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)

Therme Vals Causal Workshop
5 Aug 2013
Outline

- Brief review of counterfactuals
- A new unification of graphs and counterfactuals via node-splitting
  - Factorization and Modularity
  - Contrast with Twin Network approach
  - Some Examples and Extensions
  - Sequentially Randomized Experiments / Time Dependent Confounding
  - Dynamic Regimes
- Experimental Testability and Independence of Errors in NPSEM}s
Counterfactuals
aka Potential Outcomes
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.*
The potential outcomes framework: crop trials

Jerzy Neyman (1923):

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

\[
\begin{array}{cccc}
u_{11} & \ldots & u_{1m} \\
\vdots & & \vdots \\
u_{v1} & \ldots & u_{vm}
\end{array}
\]

Here \( u_{ij} \) is the 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.

*Popularized by Rubin (1974); sometimes called the ‘Rubin causal model’.*
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 (SUTVA).
## Potential Outcomes

<table>
<thead>
<tr>
<th>Unit</th>
<th>Potential Outcomes</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Y(x = 0)$</td>
<td>$Y(x = 1)$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Drug Response ‘Types’:

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

<table>
<thead>
<tr>
<th>$Y(x_0)$</th>
<th>$Y(x_1)$</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>Never Recover</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>Helped</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>Hurt</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Always Recover</td>
</tr>
</tbody>
</table>
## Assignment to Treatments

<table>
<thead>
<tr>
<th>Unit</th>
<th>Potential Outcomes</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Y(x = 0)$</td>
<td>$Y(x = 1)$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
### Observed Outcomes from Potential Outcomes

<table>
<thead>
<tr>
<th>Unit</th>
<th>Potential Outcomes</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Y(x = 0)$</td>
<td>$Y(x = 1)$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
### Potential Outcomes and Missing Data

<table>
<thead>
<tr>
<th>Unit</th>
<th>Potential Outcomes</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Y(x = 0)$</td>
<td>$Y(x = 1)$</td>
</tr>
<tr>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>?</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>?</td>
</tr>
</tbody>
</table>
Average Causal Effect (ACE) of $X$ on $Y$

$$\text{ACE}(X \rightarrow Y) \equiv \mathbb{E}[Y(x_1) - Y(x_0)]$$

$$= p(\text{Helped}) - p(\text{Hurt}) \quad \in [-1, 1]$$

Thus ACE($X \rightarrow Y$) 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

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

hence

$$E[Y(x_1) - Y(x_0)] = E[Y(x_1)] - E[Y(x_0)]$$

$$= E[Y(x_1) | X = 1] - E[Y(x_0) | X = 0]$$

$$= E[Y | X = 1] - E[Y | X = 0].$$

Thus if (1) holds then $\text{ACE}(X \rightarrow Y)$ is identified from $P(X, Y)$. 
Inference for the ACE without randomization

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

<table>
<thead>
<tr>
<th></th>
<th>$X = 0$</th>
<th>$X = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td>200</td>
<td>600</td>
</tr>
<tr>
<td>Drug</td>
<td>800</td>
<td>400</td>
</tr>
</tbody>
</table>

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: $0 - (600 + 800)/2000 = -0.7$
Inference for the ACE without randomization

Suppose that we do not know that $X \perp \perp Y(x_0)$ and $X \perp \perp Y(x_1)$.

General case:

$$-(P(x=0, y=1) + P(x=1, y=0)) \leq ACE(X \rightarrow Y)$$

$$ACE(X \rightarrow Y) \leq P(x=0, y=0) + P(x=1, y=1)$$

$\Rightarrow$ Bounds will always cross zero.

$\Rightarrow X \perp \perp Y(x_0)$ and $X \perp \perp Y(x_1)$ essential for drawing non-trivial causal inferences.
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* ⇒ Conceptual progress!
- The ACE is identified if $X \perp \perp Y(x_i)$ for all values $x_i$.
- Randomization of treatment assignment implies $X \perp \perp Y(x_i)$.
- 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.
Potential outcomes can be seen as a different notation for Non-Parametric Structural Equation Models (NPSEMs):

Example: $X \rightarrow Y$.

NPSEM formulation: $Y = f(X, \epsilon_Y)$

Potential outcome formulation: $Y(x) = f(x, \epsilon_Y)$

Two important caveats:

- NPSEMs typically assume all variables are seen as being subject to well-defined interventions (not so with potential outcomes)
- Pearl associates NPSEMs with Independent Errors (NPSEM-IEs) with DAGs (more on this later).
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 \textit{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
Recap: Graphical Approach to Causality

- 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 $X$. 
Edges are directed, but are they causal?

No Confounding

\[ P(X, Y) = P(X)P(Y \mid X) \]

No Confounding

\[ P(X, Y) = P(Y)P(X \mid Y) \]

- Neither factorization places any restriction on \( P(X, Y) \).
Linking the two approaches

\[ X \perp \!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\!
\![/\text{Unobserved}]
\text{H} \]
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: $X \independent Y(x = 0)$.
Also associate a new factorization:

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

where:

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

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

Thomas Richardson
Therme Vals Workshop
Slide 22
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:

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

where:

$P(Y(x = 1) = y) = P(Y = y \mid X = 1)$. 
Crucial point: $Y(x=0)$ 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 have

\[ X \perp \perp Y(x=0), Y(x=1) \]

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

\[ \Rightarrow \textbf{Single-World Intervention Graphs (SWIGs).} \]
Represent both graphs via a ‘template’

Represent both graphs via a \textit{template}:

Formally this is a ‘graph valued function’:

\begin{itemize}
\item Takes as input a specific value $x^*$
\item Returns as output a SWIG $g(x^*)$.
\end{itemize}

Each \textit{instantiation} of the template is a SWIG $g(x^*)$ that represents a different margin: $P(X, Y(x^*))$ with red nodes $x^*$ becoming constants.
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?

\[ Y(x_0, x_1) \perp\!\!\!\!\!\!\!\!\!\!\perp X_1(x_0) \mid Z(x_0), X_0, \]
Simple solution

Query does this graph imply:

\[ Y(x_0, x_1) \perp \perp X_1(x_0) \mid Z(x_0), X_0 \]

Answer: Yes – applying d-separation to the SWIG on the right we see that there is no d-connecting path from \( Y(x_0, x_1) \) given \( Z(x_0) \).

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

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

*Splitting:* Schematic Illustrating the Splitting of Node $A$
Single World Intervention Template Construction (2)

(2) Relabel descendants of fixed nodes:
A Single World Intervention Graph (SWIG) $G(a^*)$ is obtained from the Template $G(a)$ by simply substituting specific values $a^*$ for the variables $a$ in $G(a)$;

For example, we replace $G(x)$ with $G(x=0)$.

- Changing the value of a fixed variable corresponds to constructing a new graph and considering a different population, e.g. $P(X, Y(x=0))$ vs. $P(X, Y(x=1))$

- It is only the instantiated graph $G(\tilde{x})$ that represents $P(\forall(\tilde{x}))$, not the template $G(x)$.
Factorization and Modularity

**Factorization:** \( P(\mathcal{V}(\tilde{a})) \) over the counterfactual variables in \( \mathcal{G}(\tilde{a}) \) factorizes with respect to \( \mathcal{G}(\tilde{a}) \) (ignoring fixed nodes):

\[
P(\mathcal{V}(\tilde{a})) = \prod_{Y(\tilde{a}) \in \mathcal{V}(\tilde{a})} P(Y(\tilde{a}) \mid \text{pa}_{\mathcal{G}(\tilde{a})}(Y(\tilde{a})) \setminus \tilde{a}).
\]

**Modularity:** \( P(\mathcal{V}(\tilde{a})) \) and \( P(\mathcal{V}) \) are linked as follows:

\[
P(Y(\tilde{a}) = y \mid \left( \text{pa}_{\mathcal{G}(\tilde{a})}(Y(\tilde{a})) \setminus \tilde{a} \right) = q)
\]
\[
= P(Y = y \mid \left( \text{pa}_{\mathcal{G}}(Y) \setminus A \right) = q, \ (\text{pa}_{\mathcal{G}}(Y) \cap A) = \tilde{a}_{\text{pa}_{\mathcal{G}}(Y) \cap A},
\]

So the conditional density associated with \( Y(\tilde{a}_Y) \) in \( \mathcal{G}(\tilde{a}) \) is just the conditional density associated with \( Y \) in \( \mathcal{G} \) after substituting \( \tilde{a}_i \) for any \( A_i \in A \) that is a parent of \( Y \).
Applying d-separation to the graph $G(a)$

Counterfactual conditional independence relations may be obtained from the transformed graph by applying d-separation after adding fixed nodes to the conditioning set:

Given disjoint subsets $B(\tilde{a}), C(\tilde{a})$ and $D(\tilde{a})$ of random vertices (where $D(\tilde{a})$ may be empty),

if $B(\tilde{a})$ is d-separated from $C(\tilde{a})$ given $D(\tilde{a}) \cup \tilde{a}$ in $G(\tilde{a})$ \hspace{1cm} (2)

then $B(\tilde{a}) \perp \perp C(\tilde{a}) \mid D(\tilde{a}) \quad [P(V(\tilde{a}))]$

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

- intuitive since these are fixed constants in the SWIG
- Since vertices in $\tilde{a}$ have no parents, no new paths d-connect due to also conditioning on $\tilde{a}$.
  - $\Rightarrow$ If a d-separation holds in $G(\tilde{a})$ without conditioning on the fixed nodes, then it will continue to hold if we also condition on fixed nodes.
- An alternative is simply to restrict attention to paths that do not contain fixed vertices,
  - e.g. remove fixed nodes from the graph before checking d-separation.
Mediation graph (I)

Intervention on $Z$ alone.

\[
\begin{aligned}
Z & \quad \rightarrow \quad X & \quad \rightarrow \quad Y \\
\Rightarrow \quad Z & \quad \rightarrow \quad X(\tilde{z}) & \quad \rightarrow \quad Y(\tilde{z})
\end{aligned}
\]

factorization:

\[
P(Z, X(\tilde{z}), Y(\tilde{z})) = P(Z)P(X(\tilde{z}))P(Y(\tilde{z}) \mid X(\tilde{z}))
\]

modularity:

\[
\begin{aligned}
P(X(\tilde{z})=x) &= P(X=x \mid Z=\tilde{z}), \\
P(Y(\tilde{z})=y \mid X(\tilde{z})=x) &= P(Y=y \mid X=x, Z=\tilde{z}).
\end{aligned}
\]

d-separation gives:

\[
Z \perp \perp X(\tilde{z}), Y(\tilde{z}).
\]
Mediation graph (II)

Intervention on $Z$ and $X$:

\[
\begin{align*}
Z & \rightarrow X \rightarrow Y \\
& \Rightarrow \quad Z \xrightarrow[\hat{z}]{\hat{z}} X(\hat{z}) \xrightarrow[\hat{x}]{\hat{x}} Y(\hat{x}, \hat{z})
\end{align*}
\]

factorization:

\[
P(Z, X(\hat{z}), Y(\hat{x}, \hat{z})) = P(Z)P(X(\hat{z}))P(Y(\hat{x}, \hat{z}))
\]

modularity:

\[
\begin{align*}
P(X(\hat{z}) = x) &= P(X = x \mid Z = \hat{z}), \\
P(Y(\hat{x}, \hat{z}) = y) &= P(Y = y \mid X = \hat{x}, Z = \hat{z}).
\end{align*}
\]

d-separation gives:

\[
Z \perp \perp X(\hat{z}) \perp \perp Y(\hat{x}, \hat{z})
\]
No direct effect graph

\[
\begin{align*}
\text{factorization:} & \quad P(Z, X(\tilde{z}), Y(\tilde{x})) = P(Z)P(X(\tilde{z}))P(Y(\tilde{x})) \\
\text{modularity:} & \quad P(X(\tilde{z}) = x) = P(X=x \mid Z=\tilde{z}), \\
& \quad P(Y(\tilde{x}) = y) = P(Y=y \mid X=\tilde{x}). \\
\text{d-separation gives:} & \quad Z \perp \perp X(\tilde{z}) \perp \perp Y(\tilde{x})
\end{align*}
\]
Pearl (2009), Ex. 11.3.3, claims the causal DAG above does not imply:

\[ Y(x_0, x_1) \perp \perp X_1 \mid Z, X_0 = x_0. \]  

(3)

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

\[ Y(x_0, x_1) \perp \perp X_1(x_0) \mid Z(x_0), X_0, \]  

(4)

\[ \Rightarrow Y(x_0, x_1) \perp \perp X_1(x_0) \mid Z(x_0), X_0 = x_0. \]  

(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 $Y(x_0, x_1) \perp \perp X_1 \mid Z, X_0 = x_0$. This ‘extra’ independence holds in spite of d-connection because (by consistency) when $X_0 = x_0$, then $Z = Z(x_0) = Z(x_0, x_1)$. Note that $Y(x_0, x_1) \not\perp \perp X_1 \mid Z, X_0 \neq x_0$.

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

Here we can read directly from the template that $X \not\perp \!
\!\!\!\!\!\!\perp Y(\chi)$ since there is a path:

$$X \leftarrow H \rightarrow Y(\chi).$$
Adjusting for confounding

Here we can read directly from the template that

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

It follows that:

\[
P(Y(\tilde{x}) = y) = \sum_l P(Y = y \mid L = l, X = \tilde{x})P(L = l). \tag{6}
\]
Contrast with approach via removing edges

This ‘explains’ why $L$ is sufficient to control confounding under the null (where $X$ has no effect on $Y$) but not under the alternative.
Adjusting for confounding

\[ X \perp Y(\tilde{x}) \mid L. \]

Proof of identification:

\[
P[Y(\tilde{x}) = y] = \sum_{l} P[Y(\tilde{x}) = y \mid L = l]P(L = l)
\]

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

\[
= \sum_{l} P[Y = y \mid L = l, X = \tilde{x}]P(L = l) \text{ modularity}
\]
More Examples (I)

Here we can read directly from the template that

\[ X \perp \perp Y(\chi) | L. \]
More Examples (II)

Here we can read directly from the template that

\[ X \perp \! \! \! \perp Y(\chi) \mid L. \]
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 $P(D(\tilde{a}, \tilde{c}))$.
If the following holds:

\[ A \perp \perp D(\tilde{a}, \tilde{c}) \]
\[ C(\tilde{a}) \perp D(\tilde{a}, \tilde{c}) \mid B(\tilde{a}), A \]

General result of Robins (1986) then implies:

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

Does it??
Sequentially randomized experiment (II)

\[ \tilde{a} \]

\[ \tilde{c} \]

\[ \tilde{a}, \tilde{c} \]

\[ H \]

\[ A \]

\[ B(\tilde{a}) \]

\[ C(\tilde{a}) \]

\[ D(\tilde{a}, \tilde{c}) \]

**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

Thomas Richardson  Therme Vals Workshop  Slide 49
Another example

A ⊥ ⊥ D(ã, ċ)
C(ã) ⊥ ⊥ D(ã, ċ) | B(ã), A

General result of Robins (1986) then implies:

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

Does it??
Another example

General result of Robins (1986) then implies:

\[
P(D(\tilde{a}, \tilde{c}) = d) = \sum_b P(B = b \mid A = \tilde{a})P(D = d \mid A = \tilde{a}, B = b, C = \tilde{c}).
\]
General result (Robins, 1986)

Observed data:

\[ O \equiv \langle L_1, A_1, \ldots, L_K, A_K, Y \rangle. \]

If the following holds for \( k = 1, \ldots, K \)

\[ Y(a^\dagger) \perp \perp A_k(a^\dagger) | \overline{L}_k(a^\dagger), \overline{A}_{k-1}(a^\dagger); \quad (7) \]

then (under positivity):

\[ P(Y(a^\dagger) = y | \overline{L}_j(a^\dagger) = \overline{l}_j, \overline{A}_{j-1}(a^\dagger) = \overline{a}_{j-1}^\dagger) = \sum_{l_{m+1}, \ldots, l_K} p(y | \overline{l}_K, \overline{a}_K^\dagger) \prod_{j=m+1}^{K} p(l_j | \overline{l}_{j-1}, \overline{a}_{j-1}^\dagger). \quad (8) \]

Here \( \overline{A}_{j-1}(a^\dagger) \equiv \langle A_1, \ldots, A_{j-1}(a_{j-2}^\dagger) \rangle \), similarly for \( \overline{L}_{j-1}(a^\dagger) \).

The RHS of (8) is referred to as the ‘g-formula’. 
A dynamic regime $g$ is a policy that assigns treatment (usually at multiple time points) on the basis of past history;

- Including conditional on the ‘natural’ value of treatment in the absence of an intervention;

*Exercise for as long as you would have done without intervention or twenty minutes, whichever is more.*

See Young *et al.* (2012) for additional analysis.
Dynamic regimes

\[ P(Y(g)) \] is identified.
Dynamic regimes

\[ H_1 \rightarrow A_1 \leftarrow L(a_1) \rightarrow A_2(a_1) \rightarrow Y(a_1, a_2) \]

\[ H_1 \rightarrow A_1 \leftarrow L(g) \rightarrow A_2(g) \rightarrow Y(g) \]

\[ P(Y(g)) \] is not identified.
Joint Independence

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

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

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.
Assuming Independent Errors and Cross-World Independence
d-separation gives:

\[ X \perp \perp M(\tilde{x}) \perp \perp Y(\tilde{x}, \tilde{m}) \]

Pearl associates additional independence relations with this graph

\[ Y(x_1, m) \perp \perp M(x_0), X \]
\[ Y(x_0, m) \perp \perp M(x_1), X \]

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:

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

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.

$$\text{PDE} \equiv E[Y(1, M(0))] - E[Y(0)]$$

$$\text{TIE} \equiv E[Y(1, M(1)) - Y(1, M(0))]$$

$$\text{TIE + PDE} \equiv E[Y(1)] - E[Y(0)] \equiv \text{ACE}(X \rightarrow Y)$$
Pearl’s identification claim

Pearl and others claim that under “no confounding” the PDE is identified by the following mediation formula:

\[ \text{PDE}^{\text{med}}(m) = \sum_m [E[Y|x_1, m] - E[Y|x_0, m]] P(m|x_0) \]
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
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.

<table>
<thead>
<tr>
<th></th>
<th>$Y = 0$</th>
<th>$Y = 1$</th>
<th>Total</th>
<th>$\hat{P}(Y(m, x) = 1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = 0$</td>
<td>$M = 0$</td>
<td>750</td>
<td>250</td>
<td>1000</td>
</tr>
<tr>
<td></td>
<td>$M = 1$</td>
<td>600</td>
<td>400</td>
<td>1000</td>
</tr>
<tr>
<td>$X = 1$</td>
<td>$M = 0$</td>
<td>790</td>
<td>210</td>
<td>1000</td>
</tr>
<tr>
<td></td>
<td>$M = 1$</td>
<td>560</td>
<td>440</td>
<td>1000</td>
</tr>
</tbody>
</table>

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

He applies the mediation formula, yielding $\hat{PDE}^{med} = 0$.
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:

\[ \hat{PDE} = 0.153 \neq \hat{PDE}_\text{med} = 0 \]

Prof H’s conclusion was completely wrong!
Why did the mediation formula go wrong?

Dr L was right – there was a confounder:

\[ \text{but} \ldots \text{it had a special structure so that:} \]

\[ Y \perp H \mid M, X = 0 \quad \text{and} \quad M \perp H \mid X = 1 \]
Why did the mediation formula go wrong?

Dr L was right – there was a confounder:

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

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

The confounding **undetectable** by any intervention on \( X \) and/or \( M \).

Pearl: *Onus is on the researcher to be sure there is no confounding.*

*Causation should precede intervention.*
PDE identification cannot be checked via experiment

- If our only interventions are on the variables $X$ and $M$ then we cannot do an experiment to learn the PDE.

- We could learn $E[Y_{x=1, M(x=0)}]$ by intervention if we could
  - intervene and set $X$ to 0 and observe $M(0)$,
  - then return each subject to their pre-intervention state,
  - finally intervene to set $X$ to 1 and $M$ to $M(0)$ and observe $Y(1, M(0))$.

- Such an intervention strategy will usually not exist because not possible in a real-world intervention (e.g., suppose the outcome $Y$ were death).

- Because we cannot observe the same subject under both $X = 1$ and $X = 0$ (i.e. "across worlds"), no intervention will allow us to learn the distribution of mixed counterfactuals such as $Y_{x=1, M(x=0)}$:

(In the story Dr L had to introduce a new node on the graph in order to check the value of the PDE via an experiment.)
Summary of critique of Independent Error Assumption

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

⇒ Connection between experimental interventions and potential outcomes, established by Neyman has been severed;

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

What about faithfulness and causal discovery procedures?

- Such inferences are explicit that they rely on faithfulness, and are designed to guide hypothesis formation;
  - Contrast: In Pearl’s NPSEM-IE approach the simple act of using a DAG is viewed as automatically committing you to making this untestable hypothesis.

- Predictions (possibly derived assuming faithfulness) regarding intervention distributions $P(Y(x)) = P(Y|do(x))$ can be tested by randomized experiments.
How many experimentally untestable assumptions?

Assumption of independent errors implies super-exponentially many ‘cross-world’ counterfactual independence assumptions:

<table>
<thead>
<tr>
<th>No. Actual Vars.</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dim. $P(V)$</td>
<td>3</td>
<td>7</td>
<td>15</td>
<td>$2^K - 1$</td>
</tr>
<tr>
<td>No. Counterfactual Vars.</td>
<td>3</td>
<td>7</td>
<td>15</td>
<td>$2^K - 1$</td>
</tr>
<tr>
<td>Dim. Counterfactual Dist.</td>
<td>7</td>
<td>127</td>
<td>32767</td>
<td>$2^{(2^K-1)} - 1$</td>
</tr>
<tr>
<td>Dim. SWIG</td>
<td>5</td>
<td>113</td>
<td>32697</td>
<td>$(2^{(2^K-1)} - 1) - \sum_{j=1}^{K-1} (4^j - 2^j)$</td>
</tr>
<tr>
<td>Dim. NPSEM-IE</td>
<td>4</td>
<td>19</td>
<td>274</td>
<td>$\sum_{j=0}^{K-1} (2^{2^j} - 1)$</td>
</tr>
<tr>
<td>No. untestable indep. constrnts in NPSEM-IE</td>
<td>1</td>
<td>94</td>
<td>32423</td>
<td>$O(2^{2^K-2})$</td>
</tr>
</tbody>
</table>

Table: Dimensions of counterfactual models associated with complete graphs with binary variables.
In an NPSEM we define a counterfactual independence to be *logical* if it holds regardless of the distribution over counterfactuals (equivalently error terms) e.g. for binary $X$:

$$Y(x_0) \perp \perp Y(x_1) \mid X, Y$$

*Completeness Conjecture* There exists a distribution over counterfactuals that is experimentally indistinguishable from the NPSEM that assumes independent errors but in which the only non-logical independencies are those that may be derived from the SWIG.
A simple approach to unifying graphs and counterfactuals via node-splitting

The approach works via linking the factorizations associated with the two graphs

The approach provides a language that allows counterfactual and graphical people to communicate

The approach leads to many fewer untestable independence assumptions than in the NPSEM-IE approach of Pearl.

The approach also provides a way to combine information on the absence of individual and population level direct effects.
Thank You!
References


Details on Pearl’s Error

Pearl correctly states that using his Twin Network method (next slide) it may be shown that

\[ Y(x_0, x_1) \text{ is not independent of } X_1, \text{ given } Z \text{ and } X_0. \]

However, he then goes on to say (incorrectly):

*In the twin network model there is a d-connected path from \( X_1 \) to \( Y(x_0, x_1) \)... Therefore, [(3)] is not satisfied for \( Y(x_0, x_1) \) and \( X_1 \).*

[Ex. 11.3.3, p.353]

This is actually incorrect in two ways:

- \( Y(x_0, x_1) \perp\!\!\!\!\perp X_1 \mid Z, X_0 \) does not imply \( Y(x_0, x_1) \not\perp\!\!\!\!\perp X_1 \mid Z, X_0 = x_0 \)

- d-separation is not complete for Twin Networks so the presence of a d-connected path does not imply that an independence is not implied.
T and Y(z) are d-connected given X in the twin-network, but in spite of this T \perp \perp Y(z) \mid X under the associated NPSEM-IE because X(z) = X, and T and Y(z) are d-separated given X in the twin-network.
Mediation graph (I)

Intervention on $Z$ alone.

\[
\begin{array}{c}
Z 
\end{array} \rightarrow
\begin{array}{c}
X 
\end{array} \rightarrow
\begin{array}{c}
Y 
\end{array} \Rightarrow
\begin{array}{c}
Z \quad \tilde{Z} 
\end{array} \rightarrow
\begin{array}{c}
X(\tilde{Z}) 
\end{array} \rightarrow
\begin{array}{c}
Y(\tilde{Z}) 
\end{array}
\]

factorization:

\[
P(Z, X(\tilde{Z}), Y(\tilde{Z})) = P(Z)P(X(\tilde{Z}))P(Y(\tilde{Z}) | X(\tilde{Z}))
\]

modularity:

\[
P(X(\tilde{Z}) = x) = P(X = x | Z = \tilde{Z}),
\]
\[
P(Y(\tilde{Z}) = y | X(\tilde{Z}) = x) = P(Y = y | X = x, Z = \tilde{Z}).
\]

d-separation gives:

\[
Z \perp \perp X(\tilde{Z}), Y(\tilde{Z}).
\]
Mediation graph (II)

Intervention on \( Z \) and \( X \):

\[
\begin{align*}
\text{factorization:} & \quad P(Z, X(\tilde{z}), Y(\tilde{x}, \tilde{z})) = P(Z)P(X(\tilde{z}))P(Y(\tilde{x}, \tilde{z})) \\
\text{modularity:} & \quad P(X(\tilde{z})=x) = P(X=x \mid Z=\tilde{z}), \\
& \quad P(Y(\tilde{x}, \tilde{z})=y) = P(Y=y \mid X=\tilde{x}, Z=\tilde{z}).
\end{align*}
\]

d-separation gives:

\[
Z \perp \perp X(\tilde{z}) \perp \perp Y(\tilde{x}, \tilde{z})
\]
Importance of fixed nodes

Compare:

\[ P(Z, X(\tilde{z}), Y(\tilde{z})) = P(Z)P(X(\tilde{z}))P(Y(\tilde{z}) \mid X(\tilde{z})) \]
\[ P(Y(\tilde{z}) = y \mid X(\tilde{z}) = x) = P(Y = y \mid X = x, Z = \tilde{z}). \]

versus

\[ P(Z, X(\tilde{z}), Y(\tilde{z})) = P(Z)P(X(\tilde{z}))P(Y(\tilde{z}) \mid X(\tilde{z})), \]
\[ P(Y(\tilde{z}) = y \mid X(\tilde{z}) = x) = P(Y = y \mid X = x) \]
Importance of fixed nodes: leaving them out causes problems!

\[ P(Z, X(\tilde{z}), Y(\tilde{z})) = P(Z)P(X(\tilde{z}))P(Y(\tilde{z}) \mid X(\tilde{z})) \]
\[ P(Y(\tilde{z}) = y \mid X(\tilde{z}) = x) = P(Y = y \mid X = x, Z = \tilde{z}). \]

versus

\[ P(Z, X(\tilde{z}), Y(\tilde{z})) = P(Z)P(X(\tilde{z}))P(Y(\tilde{z}) \mid X(\tilde{z})) \]
\[ P(Y(\tilde{z}) = y \mid X(\tilde{z}) = x) = P(Y = y \mid X = x) \]

Red nodes are needed in order to read off modularity property from \( G(\tilde{a}) \).
No direct effect graph (I)

\[ Z \rightarrow X \rightarrow Y \quad \Rightarrow \quad Z \rightarrow \tilde{Z} \rightarrow X(\tilde{Z}) \rightarrow Y(\tilde{Z}) \]

factorization:

\[ P(Z, X(\tilde{Z}), Y(\tilde{Z})) = P(Z)P(X(\tilde{Z}))P(Y(\tilde{Z}) \mid X(\tilde{Z})) \]

modularity:

\[ P(X(\tilde{Z}) = x) = P(X = x \mid Z = \tilde{Z}), \]
\[ P(Y(\tilde{Z}) = y \mid X(\tilde{Z}) = x) = P(Y = y \mid X = x). \]

d-separation gives:

\[ Z \perp \perp X(\tilde{Z}), Y(\tilde{Z}) \]