Single World Intervention Graphs: A Primer

Thomas S. Richardson
Department of Statistics
University of Washington
thomasr@u.washington.edu

James M. Robins
Department of Epidemiology
Harvard School of Public Health
robins@hsph.harvard.edu

Abstract

We present a simple graphical theory unifying causal directed acyclic graphs (DAGs) and potential (aka counterfactual) outcomes via a node-splitting transformation. We introduce a new graph, the Single-World Intervention Graph (SWIG). The SWIG encodes the counterfactual independences associated with a specific hypothetical intervention on the set of treatment variables. The nodes on the SWIG are the corresponding counterfactual random variables. We illustrate the theory with a number of examples. Our graphical theory of SWIGs may be used to infer the counterfactual independence relations that hold among the SWIG variables under the FFRCISTG model of Robins (1986) and the NPSEM model with Independent Errors of Pearl (2000, 2009). Furthermore, in the absence of hidden variables, the joint distribution of the counterfactuals is identified; the identifying formula is the extended g-computation formula introduced in (Robins et al., 2004). As an illustration of the benefit of reasoning with SWIGs, we use SWIGs to correct an error regarding Example 11.3.3 presented in (Pearl, 2009).

1 Introduction

Potential outcomes are extensively used within Statistics, Political Science, Economics, and Epidemiology for reasoning about causation. Directed acyclic graphs (DAGs) are another formalism used to represent causal systems also extensively used in Computer Science, Bioinformatics, Sociology and Epidemiology. Given the long history and utility of both approaches – as demonstrated by many applications – it is natural to wish to unify them.

A graphical unification of existing causal models

We present a simple approach to this synthesis based on an intuitive graphical transformation: by 'splitting' treatment nodes in a causal DAG over the actual variables, we form a new graph, the Single-World Intervention Graph (SWIG). The SWIG encodes the counterfactual independences associated with a specific hypothetical intervention on the set of treatment variables. The nodes on the SWIG are the corresponding counterfactual random variables.

The factorization and Markov properties encoded in the structure of the SWIG imply and are implied by the extended g-formula of Robins et al. (2004). These two properties are satisfied by all previously proposed counterfactual causal models, including the Finest Fully Randomized Causally Interpretable Structured Tree Graphs (FFRCISTG) of Robins (1986), the Pseudo-Indeterministic Systems of Spirtes et al. (1993), the Non-Parametric Structural Equation Models with Independent Errors[1] (NPSEM-IE) consid-

[1] In (Pearl, 2000, 2009; Robins and Richardson, 2011) the acronym 'NPSEM' is used to refer to what is here termed an NPSEM-IE. We wish to emphasize here that

ered in Pearl (2000) and the Minimal Counterfactual Model (MCM) of Robins and Richardson (2011).[2] As a consequence any (counterfactual) independences or causal identification results obtained in our theory apply to the above.

Our graphical approach is as follows: given a causal DAG over the actual variables we construct a set of Single-World Intervention Graphs (SWIGs). Since the node set consists of the set of counterfactual variables corresponding to a single hypothetical intervention on a set of (possibly time varying) treatments no two SWIGs (constructed from the same initial DAG but for different interventions) will have identical node sets.

Furthermore, if the factuals on the DAG have a positive distribution, so $P(\mathbf{V}=\mathbf{v}) > 0$ for all $\mathbf{v}$ then a SWIG will not contain random variables that are related deterministically. As a consequence the graphical criterion (d-separation) for checking conditional independence among counterfactual variables (on a given SWIG) is complete. In other words, our SWIG encodes all of those independence relations (among the variables present in the SWIG) that hold for all distributions over counterfactuals in the (FFRCISTG or NPSEM-IE) model. If a counterfactual independence relation among the variables in the SWIG is not implied then there is some distribution that is in the model for which the corresponding dependence holds.

Our approach differs from previously proposed attempts to link graphs and counterfactuals such as the 'twin-network' approach of Balke and Pearl (1994), generalized to 'multi-networks' in Pearl (2009), and the 'counterfactual graph' of Shpitser and Pearl (2007, 2008). d-separation is not complete for twin-networks since they include more variables amongst which there are deterministic relations.[3] d-separation applied to the 'counterfactual graphs' introduced in (Shpitser and Pearl, 2007, 2008) is conjectured (Shpitser, 2013) to be complete for independence among *events* (or equivalently indicators $I(V=v)$). However, to use this to check for independence among *variables* requires the construction of an exponential number of counterfactual graphs. There is currently no known polynomial-time algorithm for testing independence among counterfactual variables under the NPSEM-IE. It should be noted that 'multi-networks' and 'counterfactual graphs' are designed to address a harder problem than SWIGs since their goal is to determine *all* independencies implied by an NPSEM-IE model including 'cross-world' independencies.

As an illustration of the utility of SWIGs we show that Pearl, one of the creators of the twin-network method, draws an erroneous conclusion concerning the validity of Robins' g-computation method possibly due to assuming that the presence of a d-connecting path implies lack of (context specific) independence. To successfully apply the twin-network method one must be careful to distinguish independence from context specific independence. We shall see that such is not the case if one uses SWIGS, allowing one to straightforwardly uncover Pearl's error.

Moreover the SWIG allows one to write down a factorization of the joint distribution of the counterfactuals on the graph in addition to allowing one to read off counterfactual independencies via d-separation (Pearl, 1988). Furthermore under an FFRCISTG the distribution of the counterfactuals is linked to that of the factuals through a property we refer to as 'modularity'. In particular, modularity and factorization imply the joint distribution of the counterfactuals is given by the extended g-formula. Other advantages of SWIGs include the following:

- The SWIG gives a graphical explanation as to why conditioning on variables so as to

FFRCISTGs may also be explicitly defined via a system of structural equations (with possibly dependent errors – though any such dependence is undetectable via randomized experiments). Hence Pearl's NPSEM-IE model is a strict sub-model of the FFRCISTG model. We have thus opted to refine Pearl's notation to make clear that it is solely the additional (untestable) assumptions regarding independence of the errors that distinguish the NPSEM-IE and the FFRCISTG approaches. FFRCISTGs as defined in Robins (1986) did not require that all variables could be intervened on. Thus, to be precise, the FFRCISTG models referred to in the main text of this paper and in (Robins and Richardson, 2011) are those in which all variables are subject to intervention. It is these FFRCISTGs that may be defined via a system of structural equations.

[2] Strictly MCMs (Robins and Richardson, 2011) only obey the resulting properties in the case where all intervention variables are binary.

[3] Furthermore SWIGs do not correspond to induced subgraphs of twin-networks.

‘block all back-door paths’ provides a consistent estimate of the causal effect of a variable X on Y, *both* under the null hypothesis of no causal effect, *and* under the alternative.

In (Richardson and Robins, 2013) we show the following:

- A simple modification of a SWIG allows one to encode on a single graph (and thus distinguish) the two possible causal interpretations of missing arrows: an absence of a causal effect for each individual versus the absence of an average causal effect at the population level.

- The SWIG permits the criteria for evaluation of treatment regimes or plans involving k different treatments to be checked by inspecting a single graph, whereas previous criteria (Pearl and Robins, 1995), though equivalent, require the construction and inspection of a series of k different ‘mutilated’ graphs; similar comments apply to the application of multi-networks (Pearl, 2009).

- The approach naturally extends to (possibly random) dynamic regimes where treatment is assigned (either deterministically or stochastically) on the basis of prior covariates, including the level of treatment that a patient would choose (in the absence of it being specified by the regime).

Removing a false trichotomy

A primary aim of our approach is to show that researchers in causality are *not* forced to choose between:

- Using causal graphs without counterfactuals;

- Using counterfactuals without graphs;

- Combining graphs and counterfactuals via the NPSEM-IE framework, as advocated by Pearl, thereby imposing many counterfactual independence assumptions that are not, even in principle, testable.

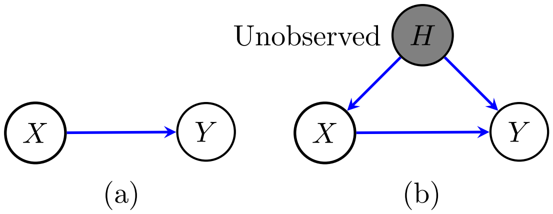


Figure 1: (a) A causal DAG representing two unconfounded variables; (b) A causal DAG representing the presence of confounding.

We believe that at least some of the motivation for using graphs without counterfactuals and vice-versa has been the misperception that to combine the two approaches necessitates the adoption of the NPSEM-IE approach and its strong assumptions that are, for many purposes, unnecessary.

In the next section we show via simple examples how to construct a SWIG and how to use it to reason with counterfactuals.

2 Motivating Examples

To motivate our development we first consider the simple graphs, shown in Figure 1. The nodes represent random variables and the graph represents a factorization of their joint density. Specifically, the DAG in Figure 1(a) is associated with the (trivial) factorization:

$$p(x, y) = p(x)p(y \mid x) \tag{1}$$

where the densities on the RHS are associated, respectively, with X and Y in the DAG.

DAGs are often given a causal interpretation. In that case the DAG in Figure 1(a) is interpreted as representing the fact that the effect of X on Y is unconfounded. (In contrast, on the DAG in Figure 1(b) the effect of X on Y is confounded by the common cause H.) Within the potential outcomes (or counterfactual) literature the absence of confounding is understood as implying (at least) the ‘weak ignorability’ conditions:

$$X \perp\!\!\!\perp Y(x = 0) \quad \text{and} \quad X \perp\!\!\!\perp Y(x = 1), \tag{2}$$

where we have supposed that X is a binary treatment variable, and that the potential outcomes

$Y(x=0)$ and $Y(x=1)$ are well-defined. Here, for example $Y(x=0)$ denotes the value of Y had, possibly contrary to fact, X been set to 0.

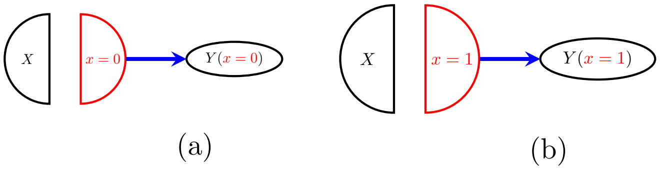
One of the primary uses of graphs, including DAGs, is to represent the conditional independence (or Markov) structure of a multivariate distribution via d-separation. Since (2) is an independence statement, one might naively think that this could be read directly from the graph in Figure 1(a). However, the absence of the variables $Y(x=0)$ and $Y(x=1)$ in the DAG in Figure 1(a) would appear to present a significant obstacle to reading the independencies (2) from this graph (!)

The node-splitting transformation

In the approach described here we address this by introducing a simple 'node splitting' operation. This is a generalization of the operation introduced in (Robins et al., 2006) to provide a graphical representation of the Effect of Treatment on the Treated; see also Evans (2012); Geneletti and Dawid (2007); Shpitser and Pearl (2009).

Applying this operation to vertex X in the DAG in Figure 1(a) results in the graphs in Figure 2, which we term Single-World Intervention Graphs (SWIGs). If the hypothetical intervention sets X to 0 then we obtain the SWIG $\mathcal{G}(x\!=\!0)$ shown in Figure 2(a), while setting X to 1 gives the SWIG $\mathcal{G}(x\!=\!1)$ in Figure 2(b). Notice that in addition to splitting the X node, the node corresponding to Y in the original DAG has been relabeled to indicate that it is now a potential outcome.[4]

By applying d-separation to the graph in Figure 2(a), we directly obtain that $X \perp\!\!\!\perp Y(x_0)$, since there are no edges emanating from the node containing X, hence there are no paths from X to $Y(x_0)$.[5] Similarly, by applying d-separation to the graph in Figure 2(b) we derive $X \perp\!\!\!\perp Y(x_1)$.



(a)

(b)

Figure 2: The single world intervention graphs (SWIGs) resulting from splitting node X in the graph in Figure 1(a), and intervening to set a particular value. (a) the SWIG $\mathcal{G}(x\!=\!0)$ corresponding to setting X to 0; (b) $\mathcal{G}(x\!=\!1)$ given by setting X to 1.

The factorization and modularity properties

In the same manner that the original DAG is associated with the joint distribution $P(X,Y)$ we associate the graphs in Figure 2 (a) and (b) with the joint distributions $P(X,Y(x_0))$ and $P(X,Y(x_1))$ respectively. Likewise, we will associate the following factorizations with these graphs:[6]

$$\begin{aligned} P(X\!=\!x, Y(x_0)\!=\!y) &= P(X\!=\!x)P(Y(x_0)\!=\!y), \\ P(X\!=\!x, Y(x_1)\!=\!y) &= P(X\!=\!x)P(Y(x_1)\!=\!y). \end{aligned} \tag{3}$$

In addition we associate the following equation

$$P(Y(x_0)\!=\!y) = P(Y\!=\!y \,|\, X\!=\!0) \text{ for all } y \tag{4}$$

with the graphical transformation from $\mathcal{G} \mapsto \mathcal{G}(x=0)$, and likewise

$$P(Y(x_1)\!=\!y) = P(Y\!=\!y \,|\, X\!=\!1) \text{ for all } y. \tag{5}$$

with the transformation from $\mathcal{G} \mapsto \mathcal{G}(x=1)$. We refer to these equalities as *modularity*[7] conditions linking the distribution of the actual variables in the DAG to the counterfactual variables in the SWIG.

Notice that these equations assert that the marginal distribution of Y resulting from an intervention in which everyone receives the value

[4] In Figure 2 all black nodes should be viewed as nodes in an ordinary DAG model (regardless of their shape). The semi-circular shape of the nodes containing X merely serves to remind us that this graph was derived by splitting X. The red nodes are constants that take on a fixed value. The primary role of these red nodes is to aid in linking the distribution of the variables after splitting to the terms in the factorization associated with the graph prior to splitting.

[5] Here we use x_i as a shorthand for $x\!=\!i$.

[6] Notice that, if we ignore the red nodes, these factorizations are simply instances of the standard DAG factorization, since in Figure 2(a) neither X nor $Y(x_0)$ have any parents (besides the red nodes).

[7] Our usage of the term modularity differs from that of Pearl, though they both derive from the same intuition.

$x = 0$ is the same as the corresponding conditional probability $P(y \mid X = 0)$, and likewise for $x = 1$.

Given the factorization (3), the modularity property follows directly from the consistency condition: $X = x$ implies $Y(x) = Y$. For example,

$$\begin{aligned} P(Y(x_0) = y) &= P(Y(x_0) = y \mid X = 0) \qquad (6) \\ &= P(Y = y \mid X = 0); \end{aligned}$$

here the first equality uses the factorization, while the second follows from consistency.

All NPSEM models satisfy consistency. In fact we will show that the factorization and modularity properties associated with a SWIG hold for an NPSEM associated with the original DAG when the errors have the independence structure specified by an FFRCISTG model, and thus for its more restrictive NPSEM-IE submodel.

The factorization and modularity properties are important because they are sufficient for deriving many identifiability results. To give a simple example, these properties allow us to identify the Effect of Treatment on the Treated (ETT):

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

Here the second equality follows from the factorizations with respect to the two graphs in Figure 2, while the third follows from modularity, i.e. (4) and (5).[8]

Single-worlds vs. multiple-worlds

The reader will notice that although we have constructed SWIGs representing the two single world distributions $P(X, Y(x_0))$ and $P(X, Y(x_1))$ we have not constructed a graph that includes *both* $Y(x_0)$ *and* $Y(x_1)$, and thus represents the joint distribution $P(X, Y(x_0), Y(x_1))$. At first sight this might strike the reader as odd, perhaps even an oversight. In fact, this is by design: in general, observed data, including that resulting from randomized experiments, only identifies the marginal single-world counterfactual distributions for which we construct graphs. It is worth noting that the independence restrictions that we encode may place inequality restrictions on the (multiple-world) joint distribution (e.g.$P(X, Y(x_0), Y(x_1))$) over all counterfactuals.[9]

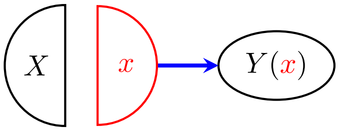


Figure 3: A template representing the two graphs in Figure 2.

As noted, the counterfactual independencies encoded in a SWIG are implied by the FFRCISTG as well as the NPSEM-IE adopted by Pearl (2000, 2009). However, the NPSEM-IE also encodes many additional cross-world restrictions on the joint distribution over all counterfactuals.

Templates

Since it is somewhat redundant to construct a different graph for every value to which we might set x, we may instead represent all such graphs via a 'template', such as shown in Figure 3. However, we note that in any instantiation of this template x should be viewed as taking a specific value: whereas the (black) random nodes in the graph vary across units in the (counterfactual) population being represented, the (red) fixed nodes take the same value. Also note that the value taken by red nodes such as x specify which particular random variables are represented by the random nodes in the template, i.e. whether $Y(x)$ represents $Y(0)$ or $Y(1)$.[10] We refer to these as '*S*ingle *W*orld *I*ntervention *T*emplates' or 'SWITs'.

[8] When X takes more than two states, the factorization and modularity assumptions associated with this graph imply that $ETT(x) \equiv E[Y(x) - Y(0)|X = x]$ equals $E[Y|X = x] - E[Y|X = 0]$.

[9] There can exist extreme distributions for which some of the inequality constraints become equalities (Pearl, 2000, 2009, §8.2).

[10] In this respect the graph differs from standard graphical models, including the conditional acyclic directed mixed graphs (CADMGs) introduced in Shpitser et al. (2011). Though CADMGs include fixed nodes, in a CADMG these nodes do not determine which other variables appear on the graph. In other words, CADMGs are not templates.

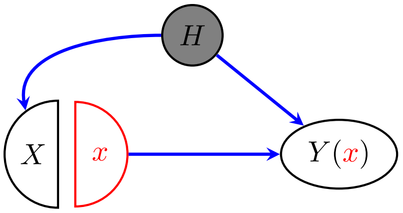


Figure 4: The template resulting from intervening on X in the graph in Figure 1(b).

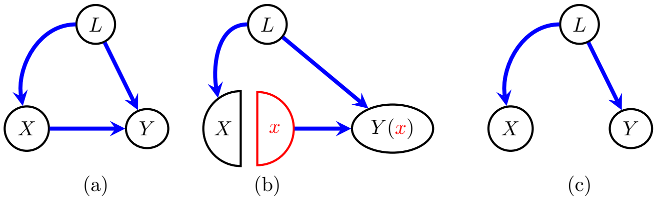


Figure 5: Adjusting for confounding. (a) The original causal graph. (b) The template $\mathcal{G}(x)$, which shows that $Y(x) \perp\!\!\!\perp X \mid L$. (c) The DAG $\mathcal{G}_{\underline{X}}$ obtained by removing edges from X advocated in Pearl (1995, 2000, 2009) to check his 'backdoor condition'.

A new graphical view of the back-door formula

In Figure 4 we show the template representing the graphs resulting from intervening on X in the graph in Figure 1(b), which intuitively represents the presence of confounding. In the potential outcomes literature, confounding is expressed as non-independence of $Y(\tilde{x})$ and X for some $\tilde{x}$. This lack of independence is consistent with $Y(x)$ and X being d-connected in the template shown in Figure 4 by the path $X \leftarrow H \rightarrow Y(x)$.

In contrast, Figure 5(a) shows a DAG in which L is observed, and is sufficient to control confounding between X and Y. From the template in Figure 5(b) we see that

$$X \perp\!\!\!\perp Y(\tilde{x}) \mid L, \tag{7}$$

often referred to as conditional ignorability, holds.[11] It is well known that this condition is sufficient for the effect of X on Y to be given via the standard adjustment formula:

$$P(Y(\tilde{x})\!=\!y) = \sum_{l} P(Y\!=\!y \mid L\!=\!l, X\!=\!\tilde{x})P(L\!=\!l). \tag{8}$$

Two further examples of graphs which imply $X \perp\!\!\!\perp Y(\tilde{x}) \mid L$ are shown in Figure 6; in these graphs H represents a hidden variable.

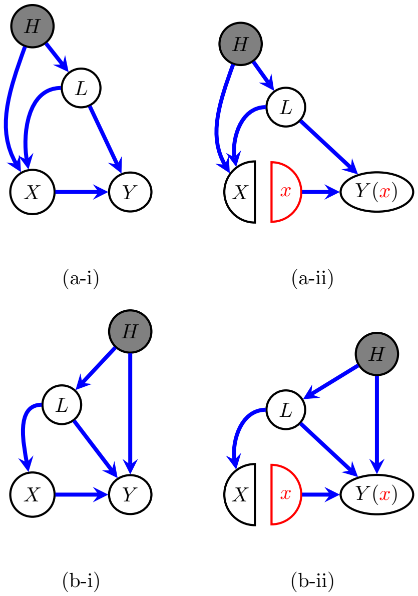
Notice that here we are able to use the graph $\mathcal{G}(\tilde{x})$ to represent the distribution $P(Y(\tilde{x}), X, L)$ in the general case where X has an effect on Y. We contrast this line of graphical reasoning with that advocated in (Pearl, 2000, 2009, p.87) in which d-separation of X and Y given L is checked in the graph $\mathcal{G}_{\underline{X}}$ obtained by removing the edges that are directed out of X; see Figure 5(c). When L is a non-descendant of X this graphical criterion is equivalent to ours, so that X is d-separated from Y given L in $\mathcal{G}_{\underline{X}}$ if and only if X is d-separated from $Y(\tilde{x})$ given L in $\mathcal{G}(\tilde{x})$; hence validity of his criterion is not at issue. However, the graph $\mathcal{G}_{\underline{X}}$ only represents the null hypothesis that X does not causally affect Y. It is only under this null hypothesis that $X \perp\!\!\!\perp Y \mid L$, corresponding to the d-separation of X and Y given L that holds in Figure 5(c). Thus the graph $\mathcal{G}_{\underline{X}}$ does not appear to offer an explanation as to why d-separation of X and Y given L in $\mathcal{G}_{\underline{X}}$ should ensure that (8) holds (even though it does) when X has an effect on Y.

Furthermore, in the general case where we are considering whether we may use the natural extension of (8) to a set of variables $\mathbf{L}$:

$$P(Y(\tilde{x})\!=\!y) = \sum_{\mathbf{l}} P(Y\!=\!y \mid \mathbf{L}\!=\!\mathbf{l}, X\!=\!\tilde{x})P(\mathbf{L}\!=\!\mathbf{l}),$$

the backdoor criterion (Pearl, 2000, 2009, p.79) requires that in addition to X and Y being d-separated given $\mathbf{L}$ in $\mathcal{G}_{\underline{X}}$, no variable in $\mathbf{L}$ may be a descendant of X. The reason for this additional

[11] For binary treatment this condition states that within levels of L, the treated and untreated are comparable, i.e. had, contrary to fact, the two groups received identical treatment, the distributions of responses would have been the same. In the experience of one of the co-authors, in both medical and epidemiologic practice, there are contexts where one may have strong suspicion that comparability does not hold even though subject-matter knowledge does not permit one to specify a full causal graph. As one example it may be unclear whether lack of comparability is due to confounding by an unmeasured common cause or also due to selection, such as M-bias; see (Hernán et al., 2004, p.621-2) for related discussion. In such cases one would take (7), rather than a graph, as a primitive.



(a-i) (a-ii)

(b-i) (b-ii)

Figure 6: Further examples of adjusting for confounding. (a-i) A graph $\mathcal{G}$; (a-ii) the template $\mathcal{G}(x)$; (b-i) A graph $\mathcal{G}'$; (b-ii) the template $\mathcal{G}'(x)$. H is an unobserved variable in $\mathcal{G}$ and $\mathcal{G}'$. Both SWITs imply $Y(x) \perp\!\!\!\perp X \mid L$.

condition is not transparent, since the inclusion of such a variable does not preclude that X and Y may be d-separated in $\mathcal{G}_{\underline{X}}$.[12] For example, consider the DAG $\mathcal{G}$ and corresponding $\mathcal{G}_{\underline{X}}$ shown in Figure 7(a) and (c). Within the framework given here there is no need to state this additional restriction. Using SWIGs we may formulate a simple adjustment criterion as follows:

Counterfactual Adjustment Criterion
If $X \perp\!\!\!\perp Y(\tilde{x}) \mid \mathbf{L}$ is implied by the SWIG $\mathcal{G}(\tilde{x})$, then

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

Notice that we have no need of any restrictions on the membership in $\mathbf{L}$ as is the case with the formulation of the backdoor criterion (Pearl, 2000, p.70). The reason why is illustrated in Figure 7. In the causal graph shown in Figure 7(a), L_1 is necessary and sufficient to control confounding, but $\{L_1, L_2\}$ is not. It may be seen directly from inspecting the template in Figure 7(b) that

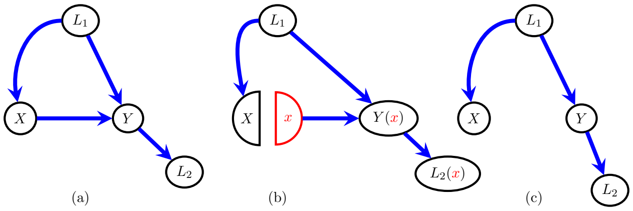


Figure 7: Simplification of the backdoor criterion. (a) The original causal graph $\mathcal{G}$. (b) The template $\mathcal{G}(x)$, which shows that $Y(x) \perp\!\!\!\perp X \mid L_1$, but does not imply $Y(x) \perp\!\!\!\perp X \mid \{L_1, L_2\}$ when there exists an arrow from X to Y, i.e. the null hypothesis is false. (c) The DAG $\mathcal{G}_{\underline{X}}$ obtained by removing edges from X advocated in Pearl (2000, 2009).

$$X \perp\!\!\!\perp Y(\tilde{x}) \mid L_1, \qquad X \perp\!\!\!\perp Y(\tilde{x}) \mid L_1, L_2(\tilde{x})$$

but the template does not imply $X \perp\!\!\!\perp Y(\tilde{x}) \mid L_1, L_2$.[13]

In contrast, under the non-counterfactual formulation of the back-door criterion (Pearl, 2000, 2009, p.78), in which the graph $\mathcal{G}_{\underline{X}}$ is formed as in Figure 7(c) an additional condition must be added, requiring that no member of $\mathbf{L}$ is a descendant of X. This extra condition is required because, as noted earlier, $\mathcal{G}_{\underline{X}}$ represents the null hypothesis of no effect of X on Y. In spite of these differences of approach we emphasize that our criterion holds if and only if Pearl's backdoor criterion holds.[14]

3 Construction of the Single-world Counterfactual Template

The SWIT $\mathcal{G}(\mathbf{a})$ resulting from intervening to set the variables in $\mathbf{A}$ to $\mathbf{a}$ in a directed acyclic graph

[12] Pearl (2009, §11.3, pp.338–344) acknowledges that the need to restrict to non-descendants is not transparent in his original derivation, and offers several alternatives.

[13] As expected, we can construct a distribution under the FFRCISTG model, and in fact in the NPSEM-IE model under which $X \perp\!\!\!\perp Y(\tilde{x}) \mid L_1, L_2$ is not true. That this independence is not implied by an NPSEM-IE associated with the graph in Figure 7(a) could also be deduced by constructing a counterfactual graph (Shpitser and Pearl, 2007, 2008) to test $X \perp\!\!\!\perp Y(\tilde{x}) \mid L_1\!=\!l_1, L_2\!=\!l_2$.

[14] Textor and Liskiewicz (2011) show that every minimal covariate adjustment set satisfies the back-door condition; see also (Shpitser et al., 2010). We conjecture that for identifying conditional effects given the adjustment set, $P(Y(x)|z)$, this criterion is complete under the FFRCISTG model, though interestingly, not the NPSEM-IE.

$\mathcal{G}$ with vertex set $\mathbf{V}$ is constructed in two steps as follows:

(1) *Split Nodes*: For every $A \in \mathbf{A}$ split the node into a random and fixed component, labelled A and a respectively.

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

Let the resulting graph be $\mathcal{G}^*$. For each random vertex V in $\mathcal{G}^*$, let $\mathbf{a}_V$ denote the subset of fixed vertices that are ancestors of V in $\mathcal{G}^*$.

(2) *Labeling:* For every random node V in $\mathcal{G}^*$, label it with $V(\mathbf{a}_V)$.

It is implicit here that if $\mathbf{a}_V = \emptyset$ then $V(\mathbf{a}_V) = V$. The resulting graph is the SWIT $\mathcal{G}(\mathbf{a})$. Let $\mathbb{V}(\mathbf{a}) \equiv \{V(\mathbf{a}_V) \mid V \in \mathbf{V}\}$ be the set of random vertices in $\mathcal{G}(\mathbf{a})$.

Note that by convention we will use $\mathbf{a}_V$ to denote the set of fixed nodes labeling the counterfactual node corresponding to V in a SWIT $\mathcal{G}(\mathbf{a})$.[15]

An *instantiation* $\mathcal{G}(\tilde{\mathbf{a}})$ of $\mathcal{G}(\mathbf{a})$ results from choosing a specific assignment of values $\tilde{\mathbf{a}}$ for the 'free variables' $\mathbf{a}$ in $\mathcal{G}(\mathbf{a})$, and appropriately replacing each occurrence of a_i with $\tilde{a}_i$ within the label for a vertex. Let $\mathfrak{A}$ denote the set of all possible instantiations of $\mathbf{a}$. Formally a template $\mathcal{G}(\mathbf{a})$ may be viewed as a graph valued function defined on the domain $\mathfrak{A}$. From this perspective $\tilde{\mathbf{a}}$ represents a specific input, and $\mathcal{G}(\tilde{\mathbf{a}})$ the resulting output.

4 Modularity and Factorization

We now describe in greater detail the properties of factorization and modularity.

Given a set of treatment variables $\mathbf{A} \subseteq \mathbf{V}$, let $\mathbb{V}(\tilde{\mathbf{a}})$ represent the set of counterfactual variables (corresponding to the actual variables $\mathbf{V}$) associated with a specific hypothetical intervention setting $\mathbf{A}$ to $\tilde{\mathbf{a}}$. The resulting counterfactual distribution $P(\mathbb{V}(\tilde{\mathbf{a}}))$ is obtained from an NPSEM by simply replacing each variable $A_i \in \mathbf{A}$ by the value assigned $\tilde{a}_i$ in the function f_V for any variable V of which A_i is a parent in $\mathcal{G}$. The distribution $P(\mathbb{V}(\tilde{\mathbf{a}}))$ of the counterfactuals $\mathbb{V}(\tilde{\mathbf{a}})$ that are vertices in the SWIG denoted $\mathcal{G}(\tilde{\mathbf{a}})$ under the NPSEM satisfies two important properties: 'factorization' and 'modularity'.

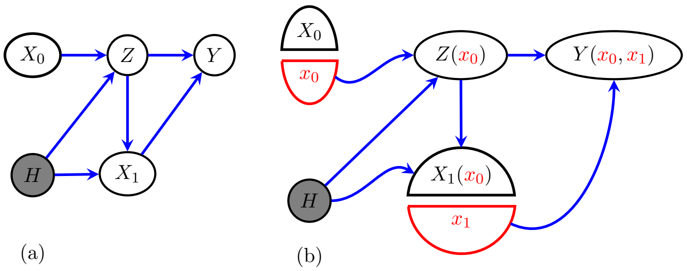


Figure 8: (a) The DAG $\mathcal{G}$, Ex. 11.3.3, Fig. 11.12 in Pearl (2009, p.353); H is unobserved ; (b) the template $\mathcal{G}(x_0, x_1)$.

The property of 'factorization' is simply that the (marginal) distribution $P(\mathbb{V}(\tilde{\mathbf{a}}))$ over the counterfactual variables present in the SWIG factors according to the respective $\mathcal{G}(\tilde{\mathbf{a}})$. This property is equivalent to the global Markov property, aka d-separation.

The 'modularity' property is that the conditional distribution associated with a counterfactual variable $Y(\tilde{\mathbf{a}})$, given its parents in $\mathcal{G}(\tilde{\mathbf{a}})$ is obtained from the conditional distribution of Y given its parents in $\mathcal{G}$. Formally, modularity may be seen as imposing a link between two sets of distributions that factor with respect to different graphs $(\mathcal{G}(\tilde{\mathbf{a}}), P(\mathbb{V}(\tilde{\mathbf{a}})))$ and $(\mathcal{G}, P(\mathbf{V}))$. This property follows immediately from the usual counterfactual consistency property and the conditional independences in the $\mathcal{G}(\tilde{\mathbf{a}})$.

5 Example

Pearl (2009) in Example 11.3.3 claims that under the NPSEM associated with the causal DAG in Figure 8(a) the following conditional independence does not hold:

$$Y(x_0, x_1) \perp\!\!\!\perp X_1 \mid Z, X_0 = x_0. \qquad (9)$$

Pearl concludes from this that a claim of Robins is false because if it were true then (9) would hold. However, direct inspection of the SWIG shown

[15] Note that this is the set of fixed nodes that are ancestors of the counterfactual node corresponding to V after having split *every* node in $\mathbf{A}$.

in Figure 8(b) shows that (9) is indeed true under this NPSEM, and that Pearl is thus incorrect. Specifically, we see by examining the template $\mathcal{G}(x_0, x_1)$ shown in Figure 8(b), that:

$$Y(x_0, x_1) \perp\!\!\!\perp X_1(x_0) \mid Z(x_0), X_0, \qquad (10)$$

from which it follows that

$$Y(x_0, x_1) \perp\!\!\!\perp X_1(x_0) \mid Z(x_0), X_0 = x_0. \qquad (11)$$

This last condition is then equivalent to (9) via the counterfactual consistency condition. Pearl made the following error. He correctly states that using his Twin Network method it may be shown that $Y(x_0, x_1)$ is not independent of X_1, given Z and X_0. However, he then goes on to say: "Therefore, [(9)] is not satisfied for $Y(x_0, x_1)$ and X_1."

As we have seen, reasoning with SWIGs immunizes us against this sort of error.

6 The FFRCISTG counterfactual model

We now give a formal definition of the FFRCISTG model (Robins, 1986) associated with a graph $\mathcal{G}$. Let $\mathbf{V}$ be the set of observed variables. In the current paper we suppose that every $V \in \mathbf{V}$ may be intervened upon. Thus we assume the counterfactual $V(\tilde{\mathbf{r}})$ for any assignment $\tilde{\mathbf{r}}$ to $\mathbf{R} \subset \mathbf{V}$ exists and is defined as follows:

(i) For each variable $V \in \mathbf{V}$ and assignment $\widetilde{\mathbf{pa}}$ to $\mathrm{pa}_{\mathcal{G}}(V)$, the parents of V in $\mathcal{G}$, we assume the existence of a counterfactual variable $V(\widetilde{\mathbf{pa}})$.

(ii) For any set $\mathbf{R}$, with $\mathbf{R} \neq \mathrm{pa}_{\mathcal{G}}(V)$, $V(\tilde{\mathbf{r}})$ is defined recursively via:

$$V(\tilde{\mathbf{r}}) = V\left(\tilde{\mathbf{r}}_{(\mathrm{pa}_{\mathcal{G}}(V)\cap\mathbf{R})}, (\mathbf{PA}_V \setminus \mathbf{R})(\tilde{\mathbf{r}})\right),$$

where $(\mathbf{PA}_V \setminus \mathbf{R})(\tilde{\mathbf{r}}) \equiv \{V^*(\tilde{\mathbf{r}}) \mid V^* \in \mathrm{pa}_{\mathcal{G}}(V), V^* \notin \mathbf{R}\}$. [16]

We may view the counterfactuals $V(\widetilde{\mathbf{pa}}_V)$, in condition (i) as primitives, from which all others are derived. For a given V the collection of such counterfactuals $\{V(\widetilde{\mathbf{pa}}_V) \mid \widetilde{\mathbf{pa}}_V\}$ may equivalently be represented via a structural equation:

$$V(\widetilde{\mathbf{pa}}_V) = f_V(\widetilde{\mathbf{pa}}_V, \boldsymbol{\epsilon}_V), \qquad (12)$$

where $\boldsymbol{\epsilon}_V$ is an error term. See also (Galles and Pearl, 1998; Halpern, 1998) for the generalization to non-recursive models.[17]

The *FFRCISTG model* is then defined as the set of distributions that satisfy the following independence assumption: For every $\mathbf{v}^{\dagger}$, the variables in the set

$$\left\{ V(\mathbf{pa}^{\dagger}_V) \;\middle|\; V \in \mathbf{V},\ \mathbf{pa}^{\dagger}_V = \mathbf{v}^{\dagger}_{\mathrm{pa}_{\mathcal{G}}(V)} \right\} \qquad (13)$$

are mutually independent.

Thus for every $\mathbf{v}^{\dagger}$ we assume that given an intervention $\mathbf{v}^{\dagger}$ to every variable in $\mathbf{V}$, the corresponding counterfactuals $V(\mathbf{pa}^{\dagger}_V)$ will be mutually independent.[18]

Our main result is then the following:

Theorem 1 *Given any distribution in the FFRCISTG model for $\mathcal{G}$, if $\mathbb{V}(\tilde{\mathbf{a}})$ is the set of counterfactual variables appearing on the SWIG $\mathcal{G}(\tilde{\mathbf{a}})$ then the marginal distribution $P(\mathbb{V}(\tilde{\mathbf{a}}))$ factorizes with respect to $\mathcal{G}(\tilde{\mathbf{a}})$; further $(P(\mathbb{V}(\tilde{\mathbf{a}})), \mathcal{G}(\tilde{\mathbf{a}}))$ obeys modularity with respect to $(P(\mathbf{V}), \mathcal{G})$.*

[16] Note that if $\mathbf{R}$ contains all of the parents of V then (ii) implies $V(\tilde{\mathbf{r}}) = V(\tilde{\mathbf{r}}_{(\mathrm{pa}_{\mathcal{G}}(V)\cap\mathbf{R})})$. Thus if we are intervening on all the parents of a variable then interventions on any other variable are irrelevant. This is referred to as the individual level 'exclusion restriction'; see Rule 1 in Pearl (2000, 2009), p.239. In addition, $A(\tilde{a}) = A$, and more generally $A(\tilde{\mathbf{r}}) = A(\tilde{\mathbf{r}}_{\mathbf{R}\setminus\{A\}})$. This usage fits with the conception that A represents the 'natural' level of treatment that the patient would receive if they were not being assigned value $\tilde{a}$. Assumption (ii) combines the assumptions described in other works as 'consistency' and 'recursive substitution'.

[17] This equivalence of potential outcomes and NPSEMs is sometimes misinterpreted as indicating that the Markov structure of all potential outcome models may be represented graphically (via a single graph). An NPSEM with FFRCISTG independence structure show this is false. An FFRCISTG does not obey the composition axiom: under the FFRCISTG $X \rightarrow Y$, we have $X \perp\!\!\!\perp Y(x_0)$, $X \perp\!\!\!\perp Y(x_1)$ but not $X \perp\!\!\!\perp Y(x_0), Y(x_1)$. This precludes representation via a (pathwise) graphical Markov property.

[18] Robins and Richardson (2011) prove that the set of independences (13) is equivalent to the set of counterfactual independence relations used in the definition of the FFRCISTG appearing in Robins (1986) and Robins and Richardson (2011); see also Appendix C in Richardson and Robins (2013).

References

Balke, A. and J. Pearl (1994). Probabilistic evaluation of counterfactual queries. In *Proceedings of the* 12^{th} *Conference on Artificial Intelligence*, Volume 1, Menlo Park, CA, pp. 230–7. MIT Press.

Evans, R. J. (2012). Graphical methods for inequality constraints in marginalized DAGs. In *Machine Learning for Signal Processing*. IEEE.

Galles, D. and J. Pearl (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science 3*(1), 151–182.

Geneletti, S. and A. P. Dawid (2007). Defining and identifying the effect of treatment on the treated. Technical Report 3, Imperial College London, Department of Epidemiology and Public Health.

Halpern, J. (1998). Axiomatizing causal reasoning. In *Proceedings of the Fourteenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)*, San Francisco, CA, pp. 202–210. Morgan Kaufmann.

Hernán, M. A., S. Hernández-Díaz, and J. M. Robins (2004). A structural approach to selection bias. *Epidemiology 15*(5), 615–625.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.

Pearl, J. (1995). Causal diagrams for empirical research (with discussion). *Biometrika 82*, 669–690.

Pearl, J. (2000). *Causality*. Cambridge, UK: Cambridge University Press.

Pearl, J. (2009). *Causality* (Second ed.). Cambridge, UK: Cambridge University Press.

Pearl, J. and J. M. Robins (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In P. Besnard and S. Hanks (Eds.), *Uncertainty in Artificial Intelligence: Proceedings of the* 11^{th} *Conference*, San Francisco, pp. 444–453. Morgan Kaufmann.

Richardson, T. S. and J. M. Robins (2013). Single World Intervention Graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Technical Report 128, Center for Statistics and the Social Sciences, University of Washington.

Robins, J. M. (1986). 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.

Robins, J. M., M. A. Hernán, and U. Siebert (2004). Effects of multiple interventions. In M. Ezzati, C. J. L. Murray, A. D. Lopez, and A. Rodgers (Eds.), *Comparative quantification of health risks : global and regional burden of disease attributable to selected major risk factors*, Volume 2, Chapter 28, pp. 2191–2230. Geneva: World Health Organization.

Robins, J. M. and T. S. Richardson (2011). Alternative graphical causal models and the identification of direct effects. In P. Shrout, K. Keyes, and K. Ornstein (Eds.), *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*, Chapter 6, pp. 1–52. Oxford University Press.

Robins, J. M., T. J. VanderWeele, and T. S. Richardson (2006). Discussion of "Causal effects in the presence of non compliance a latent variable interpretation" by Forcina, A. *Metron LXIV*(3), 288–298.

Shpitser, I. (2013). Personal e-mail communication.

Shpitser, I. and J. Pearl (2007). What counterfactuals can be tested. In R. Parr and L. van der Gaag (Eds.), *Proceedings of the 23rd Annual Conference on Uncertainty in Artificial Intelligence*, Corvallis, Oregon, pp. 437–444.

Shpitser, I. and J. Pearl (2008). What counterfactuals can be tested. *Journal of Machine Learning Research 9*, 1941–1979.

Shpitser, I. and J. Pearl (2009). Effects of treatment on the treated: Identification and generalization. In *Proceedings of the Twenty-Fifth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-09)*, Corvallis, Oregon, pp. 514–521. AUAI Press.

Shpitser, I., T. S. Richardson, and J. M. Robins (2011). An efficient algorithm for computing interventional distributions in latent variable causal models. In *27th Conference on Uncertainty in Artificial Intelligence (UAI-11)*. AUAI Press.

Shpitser, I., T. VanderWeele, and J. Robins (2010). On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, Corvallis, Oregon, pp. 527–536. AUAI Press.

Spirtes, P., C. Glymour, and R. Scheines (1993). *Causation, Prediction and Search.* Number 81 in Lecture Notes in Statistics. Springer-Verlag.

Textor, J. and M. Liskiewicz (2011). Adjustment criteria in causal diagrams: An algorithmic perspective. In *Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)*, Corvallis, Oregon, pp. 681–688. AUAI Press.