Single World Intervention Graphs (SWIGs):

Unifying the Counterfactual and Graphical Approaches to Causality

Thomas Richardson

Department of Statistics

University of Washington

Joint work with James Robins (Harvard School of Public Health)

Data, Society and Inference Seminar, Stanford

19 May 2014

Outline

- Review of counterfactuals and graphs
- A new unification of graphs and counterfactuals via node-splitting
  - Simple examples
  - General procedure
  - Factorization and Modularity Properties
  - Contrast with Twin Network approach
- Further Applications:
  - Adjustment for Confounding
  - Sequentially Randomized Experiments / Time Dependent Confounding
  - Dynamic Regimes
- Concluding remarks

The counterfactual framework: philosophy

Hume (1748) *An Enquiry Concerning Human Understanding*:

We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second, ...

...where, if the first object had not been the second never had existed.

Note: this is not one of the 3(!) causal theories Hume is famous for.

The potential outcomes framework: crop trials

Jerzy Neyman (1923):

To compare ν *varieties [on* m *plots] we will consider numbers:*

$$\left.\begin{array}{c}\overbrace{\begin{array}{ccc} U_{11} & \ldots & U_{1m} \\ \vdots & & \vdots \\ U_{\nu 1} & \ldots & U_{\nu m}\end{array}}^{\text{plots}}\end{array}\right\} \text{varieties}$$

U_{ij} is crop yield that would be observed if variety i were planted in plot j.

Physical constraints only allow one variety to be planted in a given plot in any given growing season $\Rightarrow$ Observe only one number per col.

Potential outcomes with binary treatment

For binary treatment X and response Y, we define two potential outcome variables:

- $Y(x = 0)$: the value of Y that *would* be observed for a given unit *if* assigned $X = 0$;
- $Y(x = 1)$: the value of Y that *would* be observed for a given unit *if* assigned $X = 1$;
- *WIll also write these as* $Y(x_0)$ *and* $Y(x_1)$.
- Implicit here is the assumption that these outcomes are well-defined. Specifically:
  - Only one version of treatment $X = x$
  - No interference between units / Stable Unit Treatment Value Assumption (SUTVA)
- Will use 'potential outcome' and 'counterfactual' synonymously.

Drug Response 'Types':

In the simplest case where Y is a binary outcome we have the following 4 types:

$Y(x_0)$	$Y(x_1)$	Name
0	0	Never Recover
0	1	Helped
1	0	Hurt
1	1	Always Recover

Assignment to Treatments

Unit	Potential Outcomes		Observed	
	$Y(x=0)$	$Y(x=1)$	X	Y
1	0	1	1	
2	0	1	0	
3	0	0	1	
4	1	1	1	
5	1	0	0	

Observed Outcomes from Potential Outcomes

Unit	Potential Outcomes		Observed	
	$Y(x=0)$	$Y(x=1)$	X	Y
1	0	1	1	1
2	0	1	0	0
3	0	0	1	0
4	1	1	1	1
5	1	0	0	1

Potential Outcomes and Missing Data

Unit	Potential Outcomes		Observed	
	$Y(x=0)$	$Y(x=1)$	X	Y
1	?	1	1	1
2	0	?	0	0
3	?	0	1	0
4	?	1	1	1
5	1	?	0	1

Average Causal Effect (ACE) of X on Y

$$\begin{aligned} \text{ACE}(X \to Y) &\equiv \mathbb{E}[Y(x_1) - Y(x_0)] \\ &= p(\textit{Helped}) - p(\textit{Hurt}) \qquad \in [-1, 1] \end{aligned}$$

Thus $\text{ACE}(X \to Y)$ is the difference in % recovery if everyone treated ($X = 1$) vs. if noone treated ($X = 0$).

Identification of the ACE under randomization

If X is assigned randomly then

$$X \perp\!\!\!\perp Y(x_0) \quad \text{and} \quad X \perp\!\!\!\perp Y(x_1) \tag{1}$$

hence

$$\begin{aligned} E[Y(x_1) - Y(x_0)] &= E[Y(x_1)] - E[Y(x_0)] \\ &= E[Y(x_1) \mid X = 1] - E[Y(x_0) \mid X = 0] \\ &= E[Y \mid X = 1] - E[Y \mid X = 0]. \end{aligned}$$

Thus if (1) holds then $\mathrm{ACE}(X \to Y)$ is identified from $P(X, Y)$.

Inference for the ACE without randomization

Suppose that we do not know that $X \perp\!\!\!\perp Y(x_0)$ and $X \perp\!\!\!\perp Y(x_1)$.
What can be inferred?

	$X = 0$ Placebo	$X = 1$ Drug
$Y = 0$	200	600
$Y = 1$	800	400

What is:

- The largest number of people who could be *Helped*?
  $400 + 200$
- The smallest number of people who could be *Hurt*? 0

  $\Rightarrow$ Max value of ACE: $(200 + 400)/2000 - 0 = 0.3$

Similar logic:

$\Rightarrow$ Min value of ACE: $0 - (600 + 800)/2000 = -0.7$

In general, bounds on $\text{ACE}(X \rightarrow Y)$ will always cross zero.

Summary of Counterfactual Approach

- In our observed data, for each unit one outcome will be 'actual'; the others will be 'counterfactual'.
- The potential outcome framework allows *Causation* to be 'reduced' to *Missing Data* $\Rightarrow$ Conceptual progress!
- The ACE is identified if $X \perp\!\!\!\perp Y(x_1)$ and $X \perp\!\!\!\perp Y(x_0)$
- Independences implied by Randomization of Treatment.
- Ideas are central to Fisher's Exact Test; also many parts of experimental design.
- The framework is the basis of *many* practical *causal* data analyses published in Biostatistics, Econometrics and Epidemiology.

Relating Counterfactuals and ‘do’ notation

Expressions in terms of ‘do’ can be expressed in terms of counterfactuals:

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

but counterfactual notation is more general.
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.

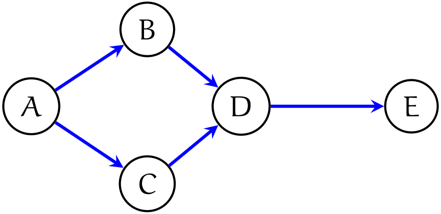
Graphs

Factorization Associated with a DAG

We associate the following factorization of a joint distribution $P(\mathbf{V})$ with a DAG:

$$P(\mathbf{V}) = \prod_{X \in \mathbf{V}} P(X \mid \mathrm{pa}(X))$$

Example:



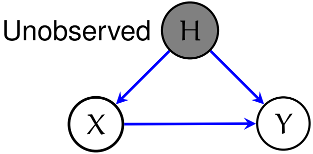
$$\begin{aligned} &P(A, B, C, D, E) \\ &\quad = \quad P(A) \times P(B \mid A) \times P(C \mid A) \times P(D \mid B, C) \times P(E \mid D) \end{aligned}$$

Graphical rule (d-separation) allows independence relations holding in a distribution that factorizes wrt a graph to be 'read' from the graph.

Ex: $\quad C \perp\!\!\!\perp B \mid A \qquad D \perp\!\!\!\perp A \mid B, C \qquad E \perp\!\!\!\perp A, B, C \mid D.$

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 $\mathrm{X} \rightarrow \mathrm{Y}$ 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:

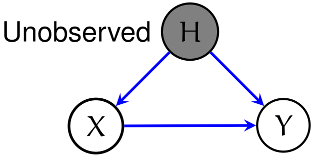
$$P(Y(x) = y) = P(Y = y \mid \text{do}(X = x)) = P(Y = y \mid X = x).$$

  here '$P(y \mid \text{do}(x))$' is defined as the distribution resulting from an intervention (or experiment) where we fix X to x.

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

Linking the two approaches

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

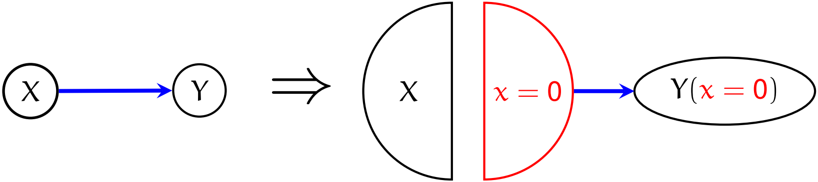


$X \not\!\perp\!\!\!\perp Y(x_0)$ *or* $X \not\!\perp\!\!\!\perp Y(x_1)$

- 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:

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

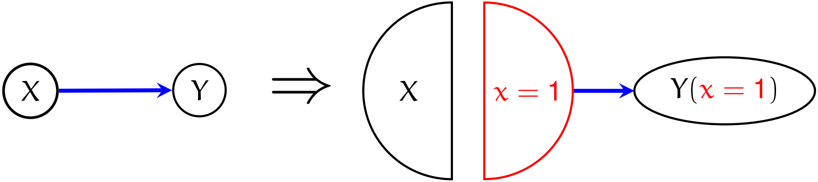
where:

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

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

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 graph.

Although we have:

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

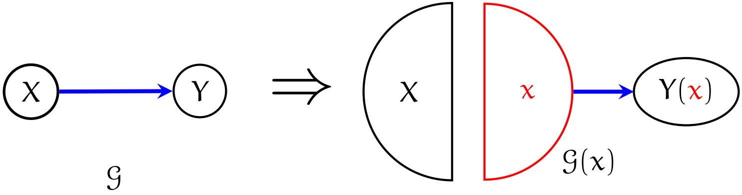
we do not assume

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

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’



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 $P(X, Y(x_1))$.

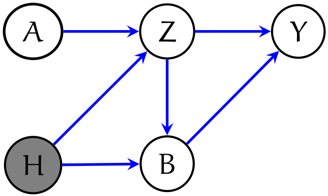
Intuition behind node splitting:

(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 $Y(x = 0)$.

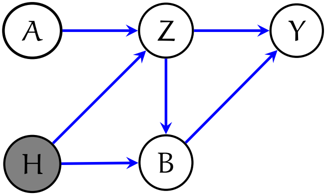
Harder Inferential problem

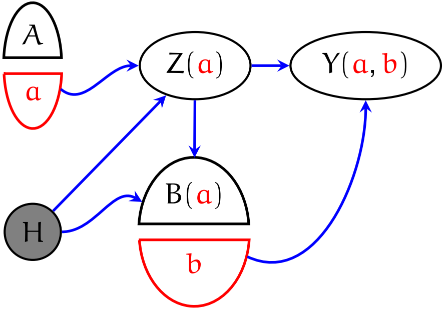


Query: does this causal graph imply:

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

Simple solution





Query does this graph imply:

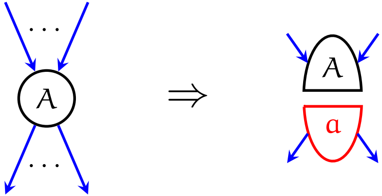
$$Y(a, b) \perp\!\!\!\perp B(a) \mid Z(a), A \qquad ?$$

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 $Z(a)$.
More on this shortly...

Single World Intervention Template Construction (1)

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

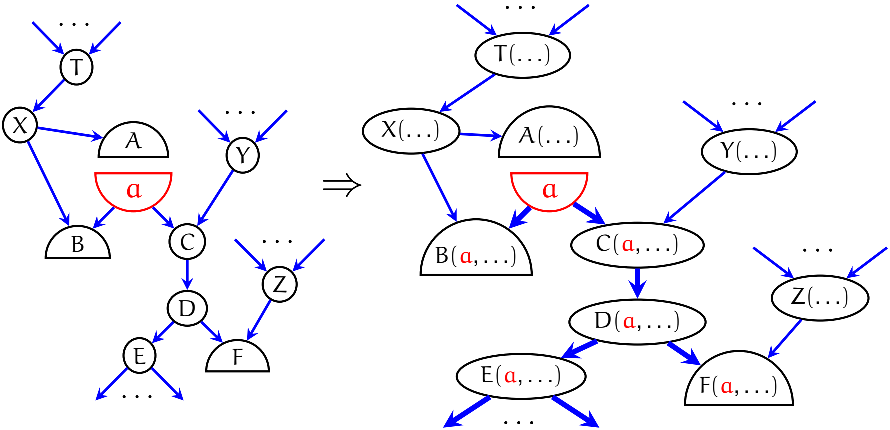


Splitting: Schematic Illustrating the Splitting of Node A

- 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:



Single World Intervention Graph

A Single World Intervention *Graph* (SWIG) $\mathcal{G}(\mathbf{a}^*)$ is obtained from the Template $\mathcal{G}(\mathbf{a})$ by simply substituting specific values $\mathbf{a}^*$ for the variables $\mathbf{a}$ in $\mathcal{G}(\mathbf{a})$;

For example, we replace $\mathcal{G}(x)$ with $\mathcal{G}(x\!=\!0)$.

Resulting SWIG $\mathcal{G}(\tilde{x})$ contains variables $\mathbb{V}(\tilde{x})$ and represents the joint: $P(\mathbb{V}(\tilde{x}))$

Factorization and Modularity

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

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

Modularity: $P(\mathbb{V}(\tilde{\mathbf{a}}))$ and $P(\mathbf{V})$ are linked as follows:
The conditional density associated with $Y(\tilde{\mathbf{a}}_Y)$ 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 $P(\mathbf{V})$ is observed then $P(\mathbb{V}(\tilde{\mathbf{a}}))$ is identified.

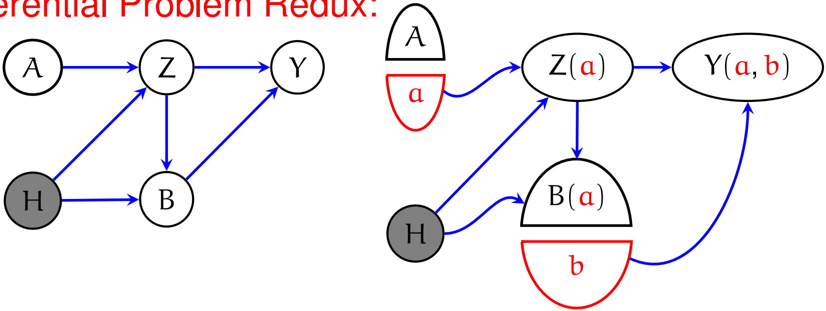
Applying d-separation to the graph $\mathrm{G}(\mathbf{a})$

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

$$\mathbb{B}(\tilde{\mathbf{a}}) \text{ is d-separated from } \mathbb{C}(\tilde{\mathbf{a}}) \text{ given } \mathbb{D}(\tilde{\mathbf{a}}) \cup \tilde{\mathbf{a}} \text{ in } \mathcal{G}(\tilde{\mathbf{a}}) \qquad (2)$$

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

Inferential Problem Redux:



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

$$Y(a,b) \perp\!\!\!\perp B \mid Z, A = a. \tag{3}$$

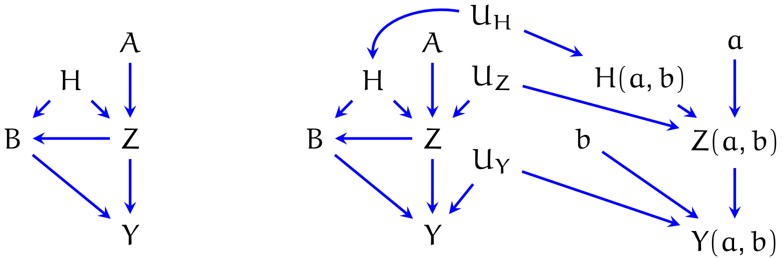
The SWIG shows that (3) does hold; Pearl is incorrect.
Specifically, we see from the SWIG:

$$Y(a,b) \perp\!\!\!\perp B(a) \mid Z(a), A \tag{4}$$

$$\Rightarrow \quad Y(a,b) \perp\!\!\!\perp B(a) \mid Z(a), A = a \tag{5}$$

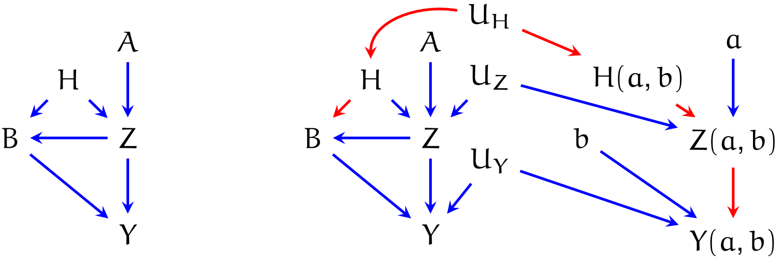
This last condition is then equivalent to (3) via consistency.
(Pearl infers a claim of Robins is false since if true then (3) would hold).

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

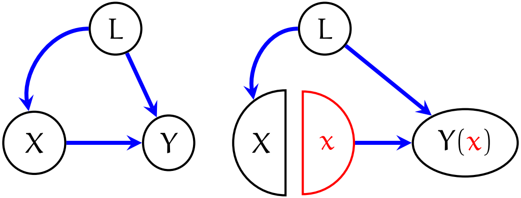


The twin network fails to reveal that $Y(a, b) \perp\!\!\!\perp B \mid Z, A = a$.
This 'extra' independence holds in spite of d-connection because (by consistency) when $A = a$, then $Z = Z(a) = Z(a, b)$.
Note that $Y(a, b) \not\!\perp\!\!\!\perp B \mid Z, A \neq a$.

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

Adjustment for Confounding

Adjusting for confounding



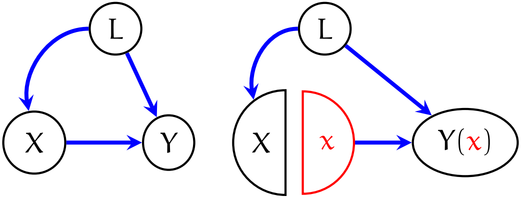
Here we can read directly from the template that

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

It follows that:

$$P(Y(\tilde{x})=y) = \sum_{l} P(Y=y \mid L=l, X=\tilde{x})P(L=l). \qquad (6)$$

Adjusting for confounding



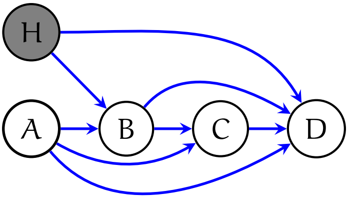
$$X \perp\!\!\!\perp Y(x) \mid L.$$

Proof of identification:

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

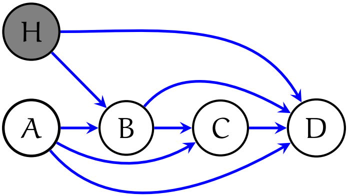
Multiple Treatments

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 $P(D(\tilde{a}, \tilde{c}))$.

Sequentially randomized experiment (I)



If the following holds:

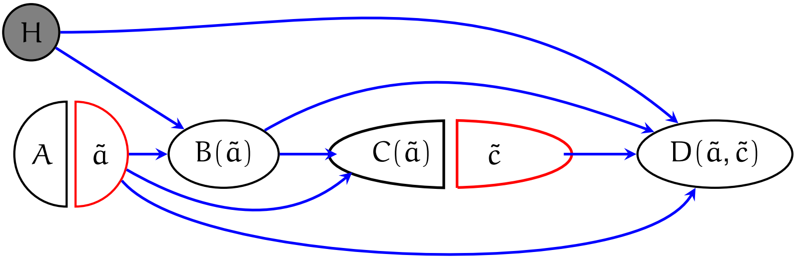
$$\begin{aligned} A &\perp\!\!\!\perp D(\tilde{a}, \tilde{c}) \\ C(\tilde{a}) &\perp\!\!\!\perp D(\tilde{a}, \tilde{c}) \mid B(\tilde{a}), A \end{aligned}$$

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)



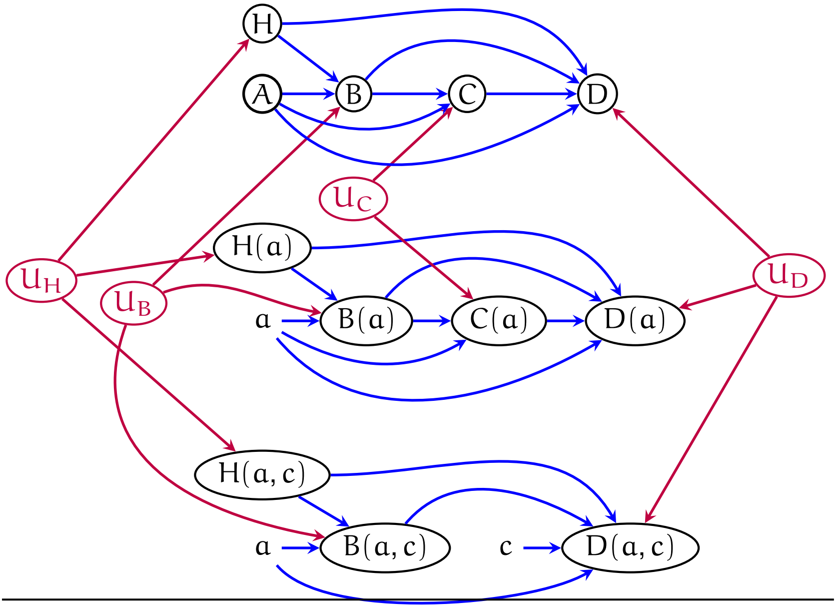
d-separation:

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

General result of Robins (1986) then implies:

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

Multi-network approach



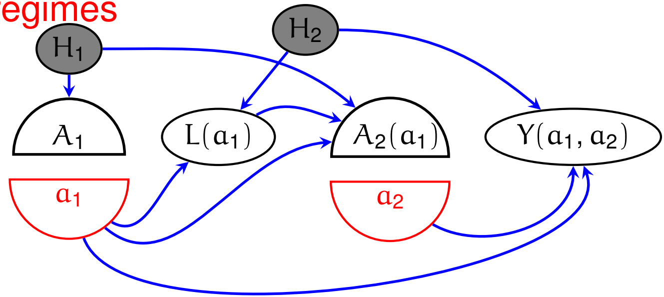
Connection to Pearl's *do*-calculus

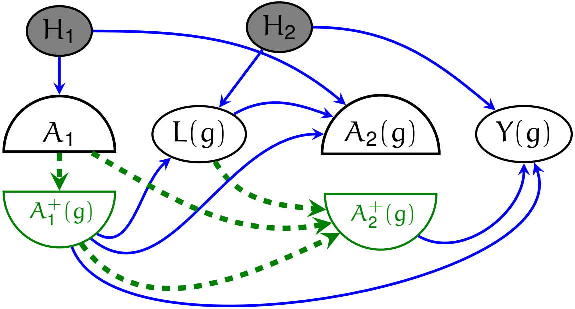
Factorization and modularity are sufficient to imply all of the identification results that hold in the *do*-calculus of Pearl (1995); see also Spirtes *et al.* (1993):

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

Dynamic regimes

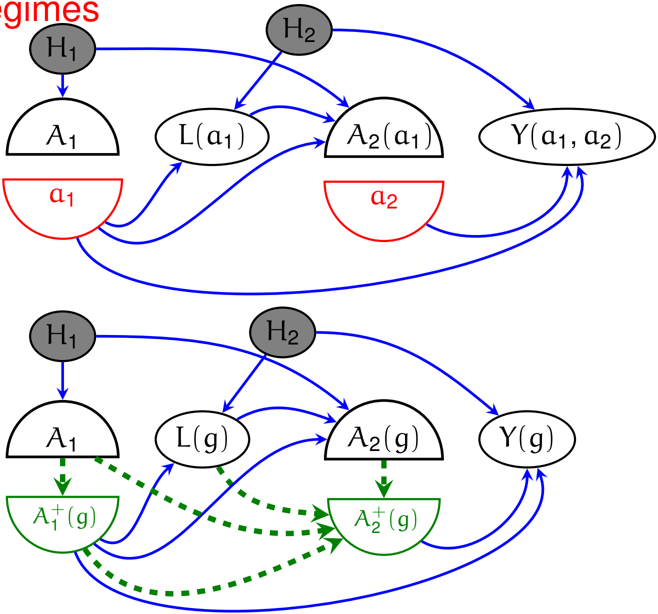
Dynamic regimes





$P(Y(g))$ is identified.

Dynamic regimes



$P(Y(g))$ is not identified.

Conclusion: Eliminating a false trichotomy

Previously the only approach to unifying counterfactuals and graphs was Pearl's approach via Non-Parametric Structural Equation Models with Independent Errors:

This gave causal modelers three options:

- Use graphs, and not counterfactuals (Dawid).
- Use counterfactuals, and not graphs (many Statisticians).
- Use both graphs and counterfactuals, but be forced to make a lot of additional assumptions that are:
  - not experimentally testable (even in principle);
  - not necessary for most identification results.

SWIGs show that one can use graphs and counterfactuals without being forced to take on additional assumptions.

Summary and Extensions

- 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) The approach also provides a way to combine information on the absence of individual and population level direct effects.
- (Not covered) Also allows to formulate models where interventions on only some variables are well-defined.

Thank You!

References

- Pearl, J. *Causality* (Second ed.). Cambridge, UK: Cambridge University Press, 2009.
- 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.
- Shpitser, I, Pearl, J. What counterfactuals can be tested. *Journal of Machine Learning Research* 9, 1941–1979, 2008.
- Spirtes, P, Glymour, C, Scheines R. *Causation, Prediction and Search.* Lecture Notes in Statistics 81, Springer-Verlag.

Assuming Independent Errors and

Cross-World Independence

Relating Counterfactuals and Structural Equations

Potential outcomes can be seen as a different notation for Non-Parametric Structural Equation Models (NPSEMs): Example: $X \rightarrow Y$.

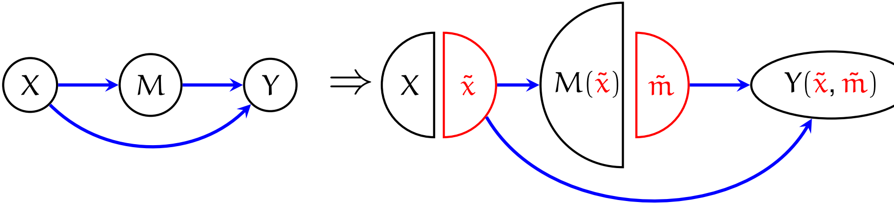
$$\begin{aligned} \text{NPSEM formulation:} \quad Y &= f(X, \epsilon_Y) \\ \text{Potential outcome formulation:} \quad Y(x) &= f(x, \epsilon_Y) \end{aligned}$$

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'.

Mediation graph

Intervention on X and M:



d-separation in the SWIG gives:

$$X \perp\!\!\!\perp M(\tilde{x}) \perp\!\!\!\perp Y(\tilde{x}, \tilde{m}), \quad \text{for } \tilde{x}, \tilde{m} \in \{0, 1\}$$

Pearl associates additional independence relations with this DAG

$$\begin{aligned} Y(x_1, m) &\perp\!\!\!\perp M(x_0), X \\ Y(x_0, m) &\perp\!\!\!\perp M(x_1), X \end{aligned}$$

equivalent to assuming independent errors, $\varepsilon_X \perp\!\!\!\perp \varepsilon_M \perp\!\!\!\perp \varepsilon_Y$.

Pure Direct Effect

Pure (aka Natural) Direct Effect (PDE): *Change in Y had X been different, but M fixed at the value it would have taken had X not been changed:*

$$\mathrm{PDE} \equiv Y(x_1, M(x_0)) - Y(x_0, M(x_0)).$$

Legal motivation [from Pearl (2000)]:

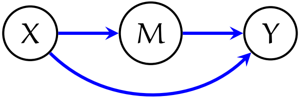
"The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same." (Carson versus Bethlehem Steel Corp., 70 FEP Cases 921, 7th Cir. (1996)).

Decomposition

PDE also allows non-parametric decomposition of Total Effect (ACE) into direct (PDE) and indirect (TIE) pieces.

$$
\begin{aligned}
\mathrm{PDE} &\equiv E\left[Y(1, M(0))\right] - E\left[Y(0)\right] \\
\mathrm{TIE} &\equiv E\left[Y\left(1, M(1)\right) - Y\left(1, M(0)\right)\right] \\
\mathrm{TIE} + \mathrm{PDE} &\equiv E\left[Y(1)\right] - E\left[Y(0)\right] \equiv ACE(X \to Y)
\end{aligned}
$$

Pearl's identification claim



Pearl (2001) shows that under the NPSEM with independent errors associated with the above graph:

the PDE is identified (!) by the following *mediation formula*:

$$\text{PDE}^{\text{med}} = \sum_{m} \left[E[Y|x_1, m] - E[Y|x_0, m] \right] P(m|x_0)$$

Critique of PDE: Hypothetical Case Study

Observational data on three variables:

- X- *treatment*: cigarette cessation
- M *intermediate*: blood pressure at 1 year, high or low
- Y *outcome*: say CHD by 2 years
- Observed data (X, M, Y) on each of n subjects.
- All binary
- X randomly assigned

Hypothetical Study (I): X randomized

		$Y=0$	$Y=1$	Total	$\hat{P}(Y=1 \mid m, x)$
$X=0$	$M=0$	1500	500	2000	0.25
	$M=1$	1200	800	2000	0.40
$X=1$	$M=0$	948	252	1200	0.21
	$M=1$	1568	1232	2800	0.44

A researcher, Prof H wishes to apply the mediation formula to estimate the PDE.

Prof H believes that there is no confounding, so that Pearl's NPSEM-IE holds, but his post-doc, Dr L is skeptical.

Hypothetical Study (II): X and M Randomized

To try to address Dr L's concerns, Prof H carries out animal intervention studies.

		$Y=0$	$Y=1$	Total	$\hat{P}(Y(m,x)=1)$
$X=0$	$M=0$	750	250	1000	0.25
	$M=1$	600	400	1000	0.40
$X=1$	$M=0$	790	210	1000	0.21
	$M=1$	560	440	1000	0.44

As we see: $\hat{P}(Y(m,x)=1) = \hat{P}(Y=1 \mid m,x)$;
Prof H is now convinced: '*What other experiment could I do ?*'

He applies the mediation formula, yielding $\widehat{\mathrm{PDE}}^{\text{med}} = 0$.
Conclusion: No direct effect of X on Y.

Failure of the mediation formula

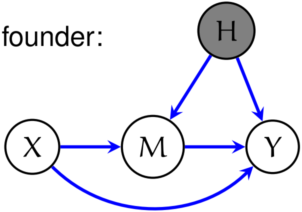
Under the true generating process, the true value of the PDE is:

$$\widehat{\mathrm{PDE}} = 0.153 \neq \widehat{\mathrm{PDE}}^{\text{med}} = 0$$

Prof H's conclusion was completely wrong!

Why did the mediation formula go wrong?

Dr L was right – there was a confounder:

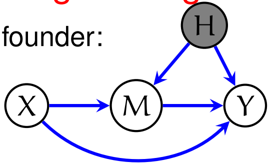


but. . . it had a special structure so that:

$$Y \perp\!\!\!\perp H \mid M, X = 0 \quad \text{and} \quad M \perp\!\!\!\perp H \mid X = 1$$

Why did the mediation formula go wrong?

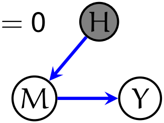
Dr L was right – there was a confounder:



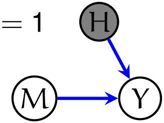
but. . . it had a special structure so that:

$$Y \perp\!\!\!\perp H \mid M, X = 0 \quad \text{and} \quad M \perp\!\!\!\perp H \mid X = 1$$

$X = 0$



$X = 1$



The confounding undetectable by any intervention on X and/or M.

Pearl: *Onus is on the researcher to be sure there is no confounding (= independent errors)*.

Causation should precede intervention.

Summary of critique of Independent Error Assumption

The independent error assumption cannot be checked by any randomized experiment on the variables in the graph.

$\Rightarrow$ Connection between experimental interventions and potential outcomes, established by Neyman has been severed;

$\Rightarrow$ Theories in Social and Medical sciences are not detailed enough to support the independent error assumption.