may be not labelled with the full set $\tilde{\mathbf{a}}$.

Factorization of counterfactual variables: Distribution $\mathrm{P}(\mathbf{V}(\tilde{\mathbf{a}}))$ over the variables in $\mathcal{G}(\tilde{\mathbf{a}})$ factorizes with respect to the SWIG $\mathcal{G}(\tilde{\mathbf{a}})$ (ignoring fixed nodes):

Modularity: $\mathrm{P}(\mathbf{V}(\tilde{\mathbf{a}}))$ and $\mathrm{P}(\mathbf{V})$ are linked as follows:
The conditional density associated with $Y(\tilde{\mathbf{a}})$ in $\mathcal{G}(\tilde{\mathbf{a}})$ is just the conditional density associated with Y in $\mathcal{G}$ after substituting $\tilde{a}_i$ for any $A_i \in \mathbf{A}$ that is a parent of Y.

Consequence: if $\mathrm{P}(\mathbf{V})$ is observed then $\mathrm{P}(\mathbf{V}(\tilde{\mathbf{a}}))$ is identified.

Applying d-separation to the graph $\mathrm{G}(\mathbf{a})$ (Part 1)

We extend the definition of d-connection to SWIGs as follows:

- A red (fixed) node is always blocked if it occurs as a non-endpoint on a path;
- A path on which one endpoint is a red (fixed) node can d-connect that node to a random node if it satisfies the usual conditions on colliders and non-colliders;

In $\mathcal{G}(\tilde{\mathbf{a}})$ if subsets $\mathbf{B}(\tilde{\mathbf{a}})$ and $\mathbf{C}(\tilde{\mathbf{a}})$ of random nodes are d-separated by $\mathbf{D}(\tilde{\mathbf{a}})$, then $\mathbf{B}(\tilde{\mathbf{a}})$ and $\mathbf{C}(\tilde{\mathbf{a}})$ are conditionally independent given $\mathbf{D}(\tilde{\mathbf{a}})$ in the associated distribution $\mathrm{P}(\mathbf{V}(\tilde{\mathbf{a}}))$.

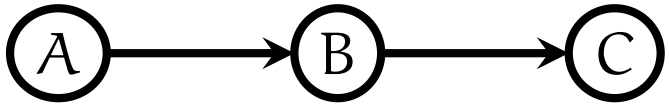
$$\mathbf{B}(\tilde{\mathbf{a}}) \text{ is d-separated from } \mathbf{C}(\tilde{\mathbf{a}}) \text{ given } \mathbf{D}(\tilde{\mathbf{a}}) \text{ in } \mathcal{G}(\tilde{\mathbf{a}}) \qquad (1)$$

$$\Rightarrow \quad \mathbf{B}(\tilde{\mathbf{a}}) \perp\!\!\!\perp \mathbf{C}(\tilde{\mathbf{a}}) \mid \mathbf{D}(\tilde{\mathbf{a}}) \quad [\mathbf{P}(\mathbf{V}(\tilde{\mathbf{a}}))].$$

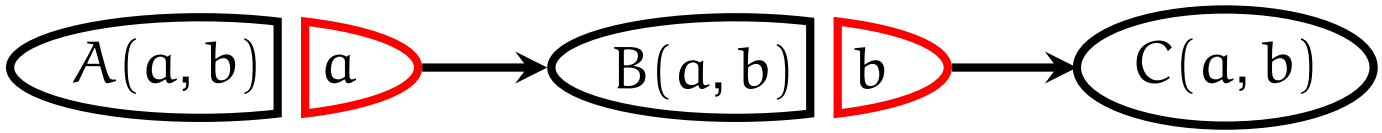
Labelling Schemes

Sometimes it is useful to use different labeling schemes:

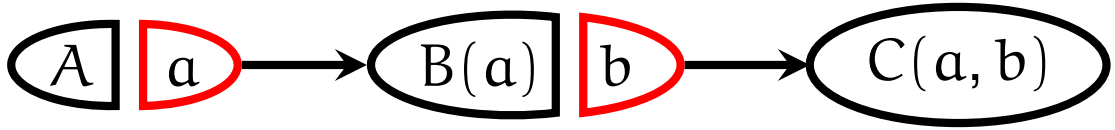
DAG



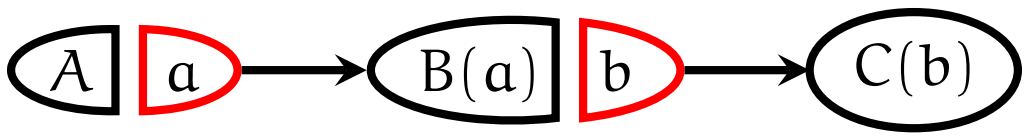
Uniform



Temporal



Ancestral



- Uniform: no equalities between potential outcomes, also when considering several SWIGs
- Temporal: time order; missing edges correspond to no direct effect at population level
- Ancestral: time order; missing edges correspond to no direct effect at individual level

SWIG Local Markov Property

Let

$$\mathcal{P}_A \equiv \{p(V(\mathbf{a})) \mid \mathbf{a} \in \mathfrak{X}_A\}, \qquad \mathcal{P}_A^{\subseteq} \equiv \bigcup_{D \subseteq A} \mathcal{P}_D \tag{2}$$

Definition

A set of potential outcome distributions $\mathcal{P}_A$ obeys *the SWIG ordered local Markov property for DAG* $\mathcal{G}$ *under* $\prec$ if for all $i \in V$, $\mathbf{a} \in \mathfrak{X}_A$, and $w \in \mathfrak{X}_{\text{pre}_\prec(i)}$,

$$p(X_i(\mathbf{a}) \mid X_{\text{pre}_\prec(i)}(\mathbf{a}) = w) \tag{3}$$

is a function only of $\mathbf{a}_{\text{pa}_\mathcal{G}(i) \cap A}$ and $w_{\text{pa}_\mathcal{G}(i) \setminus A}$.

Thus after intervening on A, the distn. of $X_i(\mathbf{a})$ given its predecessors depends solely on the values of interventions on targets in A that are parents of i, and by any other (random) variables that are parents of i but that are not intervened on (hence not in A)

Consequences

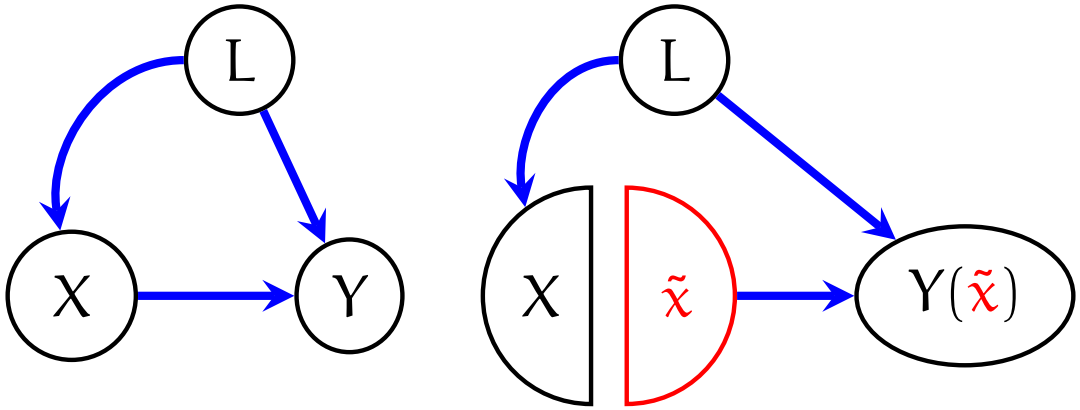
Lemma

If $\mathcal{P}_{\overline{A}}^{\subseteq}$ obeys distributional consistency and $\mathcal{P}_A$ obeys the SWIG ordered local Markov property for DAG $\mathcal{G}$ under $\prec$ then:

$$
\begin{aligned}
& p(X_i(\mathbf{a}) \mid X_{pre(i)}(\mathbf{a})) && (4)\\
&= p(X_i(\mathbf{a}_{pre(i)\cap A}) \mid X_{pre(i)}(\mathbf{a}_{pre(i)\cap A})) && (5)\\
&= p(X_i(\mathbf{a}_{pa(i)\cap A}) \mid X_{pre(i)}(\mathbf{a}_{pa(i)\cap A})) && (6)\\
&= p(X_i(\mathbf{a}_{pa(i)\cap A}) \mid X_{pa(i)}(\mathbf{a}_{pa(i)\cap A})) && (7)\\
&= p(X_i(\mathbf{a}_{pa(i)\cap A}) \mid X_{pa(i)\setminus A}(\mathbf{a}_{pa(i)\cap A})). && (8)
\end{aligned}
$$

Adjustment for Confounding

Adjusting for confounding

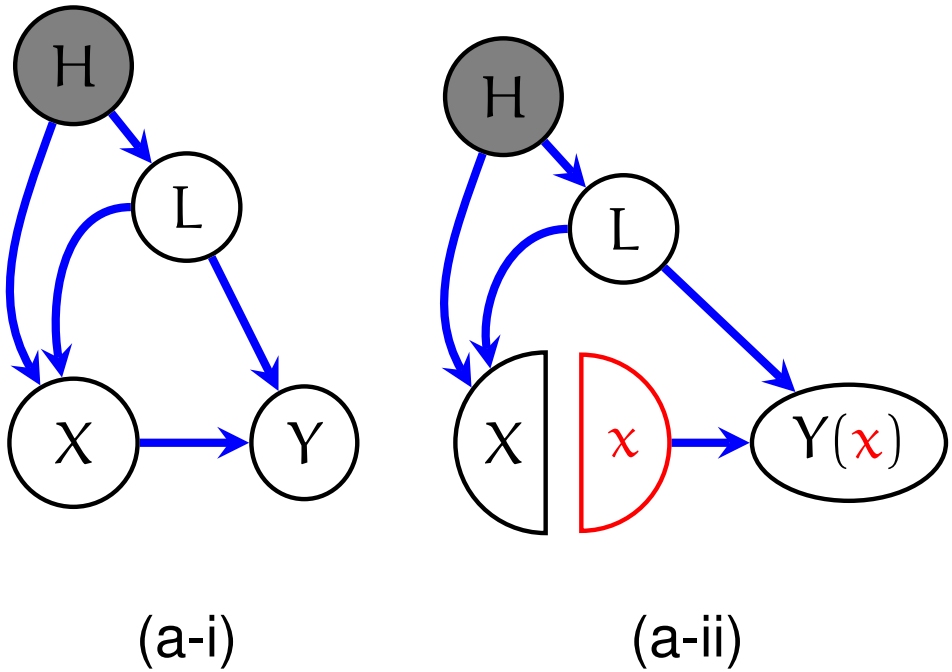


Here we can read directly from the template that

$$X \perp\!\!\!\perp Y(\tilde{x}) \mid L.$$

$$\begin{aligned} P[Y(\tilde{x}) = y] &= \sum_l P[Y(\tilde{x}) = y \mid L = l] P(L = l) \\ &= \sum_l P[Y(\tilde{x}) = y \mid L = l, X = \tilde{x}] P(L = l) \text{ indep} \\ &= \sum_l P[Y = y \mid L = l, X = \tilde{x}] P(L = l) \text{ consistency} \end{aligned}$$

More Examples (I)

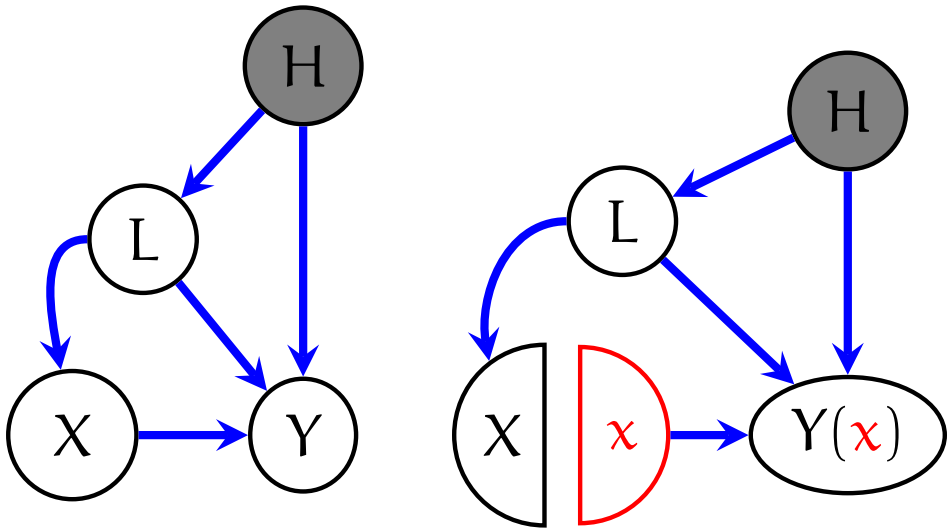


(a-i) (a-ii)

Here we can read directly from the template that

$$X \perp\!\!\!\perp Y(x) \mid L.$$

More Examples (II)



(b-i) (b-ii)

Here we can read directly from the template that

$$X \perp\!\!\!\perp Y(x) \mid L.$$

Connection to Pearl's *do*-calculus

Factorization and modularity are sufficient to imply all of the identification results that hold in the *do*-calculus of Pearl (1995); see also Spirtes *et al.* (1993):

$\mathrm{P}(Y = \mathrm{y} \mid do(\mathbf{A} = \mathbf{a}))$ is identified $\Leftrightarrow$ $\mathrm{P}(Y(\mathbf{a}) = \mathrm{y})$ is identified.

Relating Counterfactuals and 'do' notation

Expressions in terms of 'do' can be expressed in terms of counterfactuals:

$$P(Y(x) = y) \equiv P(Y = y \mid do(X = x))$$

Pearl's 'do' notation is a special case of the g-notation introduced in Robins(1986).
Counterfactual notation is more general than 'do' notation (but not g-notation!).
Ex. Distribution of outcomes that *would* arise among those who took treatment ($X = 1$) had counter-to-fact they not received treatment:

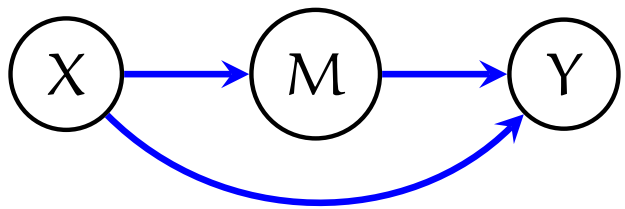
$$P(Y(x = 0) = y \mid X = 1)$$

If treatment is randomized, so $X \perp\!\!\!\perp Y(x = 0)$ then this equals $P(Y(x = 0) = y)$, but in an observational study these may be different.

Relating Counterfactuals and Structural Equations

Potential outcomes can be seen as a different notation for Non-Parametric Structural Equation Models (NPSEMs).

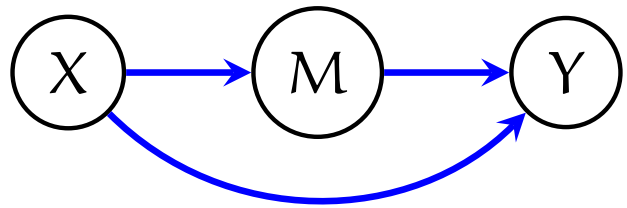
In an NPSEM model associated with a graph each variable is given by an equation expressing the variable as a function of its parents + error term



$$X = f_X(\varepsilon_X)$$
$$M = f_M(X, \varepsilon_M)$$
$$Y = f_Y(X, M, \varepsilon_Y)$$

Relating Counterfactuals and Structural Equations

In an NPSEM model associated with a graph each variable is given by an equation expressing the variable as a function of its parents + error term



But it is clearer to express with potential outcomes

$$\begin{aligned} X &= f_X(\varepsilon_X) \\ M &= f_M(X, \varepsilon_M) \\ Y &= f_Y(X, M, \varepsilon_Y) \end{aligned} \quad \Rightarrow \quad \begin{aligned} X &= f_X(\varepsilon_X) \\ M(x) &= f_M(x, \varepsilon_M) \\ Y(x, m) &= f_Y(x, m\varepsilon_Y) \end{aligned}$$

observed variables are given by: $M = M(X)$, $Y = Y(X, M(X))$.

Counterfactuals make clear equations represent invariant relationships:

intervening to set X *and* M *to* 0*, the value for* Y *will be:* $f_Y(0, 0, \varepsilon_Y)$.

(Alternative approach via crossing out equations, but this can be confusing since "Y" in the new system is not "Y" in the old system.)

Two important caveats:

- NPSEMs typically assume all variables are seen as being subject to well-defined interventions (not so with potential outcomes)
- Pearl's approach to unifying graphs and counterfactuals simply associates with a DAG the counterfactual model corresponding to an NPSEMs with Independent Errors (NPSEM-IEs) with DAGs.
  Pearl: DAGs and Potential Outcomes are 'equivalent theories'.
- However, any counterfactual independences that can be read from a SWIG will hold under the NPSEM-IE model.
  (Though in general the NPSEM-IE will imply extra independences.)

Simplifying the *do*-Calculus

Applying d-separation to the graph $\mathrm{G}(\mathbf{a})$ (Part 2)

(Malinsky, Shpitser, R, 2019; Robins 2018)

We extend the definition of d-connection to SWIGs as follows:

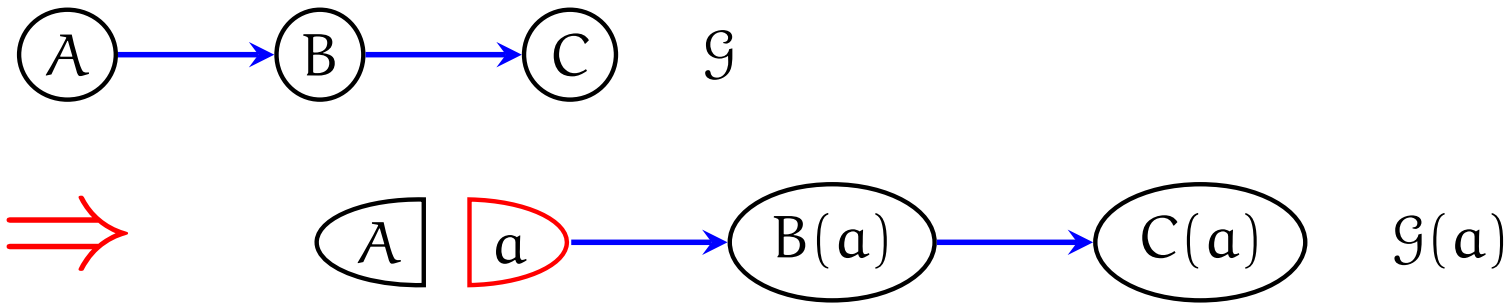
- A red (fixed) node is always blocked if it occurs as a non-endpoint on a path;
- A path on which one endpoint is a red (fixed) node can d-connect that node to a random node if it satisfies the usual conditions on colliders and non-colliders;

In $\mathcal{G}(\tilde{\mathbf{a}}, \mathrm{d})$, if fixed node $\mathbf{d}$ is d-separated from $\mathbf{B}(\tilde{\mathbf{a}}, \mathbf{d})$ given $\mathbf{C}(\tilde{\mathbf{a}}, \mathbf{d})$ then

$$\mathrm{P}(\mathbf{B}(\tilde{\mathbf{a}}, \mathbf{d}) \mid \mathbf{C}(\tilde{\mathbf{a}}, \mathbf{d})) = \mathbf{P}(\mathbf{B}(\tilde{\mathbf{a}}, \mathbf{d}') \mid \mathbf{C}(\tilde{\mathbf{a}}, \mathbf{d}')). \tag{9}$$

In other words, the conditional distribution of B given C after intervening on $\mathbf{A}$ and D does not depend on the value assigned to D.

Example of d-separation from fixed nodes



The fixed node a is d-separated from $C(a)$ given $B(a)$.
Consequently it follows that

$$P(C(\tilde{a}) \mid B(\tilde{a})) = P(C(a^*) \mid B(a^*))$$

for any values $\tilde{a}$, a^*. This may alternatively be derived:

$$\begin{aligned} P(C(\tilde{a}) \mid B(\tilde{a})) &=^{d,\mathcal{G}(\tilde{a})} P(C(\tilde{a}) \mid B(\tilde{a}), A = \tilde{a}) \\ &=^{c} P(C \mid B, A = \tilde{a}) \;=^{d,\mathcal{G}} P(C \mid B, A = a^*) \\ &=^{c} P(C(a^*) \mid B(a^*), A = a^*) \;=^{d,\mathcal{G}(\tilde{a})} P(C(a^*) \mid B(a^*)) \end{aligned}$$

via consistency and d-separation in $\mathcal{G}(\tilde{a})$ and $\mathcal{G}$.

do-calculus

Pearl (1995) formulated a set of rules that give graphical conditions allowing three transformations:

1: Removing observations

$$\begin{aligned} p(y \mid z, w, \mathsf{do}(x)) &= p(y \mid w, \mathsf{do}(x)) \\ \Leftrightarrow p(Y(x) \mid Z(x), W(x)) &= p(Y(x) \mid W(x)) \end{aligned}$$

2: Interchanging observation and intervention

$$\begin{aligned} p(y \mid z, w, \mathsf{do}(x)) &= p(y \mid w, \mathsf{do}(z), \mathsf{do}(x)) \\ \Leftrightarrow p(Y(x) \mid Z(x), W(x)) &= p(Y(x, z) \mid W(x, z)) \end{aligned}$$

3: Removing interventions:

$$\begin{aligned} p(y \mid w, \mathsf{do}(z), \mathsf{do}(x)) &= p(y \mid w, \mathsf{do}(x)) \\ \Leftrightarrow p(Y(x, z) \mid W(x, z)) &= p(Y(x) \mid W(x)) \end{aligned}$$

Do-calculus (details)

Pearl's do-calculus as originally formulated:

$$1: p(y \mid z, w, do(x)) = p(y \mid w, do(x)) \quad \text{if } (Y \perp\!\!\!\perp Z \mid W, X)_{\mathcal{G}_{\overline{X}}}$$

$$2: p(y \mid z, w, do(x)) = p(y \mid w, do(z), do(x)) \quad \text{if } (Y \perp\!\!\!\perp Z \mid W, X)_{\mathcal{G}_{\overline{X},\underline{Z}}}$$

$$3: p(y \mid w, do(z), do(x)) = p(y \mid w, do(x)) \quad \text{if } (Y \perp\!\!\!\perp Z \mid W, X)_{\mathcal{G}_{\overline{X},\overline{Z(W)}}}$$

where $\mathcal{G}_{\overline{X}}$ denotes the graph obtained from $\mathcal{G}$ by removing all edges with arrowheads into X, $\mathcal{G}_{\underline{Z}}$ denotes the graph obtained from $\mathcal{G}$ by removing all directed edges out of Z, and $Z(W)$ is all elements in Z that are not ancestors of W in $\mathcal{G}_{\overline{X}}$.

Potential Outcomes (PO) Calculus (Malinsky, Shpitser, R, 2019; Shpitser, R, Robins, 2020)

- Can use SWIGs to formulate (simpler, wlog) counterfactual versions of Pearl's rules.

1: If $Y(x)$ is d-separated from $Z(x)$ given $W(x)$ in $\mathcal{G}(x)$ then

$$p(Y(x) \mid Z(x), W(x)) = p(Y(x) \mid W(x))$$

2: If $Y(x, z)$ is d-separated from $Z(x, z)$ given $W(x, z)$ in $\mathcal{G}(x, z)$ then

$$p(Y(x, z) \mid W(x, z)) = p(Y(x) \mid W(x), Z(x) = z)$$

3: If z has no directed path to $Y(x, z)$ in $\mathcal{G}(x, z)$ then

$$p(Y(x, z)) = p(Y(x))$$

- Note: here we use non-minimal labelings: e.g. $Z(x, z)$ is the random node for Z in $\mathcal{G}(x, z)$. (This is just to make explicit which node is in which graph.)

Potential Outcomes Calculus: *TL;DR versions*

Suppressing the intervention on X to reduce clutter:

1: If Y is d-separated from Z given W in $\mathcal{G}$ then

$$p(Y \mid Z, W) = p(Y \mid W) \quad \text{(Markov property)}.$$

2: If $Y(z)$ is d-separated from $Z(z)$ given $W(z)$ in $\mathcal{G}(z)$ then

$$p(Y(z) \mid W(z)) = p(Y \mid W, Z = z) \quad \text{(generalized ignorability)}.$$

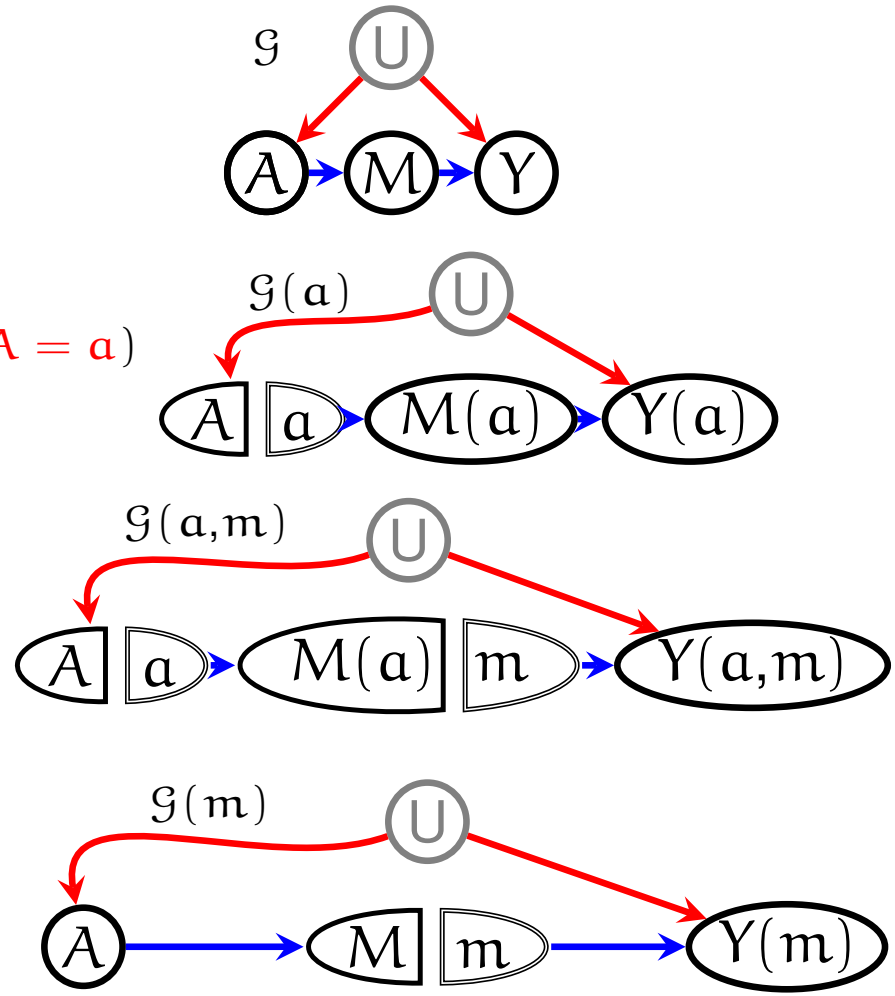
3: If z has no directed path to $Y(z)$ in $\mathcal{G}(z)$ then

$$p(Y(z)) = p(Y) \quad \text{(causal irrelevance)}.$$

po-calculus $=$ d-separation $+$ ignorability
$+$ interventions only affect causal descendants
(!)

Example Derivation (Front-Door)

$$
\begin{aligned}
& p(Y(a)) \\
&= ^{p} \sum_{m} p(Y(a)|M(a)=m)p(M(a)=m) \\
&= ^{2,\mathcal{G}(a)} \sum_{m} p(Y(a)|M(a)=m)p(M=m|A=a) \\
&= ^{2,\mathcal{G}(a,m)} \sum_{m} p(Y(a,m))p(m|a) \\
&= ^{3,\mathcal{G}(a,m)} \sum_{m} p(Y(m))p(m|a) \\
&= ^{p} \sum_{m} p(m|a) \sum_{a'} p(Y(m)|a')p(a') \\
&= ^{2,\mathcal{G}(m)} \sum_{m} p(m|a) \sum_{a'} p(Y|m,a')p(a')
\end{aligned}
$$



Joint Independence

We saw earlier that the causal DAG $X \rightarrow Y$ implied:

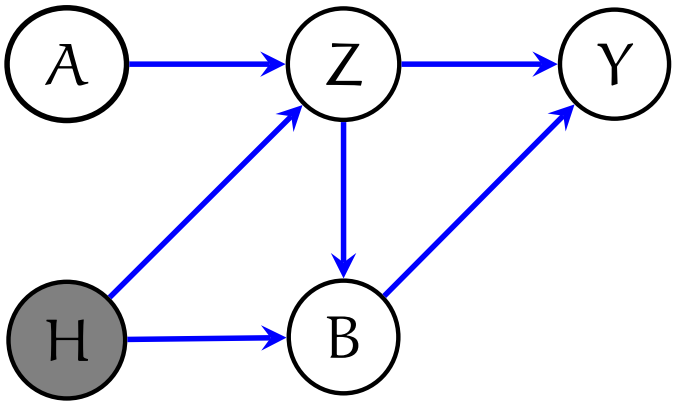
$$X \perp\!\!\!\perp Y(x_0) \qquad \text{and} \qquad X \perp\!\!\!\perp Y(x_1)$$

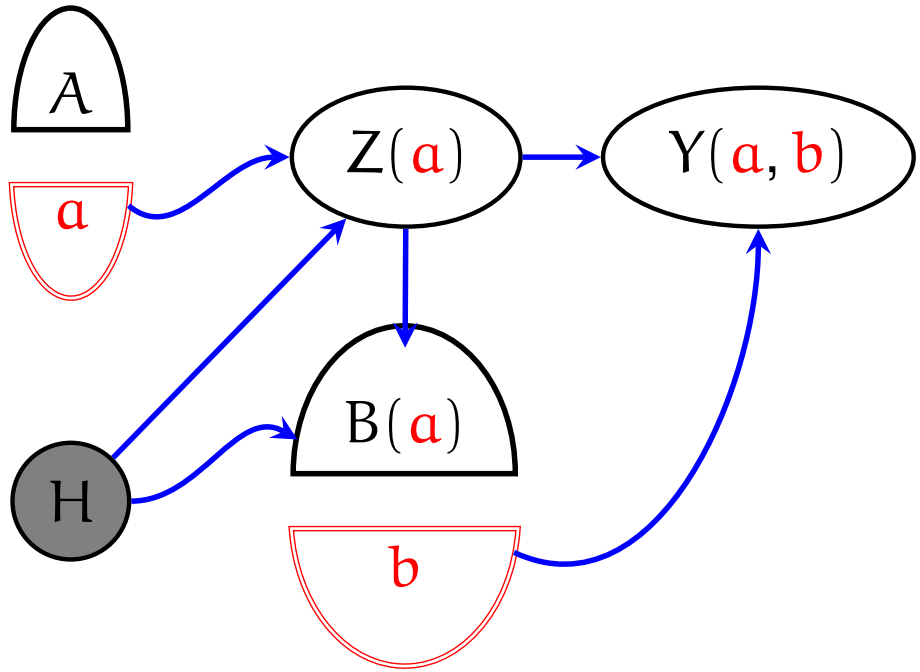
However, *joint* independence relations such as:

$$X \perp\!\!\!\perp \{Y(x_0), Y(x_1)\}$$

never follow from our SWIG transformation:
There is no way via node-splitting to construct a graph with both $Y(x_0)$, and $Y(x_1)$.
This has important consequences for the identification of direct effects.

Inferential Problem Redux:





Pearl (2009), Ex. 11.3.3, claims the causal DAG above does not imply:

$$Y(a, b) \perp\!\!\!\perp B \mid Z, A = a. \tag{10}$$

The SWIG shows that (10) does hold; Pearl is incorrect.
Specifically, we see from the SWIG:

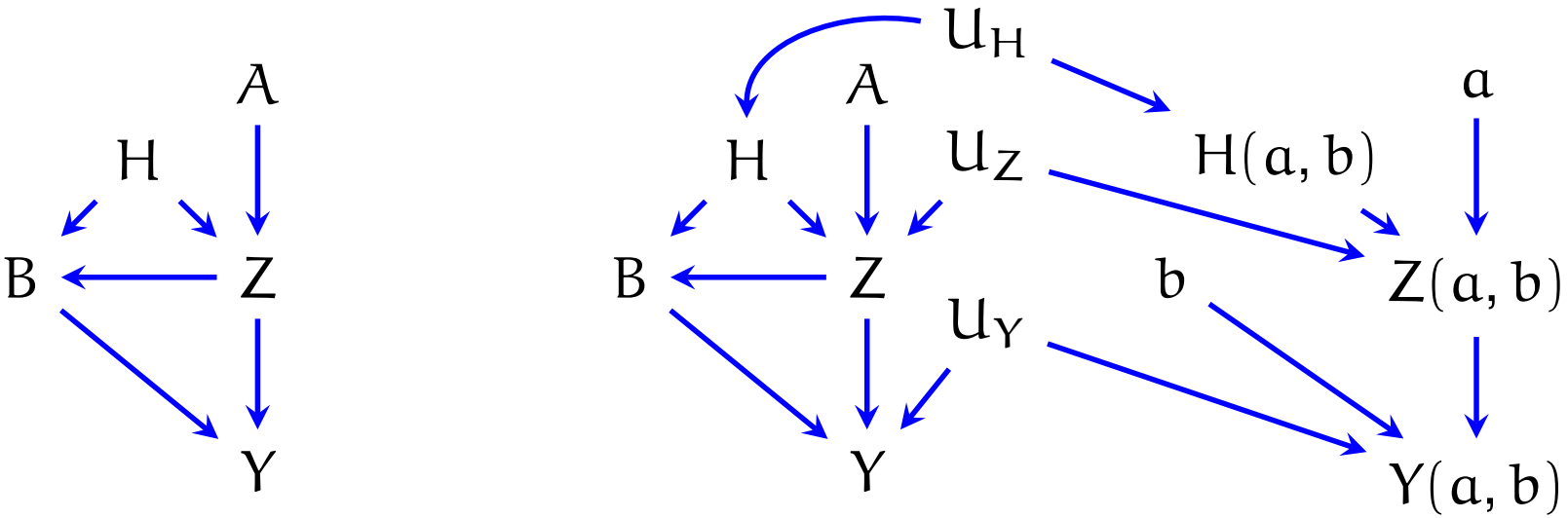
$$Y(a, b) \perp\!\!\!\perp B(a) \mid Z(a), A \tag{11}$$

$$\Rightarrow \quad Y(a, b) \perp\!\!\!\perp B(a) \mid Z(a), A = a \tag{12}$$

This last condition is then equivalent to (10) via consistency.

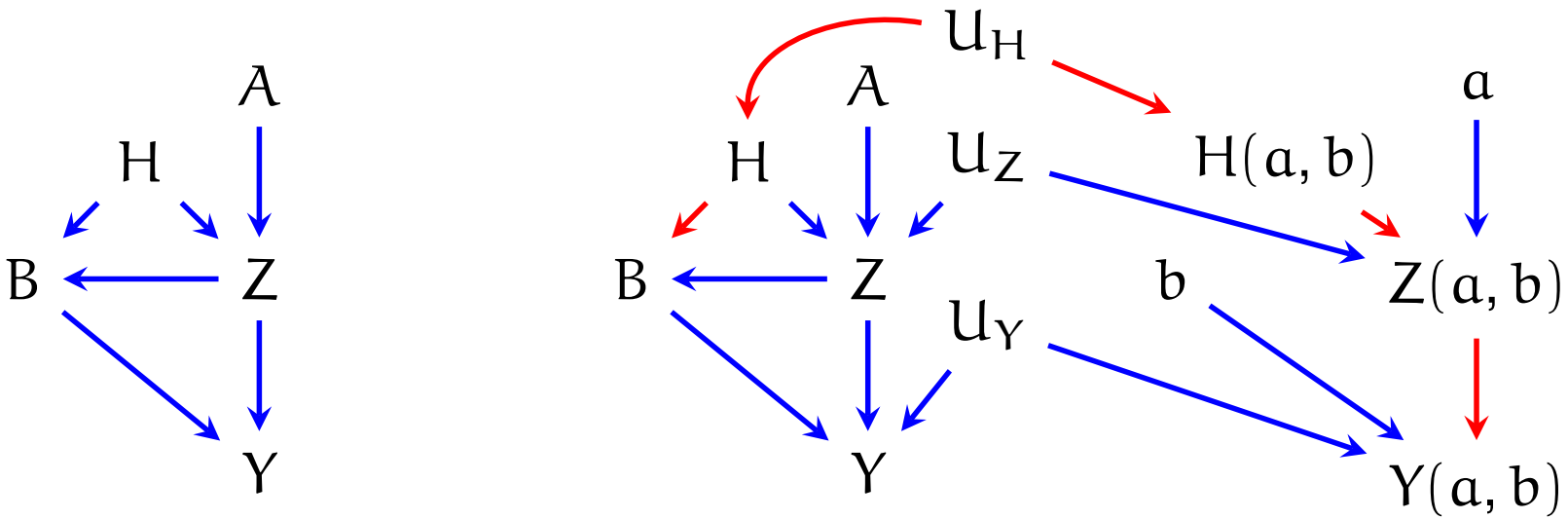
(Pearl infers a claim of Robins is false since if true then (10) would hold).

Pearl's twin network for the same problem



The twin network fails to reveal that $Y(a,b) \perp\!\!\!\perp B \mid Z, A = a$.

Pearl's twin network for the same problem

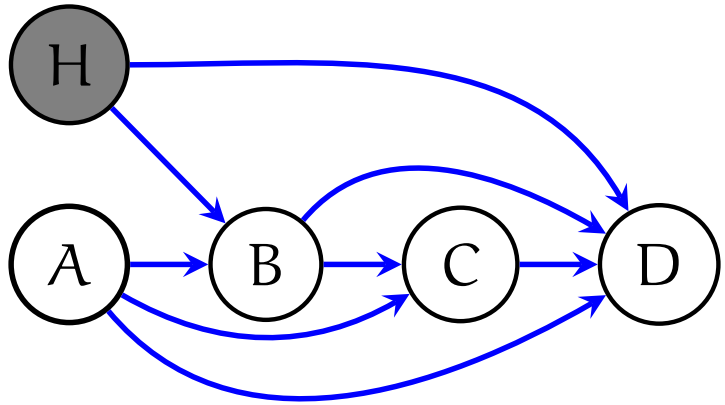


The twin network fails to reveal that $Y(a,b) \perp\!\!\!\perp B \mid Z, A = a$.
This 'extra' independence holds in spite of d-connection because (by consistency) when $A = a$, then $Z = Z(a) = Z(a,b)$.
Note that $Y(a,b) \not\perp\!\!\!\perp B \mid Z, A \neq a$.

Shpitser & Pearl (2008) introduce a pre-processing step to address this.

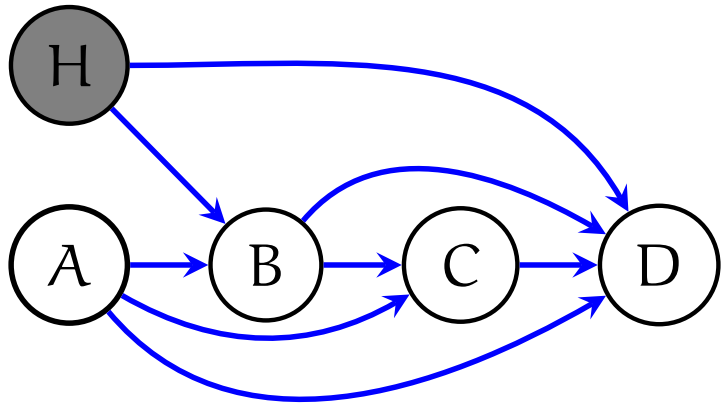
Multiple Treatments

Sequentially randomized experiment (I)



- A and C are treatments;
- H is unobserved;
- B is a time varying confounder;
- D is the final response;
- Treatment C is assigned randomly conditional on the observed history, A and B;
- Want to know $\mathrm{P}(\mathrm{D}(\tilde{a}, \tilde{c}))$.

Sequentially randomized experiment (I)



If the following holds:

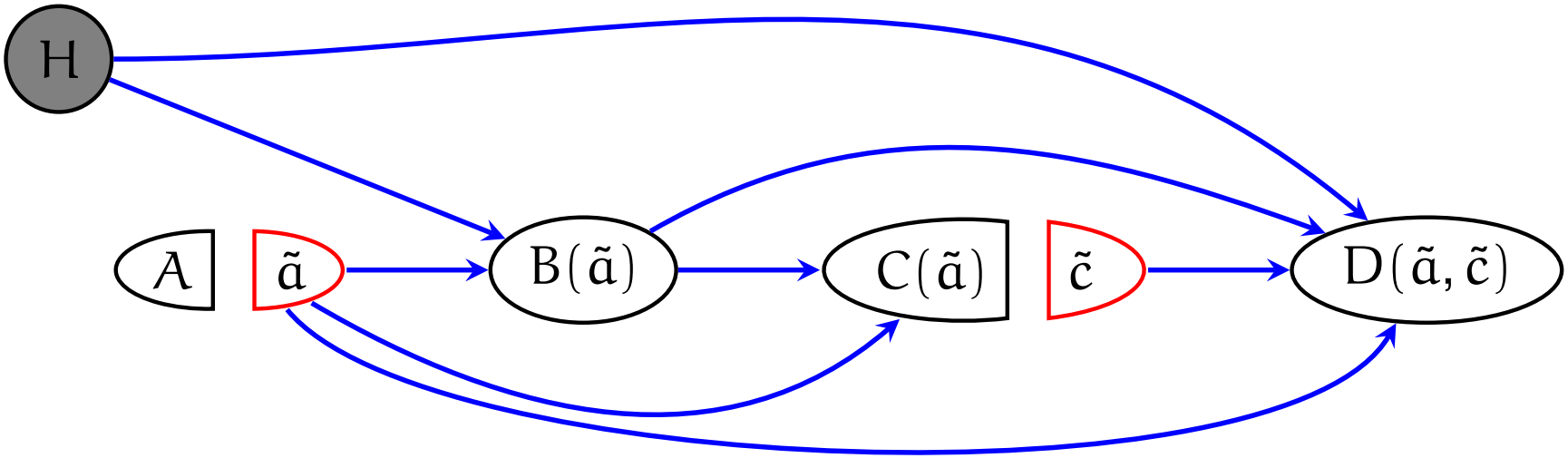
$$\begin{aligned} A &\perp\!\!\!\perp D(\tilde{a},\tilde{c}) \\ C(\tilde{a}) &\perp\!\!\!\perp D(\tilde{a},\tilde{c}) \mid B(\tilde{a}), A \end{aligned}$$

General result of Robins (1986) then implies:

$$P(D(\tilde{a},\tilde{c})=d) = \sum_b P(B=b \mid A=\tilde{a}) P(D=d \mid A=\tilde{a}, B=b, C=\tilde{c}).$$

Does it??

Sequentially randomized experiment (II)



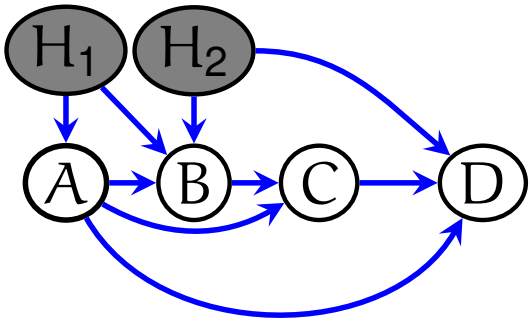
d-separation:

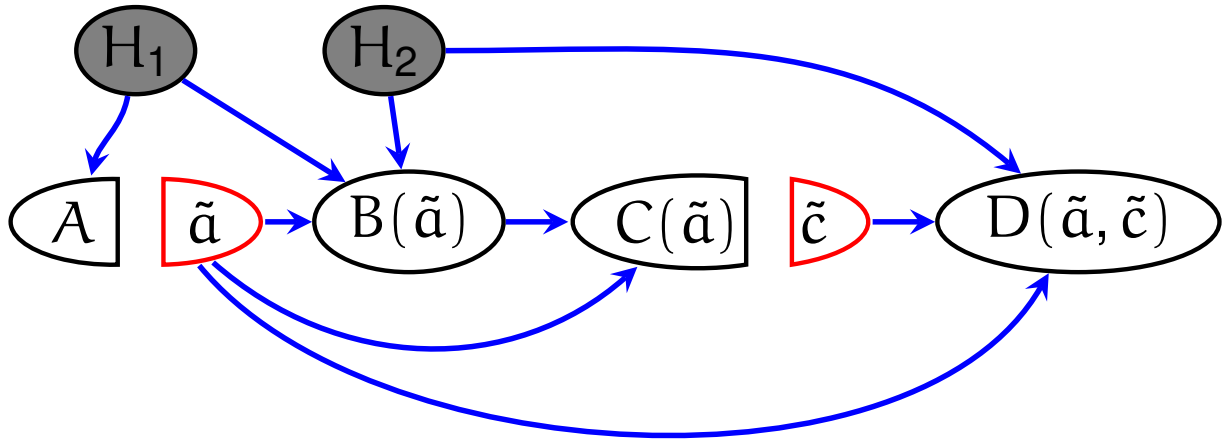
$$A \perp\!\!\!\perp D(\tilde{a}, \tilde{c})$$
$$C(\tilde{a}) \perp\!\!\!\perp D(\tilde{a}, \tilde{c}) \mid B(\tilde{a}), A$$

g-formula of Robins (1986) then implies:

$$P(D(\tilde{a}, \tilde{c}) = d) = \sum_b P(B = b \mid A = \tilde{a}) P(D = d \mid A = \tilde{a}, B = b, C = \tilde{c}).$$

Another example





$$A \perp\!\!\!\perp D(\tilde{a},\tilde{c})$$
$$C(\tilde{a}) \perp\!\!\!\perp D(\tilde{a},\tilde{c}) \mid B(\tilde{a}), A$$

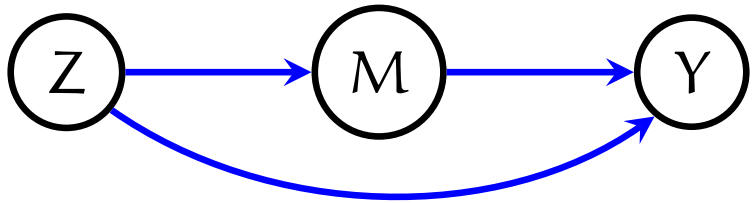
g-formula of Robins (1986) then implies:

$$P(D(\tilde{a},\tilde{c})=d) = \sum_b P(B=b \mid A=\tilde{a})P(D=d \mid A=\tilde{a}, B=b, C=\tilde{c}).$$

Can also see that this identification fails if there is a $B \to D$ edge.

Model dimensions

How many counterfactual independences? (I)



There are 7 counterfactual random variables:

$$Z,\ M(z_0), M(z_1),\ Y(m_0, z_0), Y(m_0, z_1), Y(m_1, z_0), Y(m_1, z_1)$$

Dimension of models:

- No assumptions (allowing confounding): $127 = 2^7 - 1$;
- SWIG (no confounding): 113
- Pearl's NPSEM with indep. errors (no confounding), aka SCM: 19

Number of extra counterfactual independence assumptions: 94

How many counterfactual independences? (II)

No. Actual Vars.	2	3	4	K
Dim. $\mathrm{P}(\mathbf{V})$	3	7	15	$2^{\mathrm{K}}-1$
No. Counterfactual Vars.	3	7	15	$2^{\mathrm{K}}-1$
Dim. Counterfactual Dist.	7	127	32767	$2^{(2^{\mathrm{K}}-1)}-1$
Dim. FFRCISTG	5	113	32697	$(2^{(2^{\mathrm{K}}-1)}-1)-\sum_{\mathrm{j}=1}^{\mathrm{K}-1}(4^{\mathrm{j}}-2^{\mathrm{j}}$
Dim. NPSEM-IE / SCM	4	19	274	$\sum_{\mathrm{j}=0}^{\mathrm{K}-1}(2^{2^{\mathrm{j}}}-1)$
Difference	1	94	32423	$\mathrm{O}(2^{2^{\mathrm{K}}-2})$

Table: Dimensions of counterfactual models associated with complete graphs with binary variables.

Summary so Far

- SWIGs provide a simple way to unify graphs and counterfactuals via node-splitting
- The approach works via linking the factorizations associated with the two graphs.
- The new graph represents a counterfactual distribution that is *identified* from the distribution in the original DAG.
- This provides a language that allows counterfactual and graphical people to communicate.
- (Not covered) Leads to a complete identification algorithm (Extended ID)
  - "Fixing" operation $\Rightarrow$ Splitting $+$ Marginalization
- (Not covered) Can combine information on the absence of individual and population level direct effects.
- (Not covered) Permits formulation of models where interventions on only some variables are well-defined.

References

- Pearl, J. Causal diagrams for empirical research, *Biometrika* 82, 4, 669–709, 1995.
- Richardson, TS, Robins, JM. Single World Intervention Graphs. *CSSS Technical Report No. 128* `http://www.csss.washington.edu/Papers/wp128.pdf`, 2013.
- Robins, JM A new approach to causal inference in mortality studies with sustained exposure periods – applications to control of the healthy worker survivor effect. *Mathematical Modeling* 7, 1393–1512, 1986.
- Robins, JM, VanderWeele, TJ, Richardson TS. Discussion of "Causal effects in the presence of non compliance a latent variable interpretation" by Forcina, A. *Metron* LXIV (3), 288–298, 2007.
- Malinsky, D, Shpitser, I, Richardson. TS. A potential outcomes calculus for identifying conditional path-specific effects. In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, 2019.
- Shpitser, I, Richardson, TS, Robins, JM. Multivariate Counterfactual Systems and Causal Graphical Models. Arxiv `2008.06017`.
- Spirtes, P, Glymour, C, Scheines R. *Causation, Prediction and Search.* Lecture Notes in Statistics 81, Springer-Verlag.

Distributional Consistency

Definition

The set of distributions $\mathcal{P}_A^{\subseteq}$ will be said to obey distributional consistency if, given $B_i \in A$ and $C \subseteq A \setminus \{B_i\}$, where C may be empty, for all y, b, c:

$$p(Y(b,c)=y, B_i(b,c)=b) = p(Y(c)=y, B_i(c)=b), \tag{13}$$

where $Y = V \setminus \{B_i\}$. As a special case, if C is empty then for all y, b:

$$p(Y(b)=y, B_i(b)=b) = p(Y=y, B_i=b). \tag{14}$$