Causality using Graphs

Vanessa Didelez, Karla Diaz-Ordaz, and Robin Evans

BIPS/University of Bremen, UCL, and University of Oxford

APTS Week 4, Southampton
September 2025

Outline

1. Graphs for causal models

- Graphical Motivation
- Conditional Independence
- Directed Acyclic Graphs
- Confounding and Adjustment
- Selection Bias

2. Single-World Intervention Graphs

- NPSEM-IEs
- d-separation in SWIGs
- Adjustment for Confounding

A Causal Story

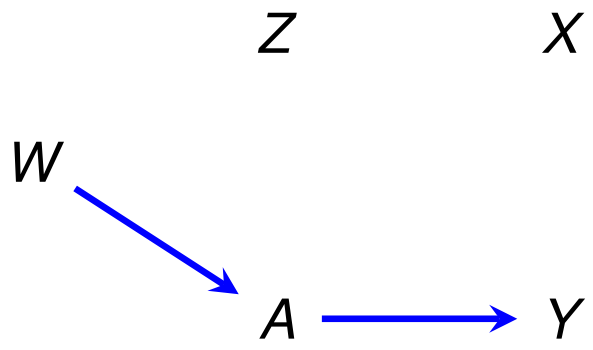
Consider the following situation.

> *An obstetrician is interested in whether giving a vitamin A supplement (A) to new mothers may help to reduce the risk of post-natal depression (Y). She implements an encouragement W to take such supplements, by offering it to a randomly selected subset of half the new mothers in her ward. This is assumed not to have any effect other than increasing the chance of mothers taking vitamin A.*
>
> *She suspects that age (Z) is also a determinant of how likely a mother is to take the supplement, and that this also affects the baby's health (X). This in turn affects the likelihood of post-natal depression, but not the probability of taking the supplement.*
>
> *There is assumed to be no direct effect of age on post-natal depression.*

How should we represent the information contained in this paragraph?

Directed Graphs

Use a graph!



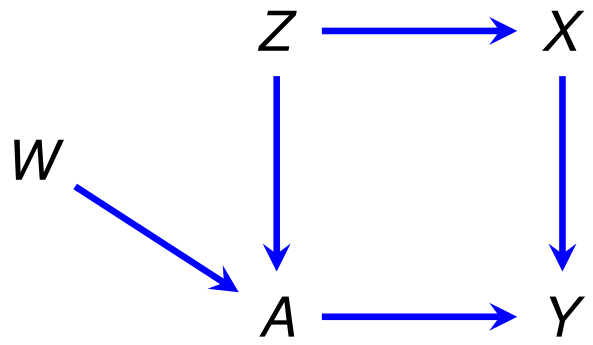
- *Z* age;
- *W* encourage;
- *X* infant health;
- *A* vitamin A;
- *Y* post-natal depression;

[*Y*] is assumed to be directly affected by the treatment (*A*).

W is **randomly** assigned, and affects only *A*.

Directed Graphs

Use a graph!



- *Z* age;
- *W* encourage;
- *X* infant health;
- *A* vitamin A;
- *Y* post-natal depression;

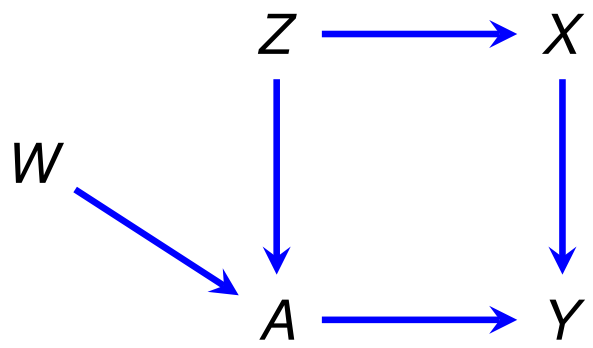
Z may determine *A* and *X*, but not *Y*.

X is predictive of *Y*, but does not affect *A*.

Interpreting a Graph

Now that we've drawn a graph, we get a nice representation of the (possible) causal structure underlying our data.

We can immediately see which paths are **causal** (they're directed!) and which are not.



This is useful **even** if we do not intend to use the graph for inference!

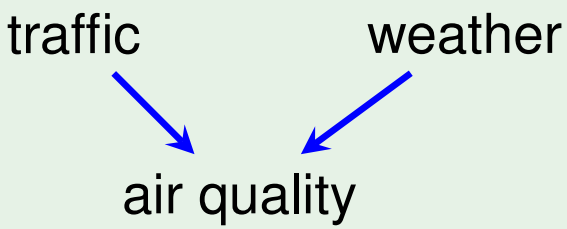
Independence

Definition

Given two random variables X and Y defined on a Cartesian product space, we say that X is **independent** of Y under p (denoted $X \perp\!\!\!\perp Y\,[p]$) if

$$p(x \mid y) = p(x).$$

Example. Air quality is adversely affected by both traffic and local weather conditions, but these two factors may not be related.



Conditional Independence

Definition

Given two random variables X and Y defined on a Cartesian product space, and a third variable Z, we say that X is **conditionally independent** of Y given Z under p (denoted $X \perp\!\!\!\perp Y \mid Z\,[p]$) if

$$p(x \mid y, z) = p(x \mid z).$$

Example. People's genes are conditionally independent of their grandparent's genes, given their parent's genes.

$$\text{GP} \longrightarrow \text{P} \longrightarrow \text{C}$$

Example. Lung cancer is conditionally independent of having yellow fingers, given one's smoking status.

$$\text{yellow} \longleftarrow \text{smoker} \longrightarrow \text{lung cancer}$$

Alternative Characterizations

Theorem

Let X, Y, Z be random variables, with joint density p. Then we can write

$$p(x, y, z) = f(x, z) \cdot g(y, z)$$

if and only if $X \perp\!\!\!\perp Y \mid Z$ [p].

This can be very useful if we only know the density up to a constant of proportionality.

Simpson's Paradox

Conditional independence is sometimes quite unintuitive.

Below is the margin of an infamous dataset on death penalty convictions in Florida between 1976 and 1987.

Death Penalty?	Defendant's Race	
	White	Black
Yes	53	15
No	430	176

White defendants are slightly more likely than black defendants to face the death penalty.

Simpson's Paradox

Here is the full dataset.

Victim's Race	Death Penalty?	Defendant's Race	
		White	Black
White	Yes	53	11
	No	414	37
Black	Yes	0	4
	No	16	139

Now we can see that if we condition on the victim's race, the dependence of the penalty applied conditional on the defendant's race is completely reversed!

Morals

Let:

- D be an indicator that the death penalty was imposed;
- V be an indicator for the race of the victim;
- R be an indicator for the race of the defendant.

By changing the numbers only very slightly, it is easy to obtain either:

$$D \perp\!\!\!\perp R \quad \text{and} \quad D \not\perp\!\!\!\perp R \mid V,$$

$$\text{or} \quad D \not\perp\!\!\!\perp R \quad \text{and} \quad D \perp\!\!\!\perp R \mid V.$$

Graphoids

Conditional independences obey several rules called **semi-graphoid axioms** (though in this context they are not really axioms!) These are:

1. Symmetry: $X \perp\!\!\!\perp Y \mid Z \Rightarrow Y \perp\!\!\!\perp X \mid Z$;
2. Decomposition: $X \perp\!\!\!\perp Y, W \mid Z \Rightarrow X \perp\!\!\!\perp Y \mid Z$;
3. Weak union: $X \perp\!\!\!\perp Y, W \mid Z \Rightarrow X \perp\!\!\!\perp W \mid Y, Z$;
4. Contraction: $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid Y, Z \Rightarrow X \perp\!\!\!\perp Y, W \mid Z$.

We can summarize axioms 2–4 as a 'chain rule':

$$X \perp\!\!\!\perp Y \mid Z \quad \text{and} \quad X \perp\!\!\!\perp W \mid Y, Z \quad \Longleftrightarrow \quad X \perp\!\!\!\perp W, Y \mid Z.$$

In addition, if $p > 0$ then we have:

5. Intersection:
$$X \perp\!\!\!\perp Y \mid W, Z \text{ and } X \perp\!\!\!\perp W \mid Y, Z \Rightarrow X \perp\!\!\!\perp Y, W \mid Z.$$

All five rules are called the **graphoid axioms**.

Graphoids

Conditional independences obey several rules called **semi-graphoid axioms** (though in this context they are not really axioms!) These are:

1. Symmetry: $X \perp\!\!\!\perp Y \quad \Rightarrow \; Y \perp\!\!\!\perp X \quad ;$
2. Decomposition: $X \perp\!\!\!\perp Y, W \quad \Rightarrow \; X \perp\!\!\!\perp Y \quad ;$
3. Weak union: $X \perp\!\!\!\perp Y, W \quad \Rightarrow \; X \perp\!\!\!\perp W \mid Y \quad ;$
4. Contraction: $X \perp\!\!\!\perp Y \quad \text{and } X \perp\!\!\!\perp W \mid Y \quad \Rightarrow \; X \perp\!\!\!\perp Y, W \quad .$

We can summarize axioms 2–4 as a 'chain rule':

$$X \perp\!\!\!\perp Y \quad \text{and} \quad X \perp\!\!\!\perp W \mid Y \quad \Longleftrightarrow \quad X \perp\!\!\!\perp W, Y \quad .$$

In addition, if $p > 0$ then we have:

5. Intersection:
$$X \perp\!\!\!\perp Y \mid W \quad \text{and } X \perp\!\!\!\perp W \mid Y \quad \Rightarrow \; X \perp\!\!\!\perp Y, W \quad .$$

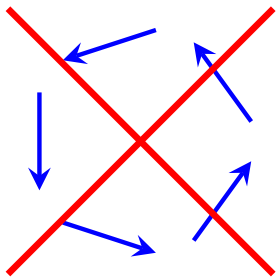
All five rules are called the **graphoid axioms**.

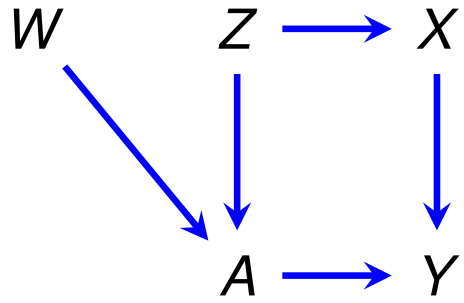
Directed Acyclic Graphs

vertices $V \in \boldsymbol{V}$

edges 

no directed cycles





directed acyclic graph (DAG), $\mathcal{G}$

We will associate the vertices/nodes with **random variables**, and the edges will denote **causal** dependence.

Terminology and Notation

For a DAG $\mathcal{G}$ with vertices $\boldsymbol{V}$...

If ...	we say...	and we write...
$X \to Y$	X is a **parent** of Y Y is a **child** of X	$X \in \mathrm{pa}_{\mathcal{G}}(Y)$ $Y \in \mathrm{ch}_{\mathcal{G}}(X)$
$X \to \cdots \to Y$ or $X = Y$	X is an **ancestor** of Y Y is a **descendant** of X	$X \in \mathrm{an}_{\mathcal{G}}(Y)$ $Y \in \mathrm{de}_{\mathcal{G}}(X)$

A **path** is a sequence of adjacent edges, without repeating a vertex. Its length is the number of edges (possibly 0).

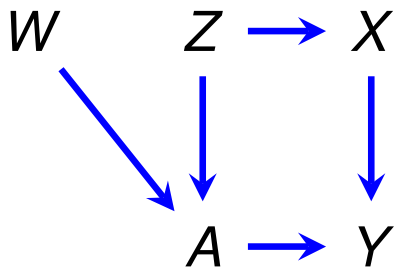
A **directed path** is a path where all the edges are oriented pointing towards the final vertex.

A **directed cycle** is a directed path from X to Y and an edge $Y \to X$.

Given a **topological ordering** (parents precede their children) of the variables $V_1, \ldots, V_p$ we write $\mathrm{pre}_<(i) = \{1, \ldots, i-1\}$ for each i.

DAG Models (aka Bayesian Networks)

graph $\mathcal{G}$



$\Longleftrightarrow$

model $\mathcal{M}(\mathcal{G})$

$$p(\boldsymbol{v}) = \prod_{v \in V} p(v \mid \mathrm{pa}_{\mathcal{G}}(v)).$$

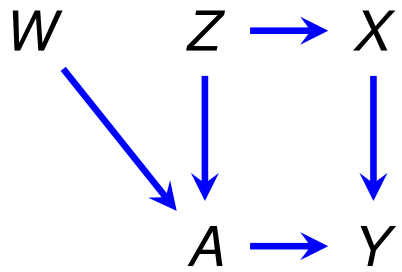
(factorization)

So in example above:

$$p(\boldsymbol{v}) = p(w) \cdot p(z) \cdot p(x \mid z) \cdot p(a \mid w, z) \cdot p(y \mid a, x).$$

DAG Models (aka Bayesian Networks)

Can also define model as a list of conditional independences:



pick a topological ordering $<$ of the graph: e.g. W, Z, X, A, Y.

Can **always** factorize a joint distribution as:

$$p(w, z, a, x, y) = p(w) \cdot p(z \mid w) \cdot p(x \mid w, z) \cdot p(a \mid w, z, x) \cdot p(y \mid w, z, a, x).$$

The model is the same as setting (e.g.)

$$p(y \mid w, z, a, x) = p(y \mid a, x) = p(y \mid \text{pa}(y)).$$

Thus $\mathcal{M}(\mathcal{G})$ is precisely distributions such that:

$$V_i \perp\!\!\!\perp \boldsymbol{V}_{\text{pre}_<(i) \setminus \text{pa}(i)} \mid \boldsymbol{V}_{\text{pa}(i)}, \qquad i = 1, \ldots, |\boldsymbol{V}|.$$

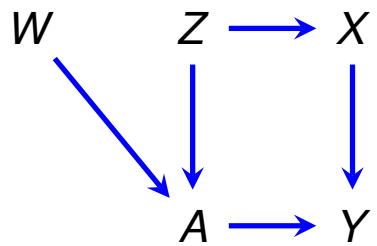
Ordered Markov Property

We say that p obeys the **ordered local Markov property** with respect to $\mathcal{G}$ and a topological ordering $<$ if:

$$V_i \perp\!\!\!\perp \boldsymbol{V}_{\mathrm{pre}_<(i) \setminus \mathrm{pa}(i)} \mid \boldsymbol{V}_{\mathrm{pa}(i)}, \qquad i = 1, \ldots, |\boldsymbol{V}|.$$

In our example, with the order W, Z, X, A, Y this means

$$Z \perp\!\!\!\perp W \qquad X \perp\!\!\!\perp W \mid Z$$
$$A \perp\!\!\!\perp X \mid W, Z \qquad Y \perp\!\!\!\perp W, Z \mid A, X.$$



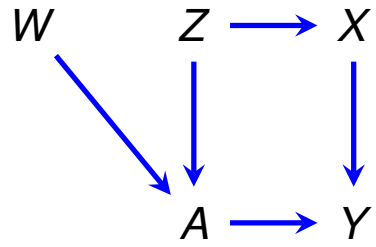
If we switch A and X, we get

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

which is equivalent (this may not be obvious, but you can check with semi-graphoids!)

d-Separation

Note that we can also obtain other independences using the graphoid axioms:



$$X \perp\!\!\!\perp W, A \mid Z \quad \text{and} \quad Y \perp\!\!\!\perp W, Z \mid X, A$$
$$\Rightarrow X \perp\!\!\!\perp W \mid Z, A \quad \text{and} \quad Y \perp\!\!\!\perp W \mid Z, X, A$$
$$\Rightarrow X, Y \perp\!\!\!\perp W \mid Z, A \qquad \Rightarrow \qquad Y \perp\!\!\!\perp W \mid Z, A.$$

Is there a way to deduce these directly?

Yes! We can use **d-separation**.

A **path** π is a sequence of adjacent edges, without repeating any vertex.

Examples:

$Z \to X \to Y$; (this is a **directed** path)

$W \to A \to Y \leftarrow X$.

d-Separation

For any path, the internal vertices are either:

- **colliders:** i.e. $\rightarrow V \leftarrow$; or
- **non-colliders:** i.e. $\rightarrow V \rightarrow$ or $\leftarrow V \rightarrow$ or $\leftarrow V \leftarrow$.

We say a path π from A to B is **open** given a set $\boldsymbol{C}$ if and only if

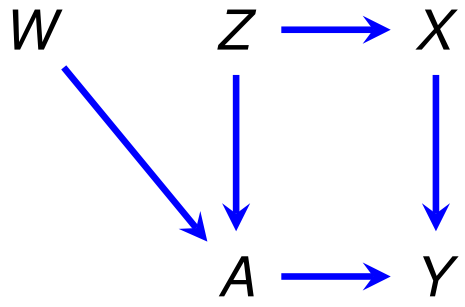
- no non-colliders on π are in $\boldsymbol{C}$; and
- every collider on π is an ancestor of something in $\boldsymbol{C}$.

Otherwise π is **blocked** (or **closed**).

Definition

We say that sets of vertices $\boldsymbol{A}$ and $\boldsymbol{B}$ are **d-separated** given $\boldsymbol{C}$ if **every** path from any $A \in \boldsymbol{A}$ to any $B \in \boldsymbol{B}$ is blocked by $\boldsymbol{C}$.

d-Separation Example



Is $\{W, A\}$ d-separated from $\{X\}$ by $\{Z\}$?

Is $\{W, A\}$ d-separated from $\{X\}$ by $\{Z, Y\}$?

Global Markov Property

Definition

A distribution P is said to obey the **global Markov property** with respect to a DAG $\mathcal{G}$ if whenever

$$\boldsymbol{A} \perp_d \boldsymbol{B} \mid \boldsymbol{C} \quad \text{in } \mathcal{G},$$

we have

$$V_{\boldsymbol{A}} \perp\!\!\!\perp V_{\boldsymbol{B}} \mid V_{\boldsymbol{C}} \quad \text{in } P.$$

In other words, d-separation implies conditional independence.

In addition to being 'sound' d-separation is **complete**: that is, any triple not d-separated is generally **not** independent.

That is: d-separation gives **all** independences implied by the model!

Markov Properties

We now have three equivalent **Markov properties** (models) which we can associate with DAGs.

Factorization. That is (if P has a density p) we have

$$p(\boldsymbol{v}) = \prod_{v \in \boldsymbol{v}} p(v \mid \mathrm{pa}_{\mathcal{G}}(v)).$$

(Ordered) Local Markov Property. For any topological ordering $\prec$, we have

$$V_i \perp\!\!\!\perp \boldsymbol{V}_{\mathrm{pre}(i;\prec)\setminus \mathrm{pa}(i)} \mid \boldsymbol{V}_{\mathrm{pa}(i)} \; [P].$$

Global Markov Property. Whenever $\boldsymbol{A}$ and $\boldsymbol{B}$ are d-separated by $\boldsymbol{C}$ in $\mathcal{G}$ then

$$V_{\boldsymbol{A}} \perp\!\!\!\perp V_{\boldsymbol{B}} \mid V_{\boldsymbol{C}} \; [P].$$

Structural Equations

An alternative model considers each variable to be generated from a **structural equation**:

$$X_v = f_v(X_{\mathrm{pa}(v)}, E_v)$$

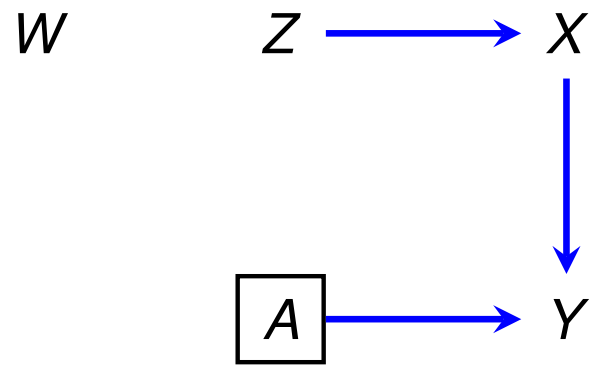
for a measurable function f_v and a noise term E_v.
The noise terms are assumed independent for DAGs.

Implications of this definition are completely equivalent to the others!

In a causal setting they are sometimes called **structural causal models** (Pearl, 2009; Peters et al., 2017), though we prefer the term **structural equation models**.

Causal Models

A DAG can also encode causal information:



If we intervene to experiment (*do*) on A, delete incoming edges.

In distribution, delete factor corresponding to A:

$$p(w, z, x, a, y) = p(w) \cdot p(z) \cdot p(x \mid z) \cdot p(a \mid w, z) \cdot p(y \mid x, a).$$

$$p(w, z, x, y \mid do(a)) = p(w) \cdot p(z) \cdot p(x \mid z) \quad \times \quad p(y \mid x, a).$$

All other factors are preserved (if causal DAG correctly specified).

Causal Effects

The function $p(\cdot \mid do(a))$ is just like any ordinary probability distribution, and obeys the same rules of conditioning and marginalization.

In particular, we can define expectations in the usual way:

$$\mathbb{E}[Y \mid do(A = a)] := \sum_y y \cdot p(y \mid do(a)).$$

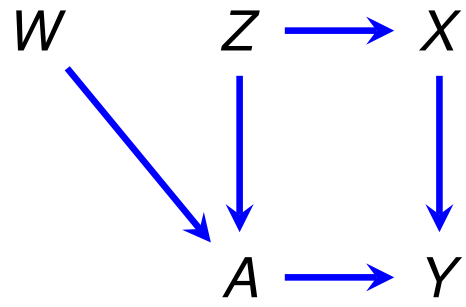
Equipped with this distribution, we can now define the **average treatment effect** of A on Y:

$$\text{ATE} := \mathbb{E}[Y \mid do(A = 1)] - \mathbb{E}[Y \mid do(A = 0)].$$

This is sometimes called the **average causal effect** (ACE) or the **total effect**.

Confounding

In this case there are other variables that causally affect both propensity to take the intervened A and our outcome Y.



For example, suppose older mothers ($Z = 1$) are more likely to take vitamin A (A), and their infants generally have worse health outcomes (X) which reduces their overall mental health level (Y).

A naïve estimate $\mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0]$ includes correlation due to this **confounding**.

This is **not** a causal quantity, since if we actually intervene to set A (e.g. by randomization), the contrast will (generally) be different.

Adjustment Using Parents

Note that we have

$$p(w, z, x, y \mid do(a)) = \frac{p(w, z, x, a, y)}{p(a \mid w, z)}.$$

Hence, to obtain (e.g.) $p(y \mid do(a))$ we just marginalize:

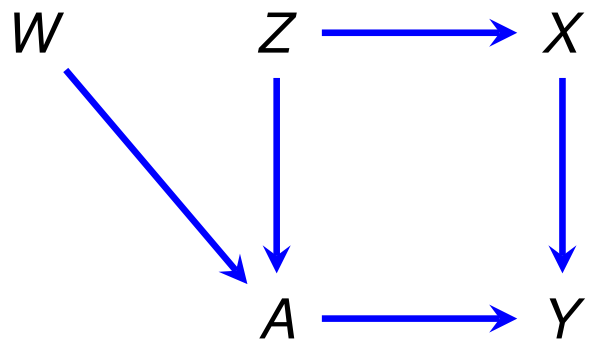
$$\begin{aligned}\sum_{w,z,x} p(y, w, z, x \mid do(a)) &= \sum_{w,z,x} \frac{p(y, w, z, x, a)}{p(a \mid w, z)} \\ &= \sum_{w,z,x} p(w, z) \cdot p(x, y \mid w, z, a) \\ &= \sum_{w,z} p(w, z) \cdot p(y \mid w, z, a).\end{aligned}$$

In this case we call $\{W, Z\}$ an **adjustment set** for the effect of A on Y.

The set of parents of a variable is **always** a valid adjustment set.

Adjustment sets are much more general that this, however.

Back-Door Paths



A **back-door** path from A to Y starts with an arrowhead at A.

Example. $A \leftarrow Z \rightarrow X \rightarrow Y$.

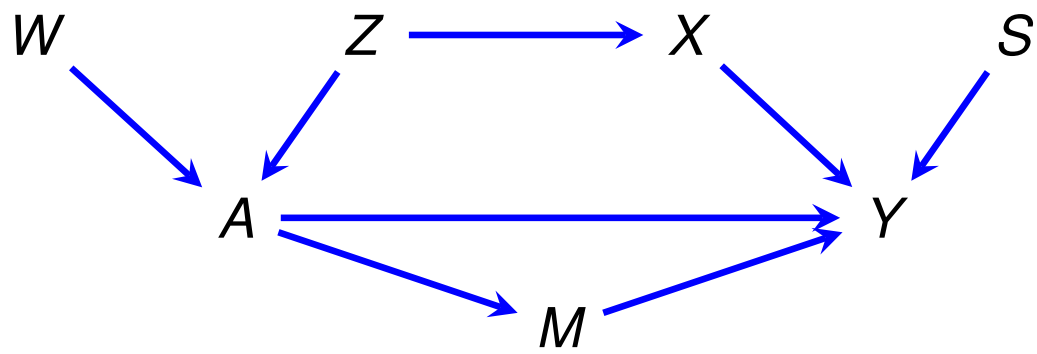
To identify $p(y \mid do(a))$ we must block all back-door paths **without** blocking any causal ones, nor inducing any selection bias.

Back-Door Criterion

Definition

A **back-door adjustment set** for the pair (A, Y) is one which:

- blocks all back-door paths from A to Y;
- does not contain any descendants of A.



Examples:

$\{Z\}$, $\{X\}$, $\{Z, X\}$
$\{W, Z\}$, $\{W, X\}$, $\{W, Z, X\}$
$\{S, Z\}$, $\{S, X\}$, $\{S, Z, X\}$
$\{S, W, Z\}$, $\{S, W, X\}$, $\{S, W, Z, X\}$.

The **optimal adjustment set** is just $\{X, S\}$.

Back-Door Adjustment

Theorem (Pearl, 1993)

Suppose that p is causally Markov w.r.t. $\mathcal{G}$, and that we are interested in the causal effect of A on Y. Then this can be identified by

$$p(y \mid do(a)) = \sum_{x_C} p(x_C) \cdot p(y \mid a, x_C),$$

provided that X_C represents a back-door adjustment set for (A, Y).

Proof.

Since the back-door adjustment set contains no descendants of A, we have $A \perp\!\!\!\perp X_C \mid X_{\mathrm{pa}(a)}$.

It is also easy to see that $Y \perp\!\!\!\perp X_{\mathrm{pa}(a)} \mid A, X_C$ if X_C blocks all back-door paths and A all causal paths.

Back-Door Adjustment (ctd.)

Then:

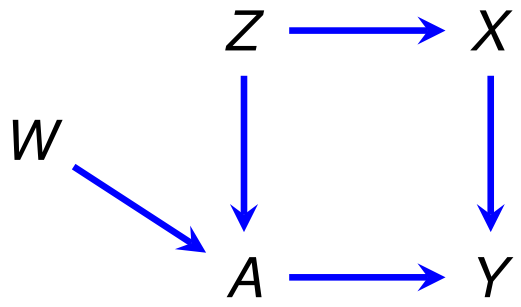
$$
\begin{aligned}
p(y \mid do(a)) &= \sum_{x_{\mathrm{pa}(a)}} p(x_{\mathrm{pa}(a)}) \cdot p(y \mid a, x_{\mathrm{pa}(a)}) \\
&= \sum_{x_{\mathrm{pa}(a)}} p(x_{\mathrm{pa}(a)}) \sum_{x_C} p(y \mid x_C, a, x_{\mathrm{pa}(a)}) \cdot p(x_C \mid a, x_{\mathrm{pa}(a)}) \\
&= \sum_{x_{\mathrm{pa}(a)}} p(x_{\mathrm{pa}(a)}) \sum_{x_C} p(y \mid x_C, a) \cdot p(x_C \mid x_{\mathrm{pa}(a)}) \\
&= \sum_{x_C} p(y \mid x_C, a) \sum_{x_{\mathrm{pa}(a)}} p(x_{\mathrm{pa}(a)}) \cdot p(x_C \mid x_{\mathrm{pa}(a)}) \\
&= \sum_{x_C} p(x_C) \cdot p(y \mid a, x_C).
\end{aligned}
$$

Running Example

Recall our running example: does vitamin A supplementation (A) improve mental health outcomes in new mothers (Y)?

There are two paths from A to Y:

- the **causal path** $A \to Y$;
- the **back-door path** $A \leftarrow Z \to X \to Y$.



The distribution we need is $p(y \mid do(a))$, and the possible back-door adjustment sets are:

$$\{Z\} \quad \{X\} \quad \{Z, X\}$$
$$\{W, Z\} \quad \{W, X\} \quad \{W, Z, X\}.$$

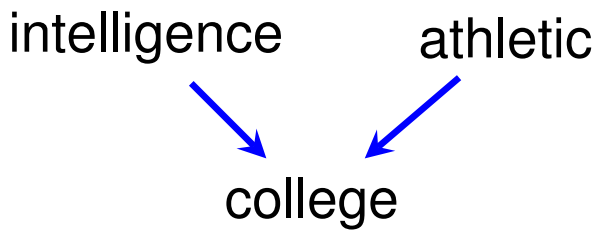
Note that $p(y \mid do(a)) \neq p(y \mid a)$ and $p(y \mid do(a)) \neq p(y)$.

References

Pearl, J. Graphical models, causality and intervention (Comment on: Bayesian analysis in expert systems) *Statistical Science*, 8(3), pp.266-269, 1993.

Selection Bias

In observational (non-randomized) data, bias can come from various sources; the most common is confounding, but **selection bias** is also a big concern.



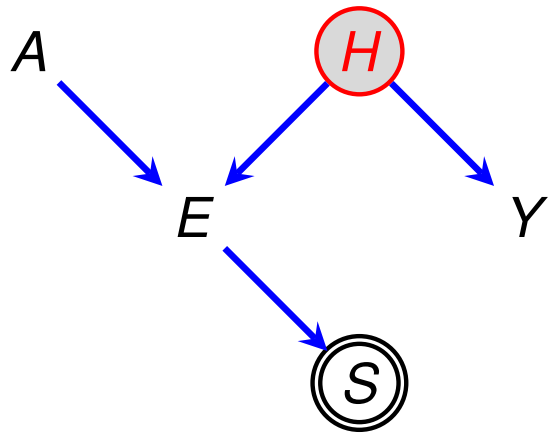
If we only observe people on a University Campus, we may incorrectly believe that intelligence and athletic ability are **negatively** correlated.

This is also referred to as **collider bias** or **Berkson's paradox**.

Side Effects

Suppose that patients may differentially drop out of a study due to side-effects.

- H — general health;
- E — side effects;
- S — drop out (only observe $S = 1$).

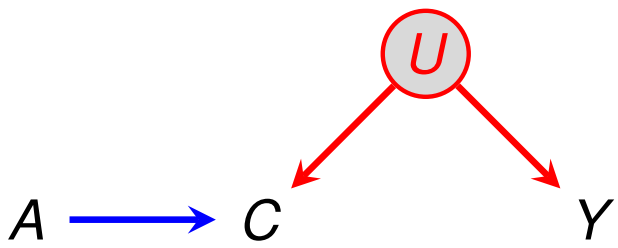


In this case we may erroneously think that the treatment is helpful, when really there is no effect.

Post-treatment Variables

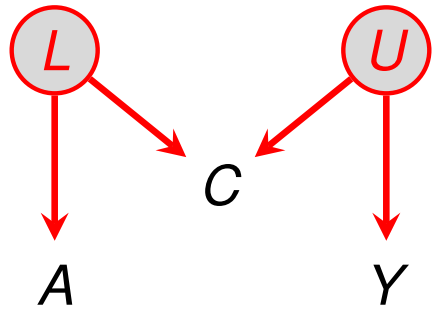
It is generally a mistake to control for **post-treatment** variables, since it may block the causal effect:

Indeed, the reverse problem can also occur!



M-bias

Another concern is so-called **M-bias**, which can arise if we try to condition on pre-treatment covariates but actually **open** a non-causal path by doing so.



Suppose that treatment A is smoking behaviour, C is childhood asthma and Y is adult asthma; L is parental smoking, U is underlying atopy.

Note that the back-door path is marginally blocked, but conditioning upon (only) C opens it!

The length of the path (four edges) means it is unlikely to be a strong bias in practice, however.

References

Pearl, J. *Causality: Models, Reasoning, and Inference.* 3rd Ed. Cambridge, 2009.

Spirtes, P, Glymour, C, Scheines R. *Causation, Prediction, and Search.* Lecture Notes in Statistics 81, Springer-Verlag, 2000.

Wright, S. The theory of path coefficients. *Genetics*, 8: 239–255, 1923.

Wright, S. The method of path coefficients. *Annals of Mathematical Statistics*, 5(3): 161–215, 1934.

Outline

1. Graphs for causal models
   - Graphical Motivation
   - Conditional Independence
   - Directed Acyclic Graphs
   - Confounding and Adjustment
   - Selection Bias

2. Single-World Intervention Graphs
   - NPSEM-IEs
   - d-separation in SWIGs
   - Adjustment for Confounding

Structural causal models

In machine learning it is common to use **structural causal models** (SCMs) to represent causal systems.

These originate with the work of Sewall Wright in the 1920s, and he referred to them as **structural equation models** (SEMs).

Each variable (say X_v) is written as a function of its parents and a noise term:

$$X_v \leftarrow f_v(X_{\text{pa}(v)}, \varepsilon_v).$$

Often the noise terms are assumed to be **independent**.

Note that we can also write this using potential outcome notation:

$$X_v = f_v(X_{\text{pa}(v)}, \varepsilon_v) = X_v(X_{\text{pa}(v)}).$$

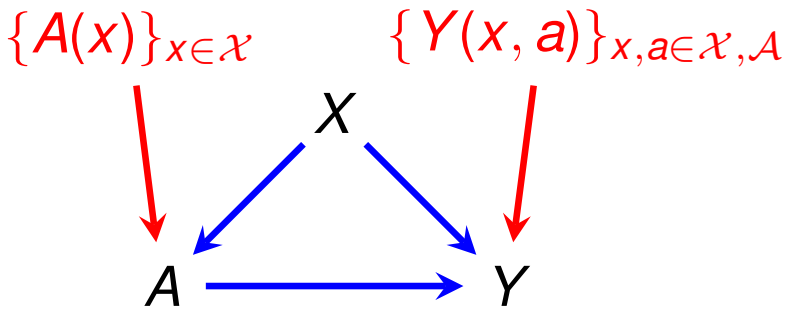
NPSEM-IEs

If the errors are independent, the model is referred to as a **non-parameteric SEM with independent errors** (NPSEM-IE) by Richardson and Robins (2013).

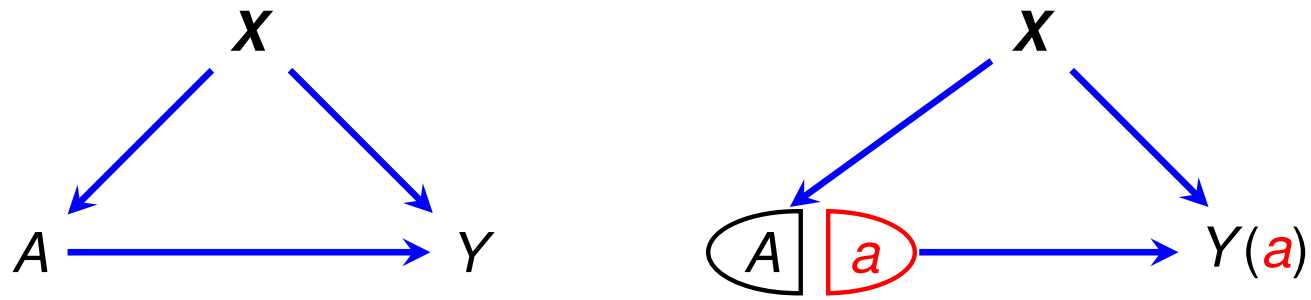
They note that it implicitly makes **cross-world** independence assumptions. For example, in the graph below, we would have (e.g.)

$$A(x') \perp\!\!\!\perp Y(x, a), \qquad \forall x, x'.$$

This is completely untestable **using any randomized trial**.



Single-World Intervention Graphs

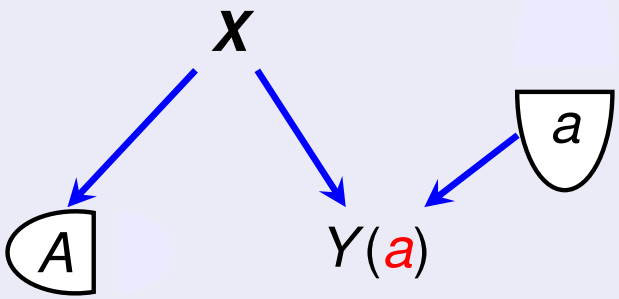


Single-World Intervention Graphs (SWIGs) combine graphs and potential outcomes so as to allow one to read off important conditions (see Richardson and Robins, 2013).

Note we can see by d-separation that the 'no unobserved confounding' assumption holds under this SWIG:

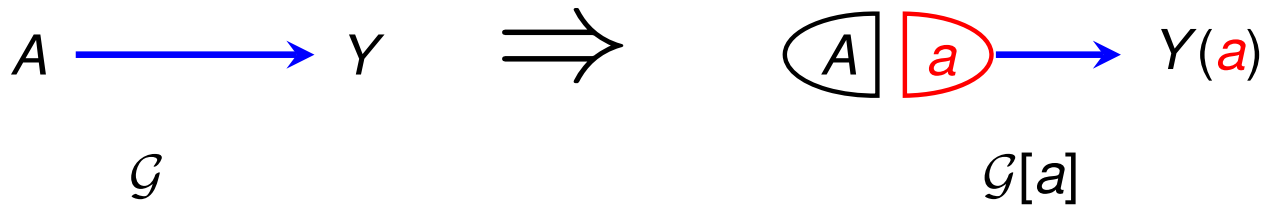
$$Y(a) \perp\!\!\!\perp A \mid \boldsymbol{X}.$$

Once nodes are split we can rearrange them:



Representing an intervention

$$P(A = a, Y = y) = P(A = a)P(Y = y \mid A = a)$$



The graph says that $Y(a) \perp_d A$, and hence:

$$P(Y(a)) = P(Y(a) \mid A = a) =^c P(Y \mid A = a), \qquad \forall a.$$

Notice that, for two distinct values a, a' of A, we **never** observe $Y(a)$ and $Y(a')$ on the same graph.

In particular, SWIGs will never say that $A \perp\!\!\!\perp \{Y(a), Y(a')\}$ if $a \neq a'$.

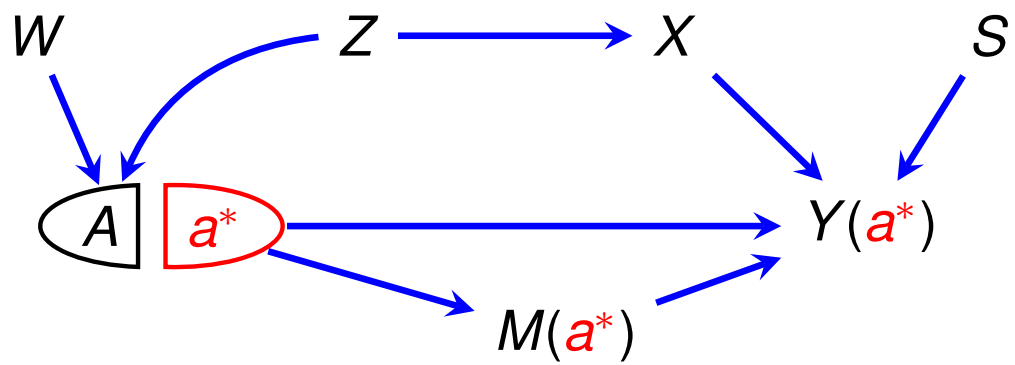
This is what is meant by **single-world** in the name of the class of graphs.

This has important consequences for the identification of direct effects.

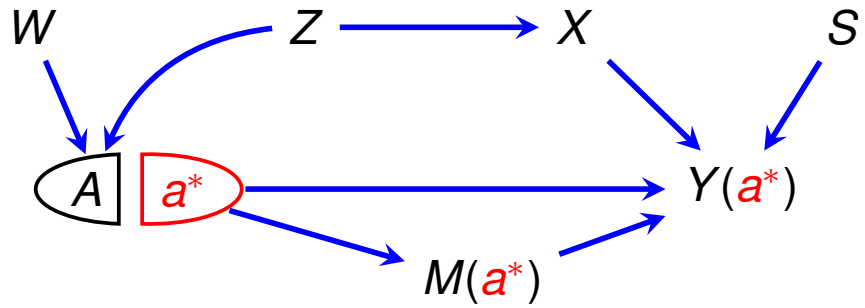
Node-splitting

What happens when we intervene in a SWIG?

1. Split the node(s) $\boldsymbol{V}_A$ being intervened on into $\boldsymbol{V}_A$ and $\boldsymbol{v}_A^*$.
2. Replace all descendants of $\boldsymbol{v}_A^*$ by $V(\boldsymbol{v}_A^*)$.
3. In the factorization, replace every instance of $\boldsymbol{v}_A$ with $\boldsymbol{v}_A^*$, and all descendants of $\boldsymbol{V}_A$ with $V(\boldsymbol{v}_A^*)$.



Node-splitting



Intervene to set $A = a^*$:

1. Add potential outcome to all descendants of A;
2. Remove any conditioning on $A = a$.

$$P(W, Z, X, S, A, M, Y) = P(W) \cdot P(Z) \cdot P(X \mid Z) \cdot P(S) \cdot \\ \times P(A \mid W, Z) \cdot P(M \mid A) \cdot P(Y \mid S, A, M)$$

$$P(W, Z, X, S, A, M(a^*), Y(a^*)) = P(W) \cdot P(Z) \cdot P(X \mid Z) \cdot P(S) \\ \times P(A \mid W, Z) \cdot P(M(a^*)) \cdot P(Y(a^*) \mid S, M(a^*)).$$

Note that we can replace **all** variables V with $V(a^*)$, but only affects the descendants of A.

Intuition behind node splitting

Question

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

Answer

Whenever a patient is observed to swallow the drug, we instantly intervene by administering a safe 'emetic' that causes the pill to be regurgitated before any drug can enter the bloodstream.

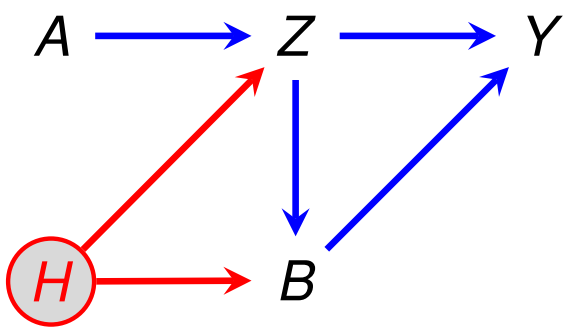
Since we assume the emetic has no side effects, the patient's recorded outcome is then $Y(a = 0)$.

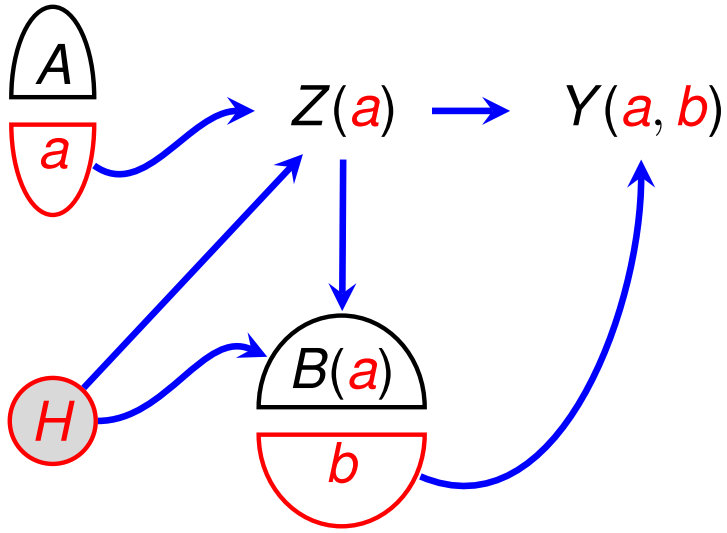
Hence the SWIG represents quantities that (at least in principle) are causally identifiable by an experiment; e.g.

$$\text{ETT} := \mathbb{E}[Y(1) \mid A = 1] - \mathbb{E}[Y(0) \mid A = 1].$$

(Robins et al. 2007)

Harder inferential problem





Query

Does this causal graph imply:

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

Answer

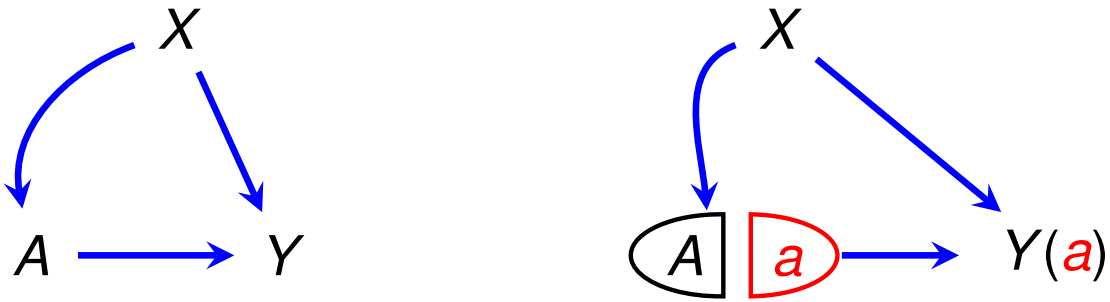
Yes! Applying d-separation to the SWIG on the right we see that there is no d-connecting path from $Y(a, b)$ given $Z(a)$.

Summary Adding Counterfactual Distributions to DAGs

Factorization of counterfactual variables: $P(\boldsymbol{V}(\boldsymbol{a}))$ factorizes with respect to the SWIG $\mathcal{G}[\boldsymbol{a}]$ (ignoring fixed nodes):

$$P\left(\boldsymbol{V}(\boldsymbol{a})\right) = \prod_{Y(\boldsymbol{a}) \in \boldsymbol{V}(\boldsymbol{a})} P\left(Y(\boldsymbol{a}) \middle| \operatorname{pa}_{\mathcal{G}[\boldsymbol{a}]}(Y(\boldsymbol{a})) \setminus \boldsymbol{a}\right).$$

Example



Suppose we want to identify the distribution of $Y(a^*)$ using these two SWIGs, but that we only observe $P(x, a, y)$.

We have that

$$
\begin{aligned}
P(Y(a)) &= \sum_x P(Y(a) \mid X = x) \cdot P(X = x) \\
&= \sum_x P(Y(a) \mid X = x, A = a) \cdot P(X = x) \\
&= \sum_x P(Y \mid X = x, A = a) \cdot P(X = x)
\end{aligned}
$$

using the law of total probability, conditional ignorability and consistency respectively.

Applying d-separation to the graph $\mathcal{G}[\boldsymbol{a}]$

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

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

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

$$\boldsymbol{B}(\tilde{a}) \text{ is d-separated from } \boldsymbol{C}(\tilde{a}) \text{ given } \boldsymbol{D}(\tilde{a}) \text{ in } \mathcal{G}[\tilde{a}] \qquad (*)$$
$$\Rightarrow \quad \boldsymbol{B}(\tilde{a}) \perp\!\!\!\perp \boldsymbol{C}(\tilde{a}) \mid \boldsymbol{D}(\tilde{a}) \quad [P(\boldsymbol{V}(\tilde{a}))].$$

Applying d-separation to the graph $\mathcal{G}[\boldsymbol{a}]$

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

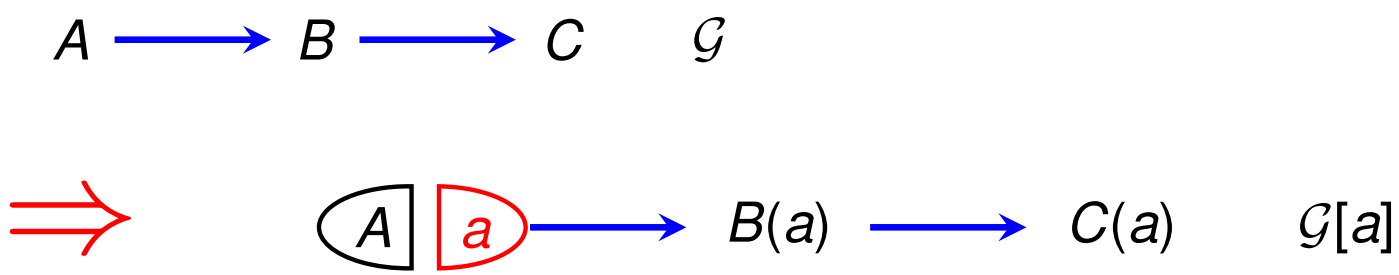
- A fixed node is always blocked if it occurs as a non-endpoint on a path;
- A path on which one endpoint is a fixed node **can d-connect** that node to a random node if it satisfies the usual conditions on colliders and non-colliders.

In $\mathcal{G}[\boldsymbol{a}, d]$, if fixed node d is d-separated from $\boldsymbol{B}(\boldsymbol{a}, d)$ given $\boldsymbol{C}(\boldsymbol{a}, d)$ then

$$P(\boldsymbol{B}(\boldsymbol{a}, d) \mid \boldsymbol{C}(\boldsymbol{a}, d)) = P(\boldsymbol{B}(\boldsymbol{a}, d') \mid \boldsymbol{C}(\boldsymbol{a}, d')).$$

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

Example of d-separation from fixed nodes



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

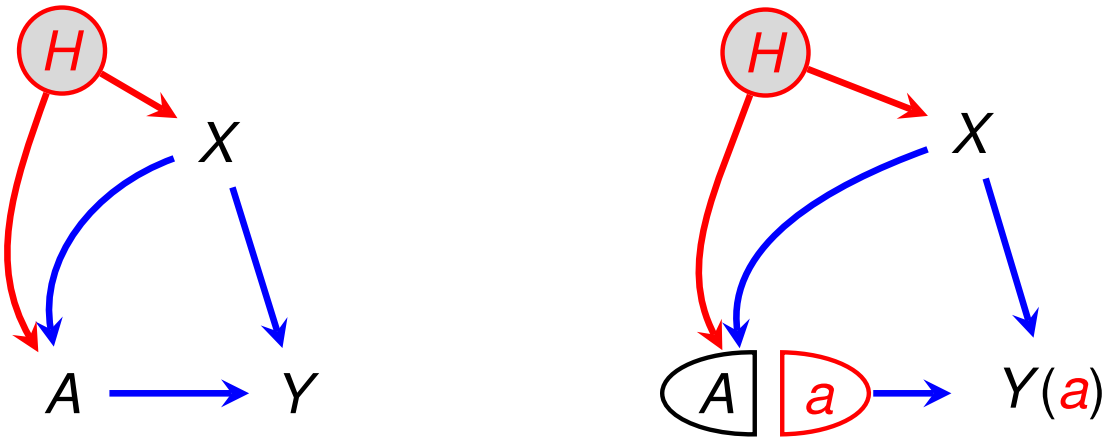
$$P(C(\tilde{a}) \mid B(\tilde{a})) = P(C(a^*) \mid B(a^*))$$

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

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

via consistency and d-separation in $\mathcal{G}[a]$ and $\mathcal{G}$.

Another Example



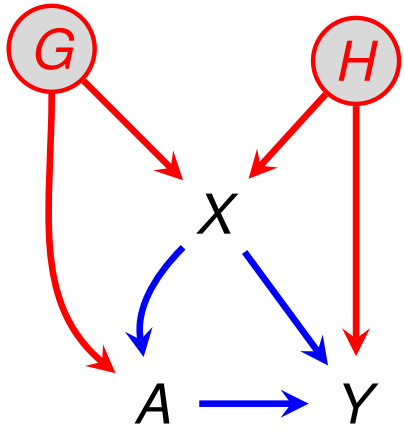
Here again we can read directly from the graph that

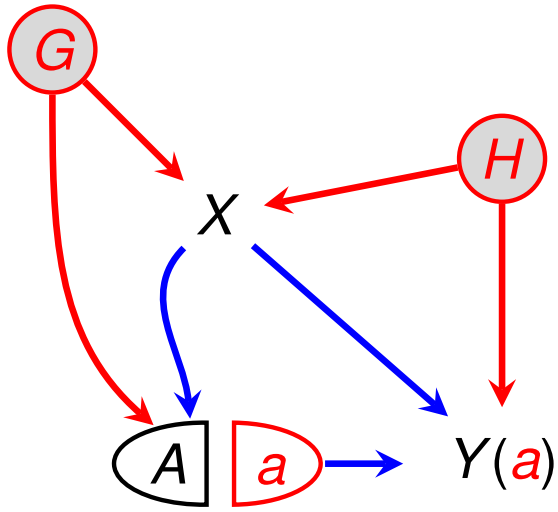
$$A \perp\!\!\!\perp Y(a) \mid X.$$

Hence

$$P(Y(a)) = \sum_x P(X = x) \cdot P(Y \mid A = a, X = x).$$

Exercise





Is it still the case that $A \perp\!\!\!\perp Y(a) \mid X$?

Summary

- SWIGs provide a simple way to unify graphs and counterfactuals via node-splitting
- The approach works via linking the factorizations associated with the SWIG to the distribution in the original DAG.
- The new graph represents a counterfactual distribution that is *identified* from the original joint distribution.
- (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 Tech. Report No. 128* `http://www.csss.washington.edu/Papers/wp128.pdf`, 2013.

Richardson, TS, Robins, JM. SWIGs: A Primer. *UAI-13*, 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.