Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality

Thomas S. Richardson
University of Washington

James M. Robins
Harvard University

Working Paper Number 128
Center for Statistics and the Social Sciences
University of Washington

30 April 2013

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 implied by the counterfactual models developed in Robins (1986, 1987). Moreover, 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). Although Robins (1986, 1987) did not use DAGs we translate his algebraic results to facilitate understanding of this prior work. An attractive feature of Robins' approach is that it largely avoids making counterfactual independence assumptions that are experimentally untestable. As an important illustration we revisit the critique of Robins' g-computation given in (Pearl, 2009, Ch. 11.3.7); we use SWIGs to show that all of Pearl's claims are either erroneous or based on misconceptions.

We also show that simple extensions of the formalism may be used to accommodate dynamic regimes, and to formulate non-parametric structural equation models in which assumptions relating to the absence of direct effects are formulated at the population level. Finally, we show that our graphical theory also naturally arises in the context of an expanded causal Bayesian network in which we are able to observe the natural state of a variable prior to intervention.

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 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); see (37) below. 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) considered in Pearl (2000) and the Minimal Counterfactual Model (MCM) of Robins and Richardson (2011).[2]

In fact, if, following Geneletti and Dawid (2007) and Robins et al. (2007)

[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 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. See Appendix C for details and the original, more general, definition.

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

we postulate the existence of additional observed variables giving the natural state that a variable would take prior to intervention, but before the natural value has any effects, then the independence and factorization properties encoded by our transformed graph are even satisfied by causal models that do not explicitly involve potential outcomes, such as 'causal Bayesian networks' (Pearl, 2000) and causal influence diagrams (Dawid and Didelez, 2008, 2010); see §9 for further details.

More specifically, 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) will have identical node sets. Furthermore, if the factuals on the DAG have a positive distribution, 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. This completeness property (with respect to the variables present in the graph) does not hold for either the 'twin network' approach of Balke and Pearl (1994); Pearl (2000) or the 'counterfactual graphs' of Shpitser and Pearl (2007, 2008).[3]

At the end of the Section 2, in §4.2 and in Appendix D we provide a number of examples where this lack of completeness for twin-networks leads Pearl, one of the creators of the twin-network method, to draw erroneous

[3]The d-separation criterion applied to the SWIG is complete (relative to the subset of variables in the SWIG) with respect to both the NPSEM-IE and FFRCISTG models associated with the original graph $\mathcal{G}$. d-separation is not complete for twin-networks since they include more variables amongst which there are deterministic relations; see §D.

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 'twin 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; see §6.

conclusions with regard to counterfactual independence.

Linking graphs and counterfactuals with minimal assumptions

As just noted, our approach differs from previously proposed attempts to link graphs and counterfactuals such as the 'twin network' approach of Balke and Pearl (1994); Pearl (2000) and the 'counterfactual graph' of Shpitser and Pearl (2007, 2008). Some of these differences reflect the fact that every counterfactual independence relation implied by our graph holds in the FFRCISTG model of Robins (1986); the other approaches encode additional counterfactual independence properties that are satisfied *only* under the more restrictive NPSEM-IE model, a strict submodel of the FFRCISTG model. It follows that any independence that holds under our model also holds under an NPSEM-IE; however, the converse is false.

More specifically, an FFRCISTG, in contrast to an NPSEM-IE:

- makes (exponentially) fewer 'cross-world' counterfactual independence assumptions that are experimentally untestable;
- in marked contrast to all graphical (Markov) models considered heretofore the FFRCISTG does not obey the composition axiom of independence[4] with respect to the full universe of variables. However, the model does obey composition with respect to subsets of counterfactual variables that are present on the same SWIG. Thus we use the usual criterion (d-separation) to read independence.

Our approach also leads naturally to the specification of models in which well-defined interventions (or potential outcomes) exist only for a strict subset of the variables as in (Robins, 1986).

Removing a false trichotomy

A primary aim of this paper is thus to show that researchers in causality are *not* forced to choose between:

- Use causal graphs without counterfactuals;
- Use counterfactuals without graphs;

[4] The 'composition axiom' is the implication: $X \perp\!\!\!\perp Y \mid Z \;\&\; X \perp\!\!\!\perp W \mid Z \Rightarrow X \perp\!\!\!\perp \{Y, W\} \mid Z$ (See Pearl, 1988, p.128); this should not be confused with the composition axiom for counterfactuals (See Pearl, 2000, 2009, p.229).

- Combine graphs and counterfactuals via the NPSEM-IE framework, as advocated by Pearl, thereby being required to make many counterfactual independence assumptions that are not potentially testable.

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.

We emphasize that we are not arguing that NPSEM-IE models are never appropriate. Rather our point is that an analyst who adopts a counterfactual theory need not also adopt an NPSEM-IE, as there exist counterfactual models – the FFRCISTGs of (Robins, 1986, 1989b) that, in contrast to NPSEM-IEs, encode few if any facts that are not potentially testable by performing an ideal Randomized Controlled Trial. In addition, the SWIGs make it easy to reason with these models. Furthermore, even though the FFRCISTG model makes fewer assumptions, most causal quantities of interest remain identified under the FFRCISTG if they are identified under the NPSEM-IE – Pure (aka Natural) Direct Effects and Principal Stratum Direct Effects being the primary exceptions; but see (Robins and Richardson, 2011) and Section 6.1.1 below for some subtleties.

To help delineate the issues involved, suppose there was good evidence for the FFRCISTG model because causal effects estimated under the model from the available observational data agreed with effects estimates from well-conducted randomized trials. Then one might believe, based on either simplicity arguments (aka a faithfulness assumption; Spirtes et al. (1993)), physical law, or Bayesian inference with a prior over structures, that it was highly likely that the corresponding NPSEM-IE model was true. In this case, if the NPSEM-IE identified the Pure Direct Effect, one might provide point estimates for this quantity. Our point is simply that such inferences rely on extrapolation and thus must be argued for in each case.

We have been arguing that an advantage of the FFRCISTG model compared to the NPSEM-IE model is that nearly every prediction from observational data of a causal effect identified under the FFRCISTG can be tested, at least in principle, by comparing the prediction with the results of a randomized controlled trial. However, this view ignores the facts that (i) such experimental trials may be infeasible or unethical; (ii) the validity of the comparison will typically require an auxiliary assumption of exchangeability between the experimental and observational population and the ability to measure all the variables included in the causal model, neither of which is likely to hold in practice; and (iii) such tests are themselves based upon

the untestable assumption that experimental interventions are ideal. Thus, many philosophers of science do not agree with a sharp separation between testable and non-testable causal predictions. Although we agree the separation may not be sharp we believe that the distinction nonetheless remains useful both conceptually and practically.

Further advantages of the new approach

Our new approach based on SWIGs has a number of further advantages over previous approaches. As discussed above the foremost advantage of our approach is that it allows one to write down a factorization of the joint distribution of the counterfactuals on the SWIG and to read off counterfactual independencies via d-separation (Pearl, 1988).

Another advantage is that a simple modification of a SWIG[5] 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.

An additional advantage is that it simplifies the approaches of (Pearl, 1995; Pearl and Robins, 1995) to the identification of intervention distributions. In particular:

- The SWIG gives a graphical explanation as to why conditioning on variables so as to '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.
- The SWIG permits the criteria for identification by the g-computation formula 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.
- The SWIG approach naturally extends to dynamic regimes where treatment is assigned (either deterministically or stochastically) on the basis of prior covariates, including the natural value of treatment that a patient would choose (in the absence of it being specified by the regime).

In the next section we give motivating examples, followed by an outline of the rest of the paper.

[5] See §7.

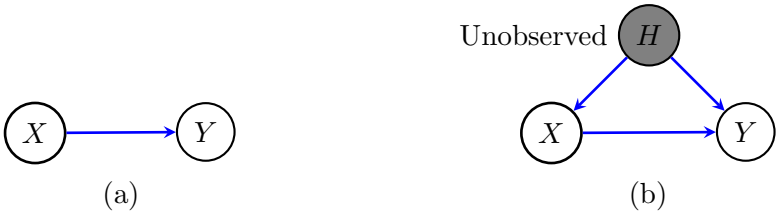


Figure 1: (a) A causal DAG representing two unconfounded variables; (b) A causal DAG representing the presence of confounding.

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.[6]

[6]Many readers may be more familiar with the 'do' notation of Pearl (1995) or the g-notation of Robins (1986, 1989a) than with the potential outcome notation adopted here. We describe their inter-relationships in order to facilitate translation. Pearl's 'do' notation is a special case of the g-notation $P(Y = y \mid g)$, introduced in Robins (1986, p.1423). Here 'g' stands for some '$g$eneralized' treatment regime or, equivalently, plan. The set of possible generalized treatment regimes 'g' included non-dynamic (aka static) regimes

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 (see Appendix A for a review). 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) present an insurmountable 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. 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.

By applying d-separation to the graph in Figure 2(a), we directly obtain that $X \perp\!\!\!\perp Y(0)$, since there are no edges emanating from the node containing X, hence there are no paths from X to $Y(0)$. Similarly, by applying d-separation to the graph in Figure 2(b) we derive $X \perp\!\!\!\perp Y(1)$.

We make a few remarks regarding the graphs 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. We keep the red nodes on the graph because their presence is needed to link the distribution of the counterfactual variables on the SWIG to the distribution of the factual

$x = x_1, \ldots, x_K$, deterministic dynamic regimes and random dynamic treatment regimes; see §5. In the case of a static regime, Robins (1989a, p.126) wrote $P(Y = y \,|\, g = (x))$. Pearl's notation for static regimes is identical except that $g = (x)$ is replaced by the equivalent $\text{do}(X = x)$.

To translate the do/g notation into counterfactual notation we then equate $P(Y = y \,|\, g = (x))$ to $P(Y(x) = y)$ (See Robins, 1989b). Pearl's 'do' notation is not as rich as the language of potential outcomes as there is no way to directly express $P(Y(x) \,|\, X = x')$ in terms of his 'do' notation. However, the g-notation of Robins (1986) does not suffer from this inexpressiveness. Specifically, although Robins defined $P(Y = y \,|\, g = (x), Z = z)$ to be $P(Y = y, Z = z | g = (x))/P(Z = z \,|\, g = (x))$ (a notation adopted by Pearl replacing 'g' with 'do'), Robins (1986) defined $P(Y = y \,|\, g = (x), [Z = z])$ to be $P(Y(x) \,|\, Z = z)$. Thus $P(Y(x) \,|\, X = x')$ can be obtained by taking Z to be X and z to be x'.

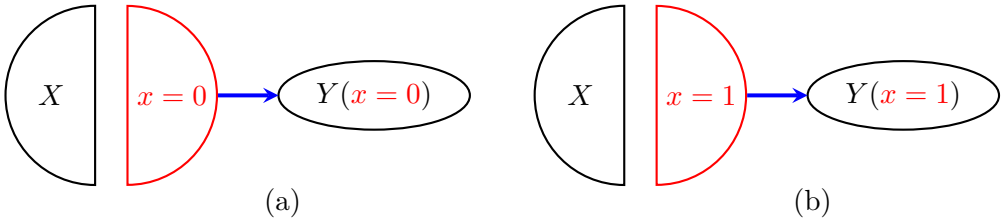


Figure 2: The single world intervention graphs (SWIGs) resulting from splitting node X in the graph in Figure 1(a), and intervening to set a particular value. (a) the SWIG $\mathcal{G}(x=0)$ corresponding to setting X to 0; (b) $\mathcal{G}(x=1)$ given by setting X to 1.

variables on the original graph; see Section 3.6.3 for further details. These nodes also play a central role in the analysis of dynamic regimes; see §5.

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 $P(X, Y(0))$ and $P(X, Y(1))$ respectively. Likewise, we will associate the following factorizations with these graphs:

$$\begin{aligned} P(X=x, Y(0)=y) &= P(X=x)P(Y(0)=y), \\ P(X=x, Y(1)=y) &= P(X=x)P(Y(1)=y). \end{aligned} \tag{3}$$

Notice that, if we ignore the red nodes, these factorizations are simply instances of the standard DAG factorization (see (104) in Appendix A), since in Figure 2(a) neither X nor $Y(0)$ have any parents (besides the red nodes). In addition, one can see that these factorizations are equivalent to the independence conditions (2).

In addition we associate the following equation

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

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

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

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 $x=0$ 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,

$$P(Y(0)=y) = P(Y(0) \mid X=0) = P(Y=y \mid X=0); \tag{6}$$

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 implied 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):

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

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 $P(X, Y(0))$ and $P(X, Y(1))$ we

[7] Our usage of the term 'modularity' differs from that of Pearl Pearl (2009), though they both derive from the same intuition.

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

have not constructed a graph that includes *both* $Y(0)$ *and* $Y(1)$, and thus represents the joint distribution $P(X, Y(0), Y(1))$. 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. However, it is worth noting that the independence restrictions that we encode may place inequality restrictions on the (multiple-world) joint distribution (e.g. $P(X, Y(0), Y(1))$) over all counterfactuals.[9]

As noted, the counterfactual independencies encoded in a SWIG are implied by the FFRCISTG and NPSEM-IE models. However, the NPSEM-IE also encodes many additional cross-world restrictions on the joint distribution over all counterfactuals. These counterfactual independencies further imply additional identification results. For example, Robins and Greenland (1992) introduced the Pure (or Natural) Direct Effect and showed it is never identified non-parametrically from randomized experiments performed on the variables on the graph. In contrast, Pearl (2001a) showed that it can be identified by the mediation formula under the NPSEM-IE corresponding to the graph in Figure 9(i).

Templates

Since it is cumbersome and 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] Again, to emphasize that each instantiation of a template only contains variables from a single world, we refer to them as '*S*ingle *W*orld *I*ntervention *T*emplates' or 'SWITs'.

[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.

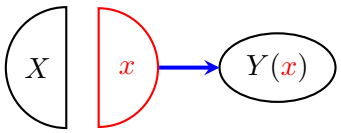


Figure 3: A template representing the two graphs in Figure 2.

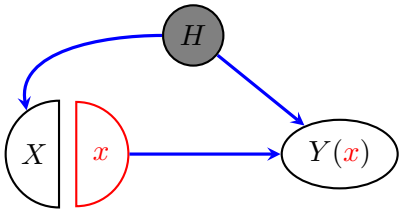


Figure 4: The template resulting from intervening on X in the graph in Figure 1(b).

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. It is well known that this condition is sufficient for the effect of X on Y to be given via the standard adjustment formula:

$$P(Y(\tilde{x})=y) = \sum_l P(Y=y \mid L=l, X=\tilde{x})P(L=l).^{11} \tag{8}$$

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.

[11] This is also a special case of Theorem 22 in Section 4.

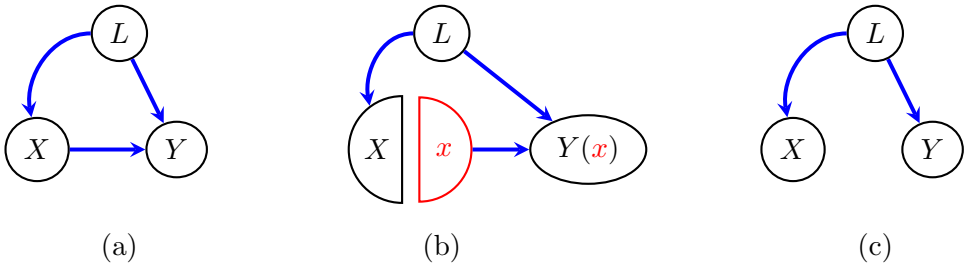


Figure 5: Adjusting for confounding. (a) The original causal graph. (b) The template $\mathcal{G}(x)$, which shows that $Y(x) \perp\!\!\!\perp X \mid L$. (c) The DAG $\mathcal{G}_{\underline{X}}$ obtained by removing edges from X advocated in Pearl (1995, 2000, 2009) to check his 'backdoor condition'.

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 $\mathbf{L}$:

$$P(Y(\tilde{x}) = y) = \sum_{\mathbf{l}} P(Y = y \mid \mathbf{L} = \mathbf{l}, X = \tilde{x}) P(\mathbf{L} = \mathbf{l}), \tag{9}$$

the backdoor criterion (Pearl, 2000, 2009, p.70) 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; see (Shpitser et al., 2012) for further discussion. The reason for this additional condition is not transparent, since the inclusion of

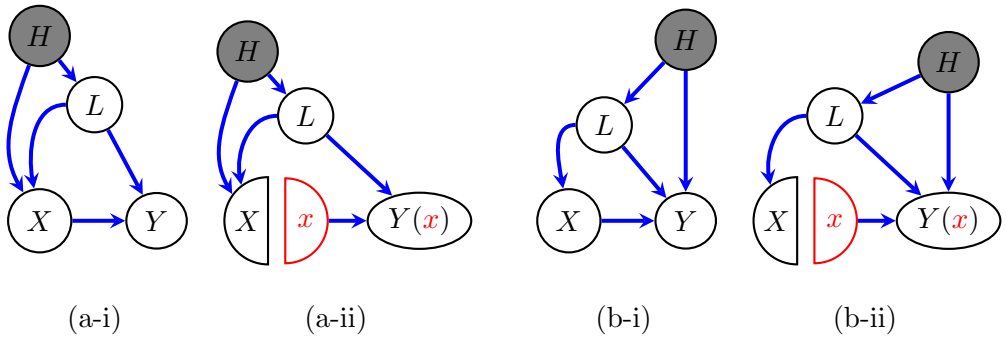


Figure 6: Further examples of adjusting for confounding. (a-i) A graph $\mathcal{G}$; (a-ii) the template $\mathcal{G}(x)$; (b-i) A graph $\mathcal{G}'$; (b-ii) the template $\mathcal{G}'(x)$. H is an unobserved variable in $\mathcal{G}$ and $\mathcal{G}'$. Both SWITs imply $Y(x) \perp\!\!\!\perp X \mid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

such a variable does not preclude that X and Y may be d-separated in $\mathcal{G}_{\underline{X}}$.[12] Within the framework given here there is no need to state this additional condition.

Using SWIGs we have the simple adjustment criterion:

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 require no restrictions on the membership in $\mathbf{L}$. 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

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

but the template does not imply $X \perp\!\!\!\perp Y(\tilde{x}) \mid L_1, L_2$. Moreover, this independence is not implied by any template constructed from $\mathcal{G}$.[13]

[12] Pearl (2009, §11.3.3, p.344) acknowledges that the need to restrict to non-descendants is not transparent in his original derivation, and offers an alternative. However, this reformulation requires two different independence relations to be checked.

[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 $X \perp\!\!\!\perp Y(\tilde{x}) \mid L_1 = l_1, L_2 = l_2$.

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.

The Advantages of Completeness

The next example shows that the completeness of d-separation for the variables arising on a SWIG can prevent errors. Pearl (2009, Ex. 11.3.3) claims that under the NPSEM associated with the causal DAG in Figure 15(a) the following conditional independence does *not* hold:

$$Y(x_0, x_1) \perp\!\!\!\perp X_1 \mid Z, X_0 = x_0. \tag{10}$$

Pearl concludes from this that a claim of Robins is false because if the claim were true then (10) would hold. However, direct inspection of the SWIG shown in Figure 15(b) shows that (10) 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 15(b), that:

$$Y(x_0, x_1) \perp\!\!\!\perp X_1(x_0) \mid Z(x_0), X_0, \tag{11}$$

from which it follows that

$$Y(x_0, x_1) \perp\!\!\!\perp X_1(x_0) \mid Z(x_0), X_0 = x_0. \tag{12}$$

This last condition is then equivalent to (10) via the counterfactual consistency condition. Pearl made the following error. He correctly states that

> $Y(x_0, x_1)$ is not independent of X_1, given Z and X_0.

However, he then goes on to mistakenly say:

> Therefore, [(10)] 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. See §4.2 for the context and implications of this error; Appendix D.4 for an exegesis and other related errors in Pearl (2000).

Overview

We now give a brief overview of the development in the remainder of the paper:

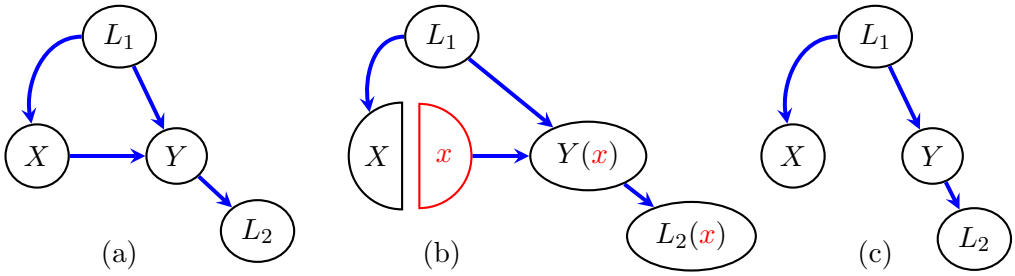


Figure 7: Simplification of the backdoor criterion. (a) The original causal graph $\mathcal{G}$. (b) The template $\mathcal{G}(x)$, which shows that $Y(x) \perp\!\!\!\perp X \mid L_1$, but does not imply $Y(x) \perp\!\!\!\perp X \mid \{L_1, L_2\}$ when there exists an arrow from X to Y, i.e. the null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

SWIGs for NPSEMs with FFRCISTG independence

In Section 3 our starting point is a non-parametric structural equation model (NPSEM) that is naturally associated with the DAG $\mathcal{G}$ (with node set $\mathbf{V}$ comprising a set of 'factual' variables): Each random variable $V \in \mathbf{V}$ is expressed as an arbitrary function f_V of the variables that are the parents of V in the graph together with an error term ε_V. Given a distribution over the error terms, the equations then specify a joint distribution over $\mathbf{V}$. We impose on these error terms the independence assumptions of the FFRCISTG model of Robins (1986); Robins and Richardson (2011). As we shall see, these assumptions are much weaker than the NPSEM-IE model of Pearl (2000, 2009), which requires that the error terms associated with different variables be independent. The latter assumption is not (even in principle) testable by any randomized experiment.

Suppose now we are given a set of treatment variables $\mathbf{A} \subseteq \mathbf{V}$; we will 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 the 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 $\mathcal{G}$. In the usual NPSEM framework it is (implicitly) assumed that all variables may (in principle) be intervened upon, and that all of the resulting counterfactuals are well-defined. Though we adopt this assumption for much of our development, we subsequently relax it and show that our main results

continue to hold (see §8 and discussion below).

In Section 3.3 we give a general graphical transformation algorithm that, given $\mathcal{G}$ and an assignment $\tilde{\mathbf{a}}$ to $\mathbf{A}$ as input, constructs a SWIG $\mathcal{G}(\tilde{\mathbf{a}})$ containing counterfactual nodes $\mathbb{V}(\tilde{\mathbf{a}})$ and fixed nodes $\tilde{\mathbf{a}}$. Given a different assignment $\mathbf{a}^\dagger$ to $\mathbf{A}$, our algorithm will generate a SWIG $\mathcal{G}(\mathbf{a}^\dagger)$ with a different node set $\mathbb{V}(\mathbf{a}^\dagger) \cup \mathbf{a}^\dagger$. We may represent the collection of all such SWIGs via a template $\mathcal{G}(\mathbf{a})$, which formally can be seen as a graph-valued function: it takes an assignment $\tilde{\mathbf{a}}$ to $\mathbf{A}$ as input and returns a graph $\mathcal{G}(\tilde{\mathbf{a}})$.

We prove in Propositions 11 and 16 that the distribution $P(\mathbb{V}(\tilde{\mathbf{a}}))$ of the counterfactuals $\mathbb{V}(\tilde{\mathbf{a}})$ that are vertices in the SWIG $\mathcal{G}(\tilde{\mathbf{a}})$ under the FFRCISTG NPSEM satisfies two important properties that we call factorization and modularity:

> 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; see §3.5.
>
> 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 $\mathcal{G}$; see §3.6. 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 $(\mathcal{G}, P(\mathbf{V}))$.

We will show that many important results and properties follow directly from the properties of factorization and modularity alone. In particular, these properties hold if and only if the joint density $P(\mathbb{V}(\tilde{\mathbf{a}}))$ is equal to a specific function of the joint distribution of the factuals: the extended g-formula density of Robins et al. (2004, p.2222); see (37) below.[14] Since the NPSEM-IE is a submodel of the FFRCISTG model of Robins (1986) (associated with the same graph), it follows that the NPSEM-IE also satisfies factorization and modularity.

In Section 11.3.7 of his second edition Pearl (2009) offers three critiques of the g-formula based methodology developed in Robins (1986, 1987) for estimation of the causal effects of time-varying treatments. In Section 4 we demonstrate that all of Pearl's claims are either erroneous or based on misconceptions. In Section 4.1 we provide the necessary background by

[14] The 'g-formula' is also referred to as the 'g-computation formula'.

reviewing Robins' g-formula methodology with the aid of SWIGs . In Section 4.2 we show that Pearl's claims are erroneous: two of the false claims are the direct result of mathematical errors made by Pearl; the third is the result of Pearl's misreading or lack of awareness of an example in Robins (1986).

In Section 5 we show how our graphical formalism may be easily extended to represent dynamic regimes. This includes regimes where treatment assignment depends on the value that a variable would have in the absence of an intervention. For example, consider the regime '*Exercise for as long as you would have done without intervention or twenty minutes, whichever is more*'. We show how the identification results described in §4.2 generalize to this setting.

NPSEMs with population-level exclusion restrictions

As we have mentioned, the motivation for adopting the FFRCISTG model, rather than the NPSEM-IE is in large part driven by *epistemic* considerations: there is no randomized experiment that may be performed (even in principle) to verify the independence assumptions that are made by an NPSEM-IE. However, this same criticism may also be made regarding the assumptions associated with the absence of a directed edge in an NPSEM even under the FFRCISTG assumptions.[15]

As we have described above, in an NPSEM each variable is given by a function f_V expressed as an arbitrary function of the variables that are its *parents* in the graph, together with an error term.[16] Thus the absence of an edge in a DAG $\mathcal{G}$ implies that there is no *individual level direct effect*, also known as an exclusion restriction. However, there is no consistent test for the absence of an individual level effect, from randomized experiments. This is because it is only the average or population-level effect that is identified, and it is possible for the population level effects to be zero, and yet the individual level effects to be non-zero.[17] Thus although the FFRCISTG assumptions relating to independence of errors in the NPSEMs considered in §3 and (Robins and Richardson, 2011) may be verified via randomized experiment this is not true of the restrictions relating to the absence of direct effects in these models.[18] To make this issue concrete, we will show a simple

[15] A point also noted by Pearl (2010) in a response to (Robins and Richardson, 2011).

[16] As noted above, the FFRCISTG and NPSEM-IE differ regarding the independence assumptions that they impose on the distribution of the errors.

[17] In the simple case of a binary treatment A and binary response Y this corresponds to the situation in which there are equal, but non-zero, proportions of patients who are helped and hurt, so that $P(Y(a=0)=0, Y(a=1)=1) = P(Y(a=0)=1, Y(a=1)=0) \neq 0$.

[18] In graphical terms, the NPSEM FFRCISTG models of §3 and (Robins and Richardson,

graph under which the Effect of Treatment on the Double Treated

$$E[Y(a_1=1, a_2=1) - Y(a_1=0, a_2=0) \mid A_1=1, A_2=1] \tag{13}$$

is identified via individual-level exclusion restrictions, yet fails to be identified under population-level exclusions. Were we to assume the absence of an individual-level effect, when this was not the case, then we could draw erroneous inferences, and yet there would be no randomized experiment that would allow us to detect this error.[19]

In section 7 we address this epistemological issue by constructing a counterfactual theory that interprets the absence of an edge in a graph in terms of the absence of a population level direct effect. In other words, the absence of an edge from X to Y implies that interventions on X will have no effect on the *distribution* of Y, but does not assume that $Y(x) = Y(x')$, so that there may still be an effect at the individual level.

Our theory still assumes an underlying NPSEM, but does not assume that missing edges in the graph imply the absence of variables from equations in the NPSEM. Thus the relationship between the DAG and the NPSEM in this theory differs from the standard approach: missing edges in the DAG do not lead to variables being omitted from equations in the NPSEM. Fortunately, accommodating this weaker interpretation of missing edges requires only a minor change to the SWIG construction procedure (in fact, the resulting construction is simpler!). Since the resulting distributions $P(\mathbb{V}(\tilde{\mathbf{a}}))$ implied by the NPSEM will still obey modularity and factorization with respect to $\mathcal{G}(\tilde{\mathbf{a}})$, the theory developed in Section 4 in the traditional NPSEM setting – where the absence of an edge corresponds to the absence of an individual direct effect – goes over to this more general context with very little change.[20]

Counting the untestable assumptions implied by an NPSEM-IE

In Section 6 we conclude our comparison of the FFRCISTG and NPSEM-IE independence assumptions by counting the dimension of the associated

2011) require experimentally verifiable conditions for the absence of 'confounding arcs' ($\leftrightarrow$) but not for the absence of directed edges ($\rightarrow$).

[19] Thus the Effect of Treatment on the Double Treated plays the same role in the FFRCISTG theory with individual-level exclusions that the Pure Direct Effect plays for the NPSEM-IE. Namely, it is a causally interpretable quantity that is identified via assumptions that are not experimentally testable.

[20] These SWIGs may also be seen as indicating that Pearl's 'causal hierarchy' that ties counterfactuals to individual level exclusion restrictions is an oversimplification.

counterfactual models in the case of complete graphs with binary variables. This reveals that the NPSEM-IE model imposes super-exponentially more independence assumptions that are experimentally untestable. This leads us to question whether NPSEM-IEs and potential outcomes are functionally 'equivalent' theories as is claimed by (Pearl, 2000, 2009, 2012a). In particular, the FFRCISTG model is a simple and natural potential outcome model whose conditional independence structure cannot be faithfully represented by any NPSEM-IE.

Ontological Issues

As stated earlier, the issues relating to independence assumptions and exclusion restrictions are epistemic: if there is no experiment that can be performed (even in principle) to verify a particular assumption, then one might naturally be concerned about its adoption.

However, it may be the case that one does not think that the counterfactual corresponding to an intervention on a particular variable is well-defined.[21] Therefore in Section 8 we consider the situation in which interventions are only defined on a strict subset of variables, so that not all the counterfactual variables that would be implied by an NPSEM exist. We show that SWIGs may easily be applied in this setting. We discuss but leave open the question as to the meaning of directed edges emanating from variables that cannot be intervened upon.

In Section 9, we consider an even more restrictive ontological position whereby no counterfactuals are assumed to exist. We define an extended agnostic model (equivalently extended causal Bayes net), and show that our SWIG (under a suitable labeling) describes the structure of intervention distributions. Thus the graphical structure of a SWIG together with its associated factorization and Markov property naturally arises in the context of a purely interventional theory.

Appendices

In Appendix A we review graphical models based on directed acyclic graphs, specifically the concept of d-separation. Appendix B contains some techni-

[21] In the analysis of mediation and 'censoring by death' considerations of this kind have motivated the use of so-called 'principal-stratum direct effects' introduced in (Robins, 1986, Sec. 12.2), but named and popularized by Frangakis and Rubin (2002) as an alternative to 'controlled direct effects' since the latter, but not the former, require well-defined interventions on intermediate variables; see (Robins and Richardson, 2011) §2.3 for further discussion.

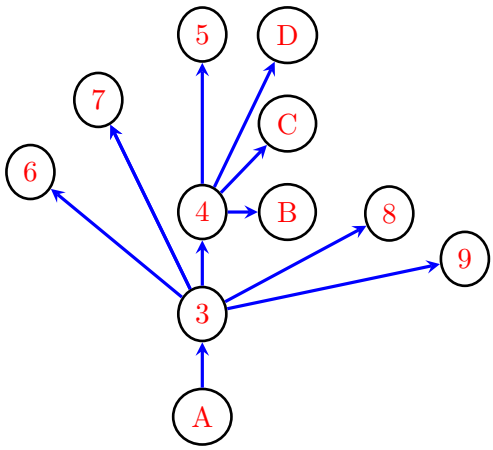


Figure 8: Organization of the paper.

cal proofs. Appendix C revisits several additional classes of counterfactual model introduced in Robins (1986), using SWIGs to illustrate the relevant distinctions. In Appendix D we give review the different proposals to unify graphs and counterfactuals presented in Pearl (2000) and Pearl (2009). We also chronicle a number of problems that have arisen in specific applications owing to the presence of deterministic relationships and context specific independence relations present in these graphs.

3 SWIGs applied to NPSEMs

In this section we develop the theory of SWIGs in the context of a non-parametric structural equation model. Throughout this section we will relate NPSEMs and graphs in the usual way: there is an equation for each variable in the model, specifying that variable as a function of its parents in the graph. Thus the *absence* of an edge $X \to Y$ in a graph $\mathcal{G}$ indicates the absence of a direct effect at the individual level, corresponding to the absence of X from the function f_Y in the structural equation for Y.[22]

3.1 Counterfactual Existence Assumptions

Let $\mathbf{V}$ be an underlying set of random variables. Throughout Section 3, we will assume that every variable $V \in \mathbf{V}$ may be intervened upon.[23] Thus we will assume the counterfactual $\mathbb{V}(\tilde{\mathbf{r}})$ for any assignment $\tilde{\mathbf{r}}$ to a subset $\mathbf{R} \subset \mathbf{V}$ exists and is defined as follows:

Definition 1 (NPSEM Counterfactual Existence Assumption)**.**

(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:*

$$V(\tilde{\mathbf{r}}) = V\left(\tilde{\mathbf{r}}_{(\mathrm{pa}_{\mathcal{G}}(V)\cap\mathbf{R})}, (\mathbf{PA}_V \setminus \mathbf{R})(\tilde{\mathbf{r}})\right), \tag{14}$$

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

Note that if $\mathbf{R}$ contains all of the parents of V then assumption (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 the value $\tilde{a}$.

Assumption (ii) combines the assumptions described in other works as 'consistency' and 'recursive substitution'.[24] These assumptions are common to both the NPSEM-IE model and the FFRCISTG model.

[22] In Section 7 we will consider graphs under a weaker causal interpretation whereby a missing edge corresponds merely to the absence of a direct effect at the population level.

[23] We will relax this assumption in the Section 8.

[24] See also the 'composition' property in Pearl (2000, 2009), p.229.

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[25] of such counterfactuals $\{V(\widetilde{\mathbf{pa}}_V) \mid \widetilde{\mathbf{pa}}_V\}$ may equivalently be represented via a structural equation:

$$V(\widetilde{\mathbf{pa}}_V) = f_V(\widetilde{\mathbf{pa}}_V, \boldsymbol{\epsilon}_V), \tag{15}$$

where $\boldsymbol{\epsilon}_V$ is an error term.[26]

3.1.1 Examples

Consider the DAG in Figure 9(i) which is associated with the following structural equations:

$$\begin{aligned} Z &= f_Z(\epsilon_Z), \\ M(z) &= f_M(z, \epsilon_M) \\ Y(z, m) &= f_Y(z, m, \epsilon_Y) \end{aligned}$$

by (i) in the Counterfactual Existence Assumption. These equations define the counterfactual variables Z, $M(z)$ and $Y(z, m)$. Condition (ii) implies the following equalities hold for all values of z, m, y:

$$\begin{aligned} Z(m, y) &= Z, & \qquad Z(m) &= Z, \\ M(z, y) &= M(z), & \qquad M(y) &= M(Z). \end{aligned} \tag{16}$$

Note that these restrictions simply express the concept that the future does not cause the past. Similarly, for all m and z we have the following:

$$\begin{aligned} Y(m) &= Y(m, Z), \\ Y(z) &= Y(z, M(z)), \\ Y &= Y(Z, M(Z)). \end{aligned}$$

By contrast the DAG in Figure 10(i) would be associated with the following structural equations:

$$\begin{aligned} Z &= f_Z(\epsilon_Z), & \qquad M(z) &= f_M(z, \epsilon_M), \\ Y(m) &= f_Y(m, \epsilon_Y), \end{aligned}$$

[25] Formally, a stochastic process to allow there to be uncountably many assignments $\widetilde{\mathbf{pa}}_m$.

[26] Given $\{V(\widetilde{\mathbf{pa}}_V) \mid \widetilde{\mathbf{pa}}_V\}$ we may trivially obtain such a representation by defining the error term $\boldsymbol{\epsilon}_V$ to be this set and f_V to be such that $f_V(\mathbf{pa}_V, \epsilon_V) \equiv (\boldsymbol{\epsilon}_V)_{\mathbf{pa}_V} = V(\mathbf{pa}_V)$.

which define the variables Z, $M(z)$ and $Y(m)$. The equalities (16) continue to hold but for Y we now have:

$$\begin{aligned} Y(z) &= Y(M(z)), \\ Y &= Y(M(Z)), \\ Y(z,m) &= Y(m). \end{aligned}$$

Note that the last equality here expresses the absence of a direct effect (at the individual-level) of Z on Y.

3.2 Definition of the FFRCISTG and NPSEM-IE models

The models that we consider correspond to sets of distributions over the set of counterfactuals whose existence is given by Definition 1.

Distributions in the FFRCISTG model will in addition make the following independence assumption. For every $\mathbf{v}^\dagger$ we assume that given an intervention $\mathbf{v}^\dagger$ to every variable in $\mathbf{V}$, the corresponding counterfactuals $V(\mathbf{pa}^\dagger)$ will be mutually independent. Formally we make the following:

Definition 2 (FFRCISTG Independence Assumption[27])**.**

$$\textit{For every } \mathbf{v}^\dagger, \textit{the variables } \left\{ V(\mathbf{pa}_V^\dagger) \;\middle|\; V \in \mathbf{V},\ \mathbf{pa}_V^\dagger = \mathbf{v}^\dagger_{\mathrm{pa}_\mathcal{G}(V)} \right\} \tag{17}$$

are mutually independent.

The NPSEM-IE model is the submodel of the FFRCISTG obeying the stronger independence assumption:[28]

Definition 3 (NPSEM-IE Independence Assumption)**.**
The variables $\{\boldsymbol{\epsilon}_V \mid V \in \mathbf{V}\}$ *are mutually independent.*
This is equivalent to:

$$\textit{The sets of variables } \left\{ \{V(\mathbf{pa}_V^\dagger) \mid \textit{for all } \mathbf{pa}_V^\dagger\} \;\middle|\; V \in \mathbf{V} \right\} \tag{18}$$

are mutually independent.

It is not hard to see that (18) implies (17), but not vice versa. We will investigate this distinction in much greater detail in Section 6 below. In the form stated here, neither assumption is particularly easy to interpret. However, we will see below that the logical implications of (17) are captured via the associated SWIGs.

[27] Robins and Richardson (2011) prove that the set of independences (17) 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 Definition 62 in Appendix C.

[28] See Pearl (2000, 2009) p.101, (3.56) and footnote 14, and p.239 Rule 2.

3.2.1 Examples

The FFRCISTG Independence Assumption for the graph in Figure 9(i) requires that for each pair of values z, m, the following three variables are jointly independent:

$$Z \perp\!\!\!\perp M(z) \perp\!\!\!\perp Y(z,m).$$

Thus, in particular if Z and M are binary, then the FFRCISTG assumption requires the following four joint independences:

$$\begin{aligned} &Z \perp\!\!\!\perp M(z=0) \perp\!\!\!\perp Y(z=0,m=0), \qquad Z \perp\!\!\!\perp M(z=0) \perp\!\!\!\perp Y(z=0,m=1),\\ &Z \perp\!\!\!\perp M(z=1) \perp\!\!\!\perp Y(z=1,m=0), \qquad Z \perp\!\!\!\perp M(z=1) \perp\!\!\!\perp Y(z=1,m=1). \end{aligned}$$

In contrast the NPSEM-IE Independence Assumption requires that the following sets of variables are jointly independent:

$$Z \perp\!\!\!\perp \{M(z) \text{ for all values of } z\} \perp\!\!\!\perp \{Y(z,m) \text{ for all values of } z,m\}$$

Thus if Z and M are binary then the NPSEM-IE assumption requires the following joint independence:

$$\begin{aligned} Z \perp\!\!\!\perp \{M(z=0), M(z=1)\} \perp\!\!\!\perp \{&Y(z=0,m=0), Y(z=0,m=1),\\ &Y(z=1,m=0), Y(z=1,m=1)\} \end{aligned}$$

For example, the NPSEM-IE implies that $M(z) \perp\!\!\!\perp Y(z',m)$ while this is not implied by the FFRCISTG assumption.

Likewise for the DAG in Figure 10(i), the FFRCISTG assumption implies:

$$Z \perp\!\!\!\perp M(z) \perp\!\!\!\perp Y(m), \qquad \text{for each pair of values } (z,m),$$

while the NPSEM-IE assumption requires:

$$Z \perp\!\!\!\perp \{M(z) \text{ for all values of } z\} \perp\!\!\!\perp \{Y(m) \text{ for all values of } m\}.$$

3.3 The template $\mathcal{G}(\mathbf{a})$ for intervention on set A

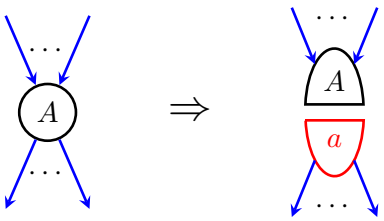
We now describe the procedure for constructing a Single-World Intervention Template (SWIT), denoted $\mathcal{G}(\mathbf{a})$. An instantiation of the template, will be a Single-World Intervention Graph (SWIG), denoted $\mathcal{G}(\tilde{\mathbf{a}})$. The node set for $\mathcal{G}(\tilde{\mathbf{a}})$ consists of the set of potential outcome variables corresponding to a single hypothetical intervention setting $\mathbf{A}$ to $\tilde{\mathbf{a}}$, together with a set of fixed vertices $\tilde{\mathbf{a}}$. $\mathcal{G}(\tilde{\mathbf{a}})$ encodes counterfactual independence relations that will

hold among the set of potential outcome variables $\mathbb{V}(\tilde{\mathbf{a}})$ under an NPSEM obeying the FFRCISTG assumption (17). Since the NPSEM-IE independence assumption (18) implies the FFRCISTG assumption (17), it further follows that any independence relation given by $\mathcal{G}(\tilde{\mathbf{a}})$ will also hold under this submodel.

3.3.1 Construction of Single-world Counterfactual Templates (Individual-level exclusions)

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



Splitting: Schematic Illustrating the Splitting of Node A

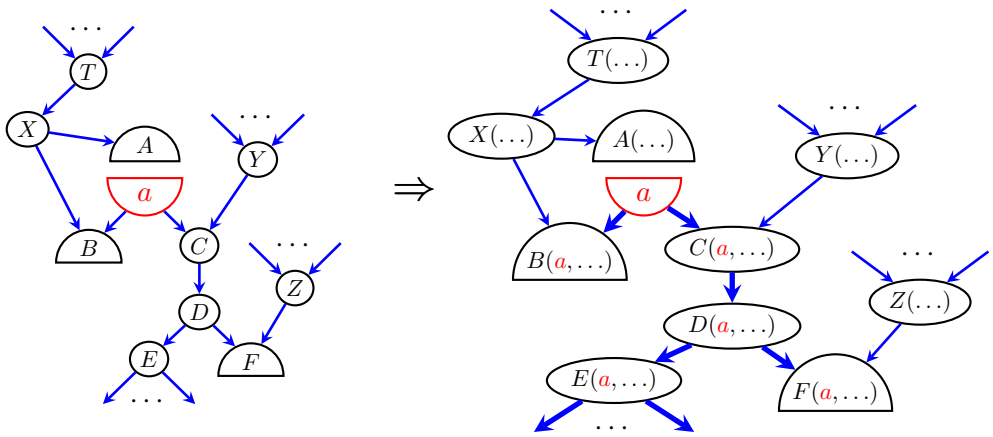
Thus the random half inherits all edges directed into A in $\mathcal{G}$; the fixed half inherits all edges directed out of A.

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 $V(\mathbf{a}_V)$ (see the schematic below).

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 $\mathcal{G}(\mathbf{a})$. 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 $\mathbf{A}$.



labeling: Schematic showing the nodes $V(\mathbf{a}_V)$ in $\mathcal{G}(\mathbf{a})$ for which $a \in \mathbf{a}_V$.

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 corresponding output.

3.3.2 Examples of SWITs

We illustrate the construction of templates via three examples in Figure 9, a complete DAG corresponding to the situation where M 'mediates' the effect of a treatment Z on a response Y, and there is no confounding, and three in Figure 10, where we further assume that there is 'no direct effect' of Z on Y, at the individual level, so that $Y(\tilde{z}, \tilde{m}) = Y(\tilde{m})$.

The subsets of $\mathbf{a}$ labeling each vertex are summarized in Table 1. Note that in the templates in Figure 9, since Y is a descendant (in fact, child) of both Z and M in all three graphs, Y is labelled with all variables that are intervened on. Thus in template (iv) in Figure 9 $\mathbf{a}_Y = \mathbf{a} = \{z, m\}$. In contrast, in template (iv) in Figure 10, where both Z and M are intervened on, so $\mathbf{a} = \{z, m\}$, we have $\mathbf{a}_Y = \{m\}$ since after splitting M, Y is no longer a descendant of Z.

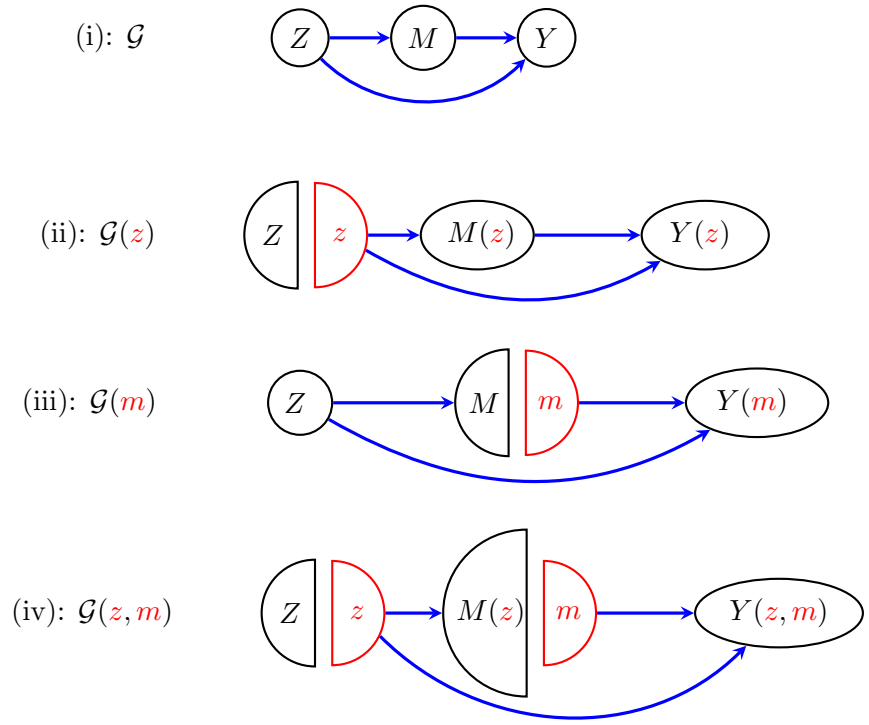


Figure 9: (i) A DAG $\mathcal{G}$ with treatment (Z), mediator (M) and response (Y) in the absence of confounding. Templates: (ii) $\mathcal{G}(z)$; (iii) $\mathcal{G}(m)$; (iv) $\mathcal{G}(z,m)$.

	$\mathbf{a}$	Figure 9	Figure 10
(ii)	$\{z\}$	$\mathbf{a}_Z=\emptyset;\ \mathbf{a}_M=\mathbf{a}_Y=\{z\}$	$\mathbf{a}_Z=\emptyset;\ \mathbf{a}_M=\mathbf{a}_Y=\{z\}$
(iii)	$\{m\}$	$\mathbf{a}_Z=\mathbf{a}_M=\emptyset;\ \mathbf{a}_Y=\{m\}$	$\mathbf{a}_Z=\mathbf{a}_M=\emptyset;\ \mathbf{a}_Y=\{m\}$
(iv)	$\{z,m\}$	$\mathbf{a}_Z=\emptyset;\ \mathbf{a}_M=\{z\};\ \mathbf{a}_Y=\{z,m\}$	$\mathbf{a}_Z=\emptyset;\ \mathbf{a}_M=\{z\};\ \mathbf{a}_Y=\{m\}$

Table 1: The sets labeling nodes in the SWITs in Figures 9 and 10.

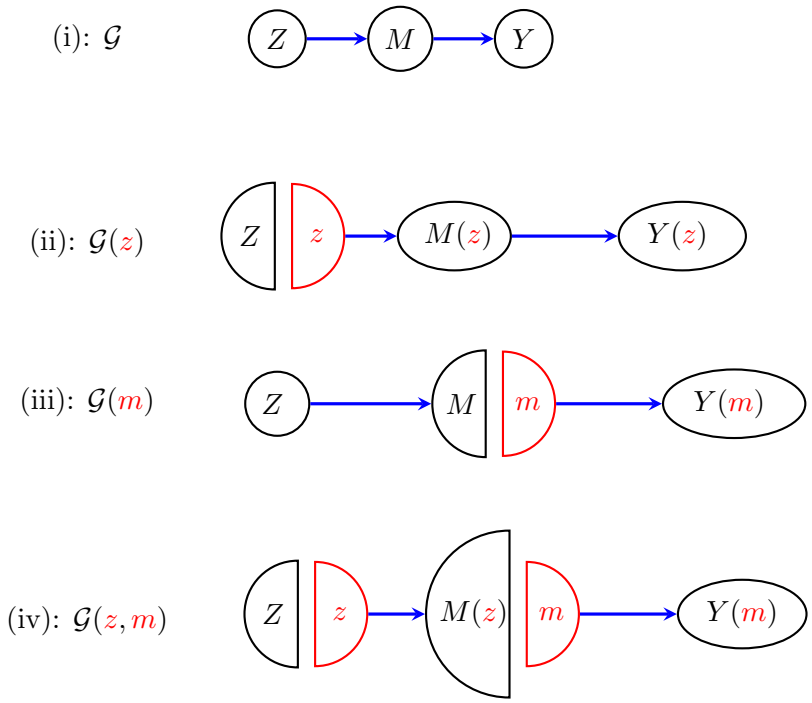


Figure 10: (i) The DAG $\mathcal{G}$ from Figure 9 under the additional assumption that there is no direct effect of treatment (Z) on the response (Y). Templates: (ii) $\mathcal{G}(z)$; (iii) $\mathcal{G}(m)$; (iv) $\mathcal{G}(z, m)$.

3.3.3 Graphical Properties of SWITs

The next three Propositions follow immediately from the construction procedure.

Proposition 4. *In $\mathcal{G}(\mathbf{a})$, every node $Y(\mathbf{a}_Y)$ that is a descendant of a fixed node a_i, is labelled with a_i, so $a_i \in \mathbf{a}_Y$.*

Proposition 5. *There is a one-to-one correspondence between random vertices in $\mathcal{G}(\mathbf{a})$ and vertices in $\mathcal{G}$ given by $Y \mapsto Y(\mathbf{a}_Y)$.*

Proposition 6. *There is a one-to-one correspondence between edges in $\mathcal{G}$ and edges in $\mathcal{G}(\mathbf{a})$:*

$$X \to Y \text{ in } \mathcal{G} \qquad \mapsto \qquad \begin{cases} x \to Y(\mathbf{a}_Y) & \text{in } \mathcal{G}(\mathbf{a}) \quad \text{if } X \in A; \\ X(\mathbf{a}_X) \to Y(\mathbf{a}_Y) & \text{in } \mathcal{G}(\mathbf{a}) \quad \text{if } X \notin A. \end{cases} \tag{19}$$

(Recall that it is possible that $\mathbf{a}_V = \emptyset$, in which case $V(\mathbf{a}_V) = V$.)

Proof: This follows since neither the process of node splitting nor labeling removes any edges that were present in $\mathcal{G}$. □

In what follows we will use blackboard bold, e.g. $\mathbb{B}(\mathbf{a})$ to refer to the set of random vertices in $\mathcal{G}(\mathbf{a})$ corresponding to a given set $\mathbf{B}$ in $\mathcal{G}$, so that $\mathbb{B}(\mathbf{a}) \equiv \{B(\mathbf{a}_B) \mid B \in \mathbf{B}\}$, and similarly for other sets.

When a node A is split, those vertices that, prior to the transformation had A as a parent, now have the fixed vertex a as a parent. It thus follows that in $\mathcal{G}(\mathbf{a})$ the parents of a random vertex $V(\mathbf{a}_V)$ that are *not* fixed, correspond to those parents of V in $\mathcal{G}$ that are *not* in $\mathbf{A}$:

Proposition 7. *In $\mathcal{G}(\mathbf{a})$, the set of parents of $Y(\mathbf{a}_Y)$ that are random consists of the vertices in $\mathcal{G}(\mathbf{a})$ corresponding to vertices in $\mathrm{pa}_{\mathcal{G}}(Y) \setminus \mathbf{A}$.*[29] *This may be expressed as follows:*

$$\{V(\mathbf{a}_V) \mid V(\mathbf{a}_V) \to Y(\mathbf{a}_Y) \text{ in } \mathcal{G}(\mathbf{a})\} = \mathbb{pa}_{\mathcal{G},Y}(\mathbf{a}) \setminus \mathbb{A}(\mathbf{a}),$$

where: $\mathrm{pa}_{\mathcal{G},Y} \equiv \mathrm{pa}_{\mathcal{G}}(Y)$, $\mathbb{pa}_{\mathcal{G},Y}(\mathbf{a}) \equiv \{V(\mathbf{a}_V) \mid V \in \mathrm{pa}_{\mathcal{G}}(Y)\}$, *and* $\mathbb{A}(\mathbf{a}) \equiv \{V(\mathbf{a}_V) \mid V \in \mathbf{A}\}$. *Equivalently it may be written:*

$$\mathbb{pa}_{\mathcal{G},Y}(\tilde{\mathbf{a}}) \setminus \mathbb{A}(\tilde{\mathbf{a}}) = \{V(\tilde{\mathbf{a}}_V) \mid V \in \mathrm{pa}_{\mathcal{G}}(Y) \setminus \mathbf{A}\}.$$

[29] Note that in our notation for any graph $\mathcal{G}$, $\mathrm{pa}_{\mathcal{G}}(Y)$ is a set of random variables, the parents of Y in $\mathcal{G}$. In particular although $\mathrm{pa}(\cdot)$ is lower case this is a set of random variables, not an instantiation of this set.

We will also need the following:

Proposition 8. *If* $X(\mathbf{a}_X) \to \cdots \to Y(\mathbf{a}_Y)$ *in* $\mathcal{G}(\mathbf{a})$ *then* $\mathbf{a}_X \subseteq \mathbf{a}_Y$.

Proof: Immediate by the definition of $\mathbf{a}_V$ since any fixed vertex that is an ancestor of $X(\mathbf{a}_X)$ in $\mathcal{G}(\mathbf{a})$ is also an ancestor of $Y(\mathbf{a}_Y)$. □

For examples of this Proposition where $\mathbf{a}_X \subset \mathbf{a}_Y$ see the nodes the nodes L and $Y(x)$ in Figure 5(b), also $L_2(a_1)$ and $Y(a_1, a_2)$ in Figure 14(b) below; for examples where $\mathbf{a}_X = \mathbf{a}_Y$ see the nodes $M(z)$ and $Y(z)$ in Figures 9(ii) and 10(ii) below.

Proposition 9. *In* $\mathcal{G}$*, let* A_i *be the vertex in* $\mathcal{G}$ *corresponding to a fixed node* $a_i \in \mathbf{a}_V$*, the set labeling* $V(\mathbf{a}_V)$ *in the SWIT* $\mathcal{G}(\mathbf{a})$*. Then* A_i *is an ancestor of* V *in* $\mathcal{G}$*; further, there is a directed path from* A_i *to* V *on which no other vertex is in* $\mathbf{A}$*.*

Proof: If $a_i \in \mathbf{a}_V$ then there is a directed path from a_i to V in $\mathcal{G}^*$, the graph resulting from the node splitting step. Since this directed path is still present in $\mathcal{G}^*$, it was also present in $\mathcal{G}$, and no node on that path is in $\mathbf{A}$, since otherwise the directed path would not be present in $\mathcal{G}^*$. □

3.4 The interpretation of node labels in $\mathcal{G}(\tilde{\mathbf{a}})$

Formally, a vertex containing $V(\tilde{\mathbf{a}}_V)$ in a SWIG $\mathcal{G}(\tilde{\mathbf{a}})$ should be viewed as representing an equivalence class of counterfactual variables that are equal for every unit in the population. Specifically, the node $V(\tilde{\mathbf{a}}_V)$ in the SWIG $\mathcal{G}(\tilde{\mathbf{a}})$ represents the set:

$$\left\{ V(\tilde{\mathbf{a}}_V, \mathbf{b}^\dagger) \mid \mathbf{B} \subseteq (\mathbf{A} \setminus (\mathbf{A}_V \cup \{V\})) \right\}. \tag{20}$$

Thus, in words, $V(\tilde{\mathbf{a}}_V)$ in $\mathcal{G}(\tilde{\mathbf{a}})$ represents the set of counterfactual variables resulting from assigning $\tilde{\mathbf{a}}_V$ to the variables in the subset of treatments $\mathbf{A}_V$ that are (proper) ancestors of $V(\mathbf{a}_V)$ in $\mathcal{G}(\mathbf{a})$ and assigning any other value to any subset of the other treatments in $\mathbf{A}$.

The set (20) captures equalities that arise both from the constraints of time order, i.e. that later treatments cannot affect earlier counterfactuals, and from the absence of direct effects, relative to the other treatment variables in $\mathbf{A}$, as reflected in the set of structural equations.

The subset $\mathbf{A}_V \subseteq \mathbf{A}$ is the minimal subset of treatments sufficient to ensure that $V(\mathbf{a}^\dagger) = V(\mathbf{a}^\dagger_V)$ for every $\mathbf{a}^\dagger$ and every unit of the populations. There may be situations in which we might find it useful to label a node with

another variable from the class (20); however, in this paper we will always use the minimal subset as the label. This allows us to unify the semantics for SWIGs representing individual level and population level exclusion restrictions, given in §7. (Under the population interpretation, equalities between counterfactuals will only arise due to considerations of time-order.)

For most purposes it is sufficient to think of a node $V(\tilde{\mathbf{a}}_V)$ in a SWIG as (solely) representing $V(\tilde{\mathbf{a}}_V)$.

3.4.1 Examples of equalities implied by node labels

Consider a SWIG $\mathcal{G}(\tilde{m})$ that is an instantiation of the template in Figure 9(iii). Formally the node Z in $\mathcal{G}(\tilde{m})$ represents the equivalence class of variables

$$\{Z\} \cup \{Z(m'), \text{ for any } m'\}. \tag{21}$$

This set of variables may be seen as capturing the requirement that the future does not affect the past.

For a second example, consider a SWIG $\mathcal{G}(\tilde{z}, \tilde{m})$ that is an instantiation of the template in Figure 10(iv), derived from a DAG in which there is no $Z \to Y$ edge. In $\mathcal{G}(\tilde{z}, \tilde{m})$ the node labelled $Y(\tilde{m})$ represents the equivalence class of variables:

$$\{Y(\tilde{m})\} \cup \{Y(z', \tilde{m}), \text{ for any } z'\}. \tag{22}$$

The equality of the counterfactuals in this class follows from the assumption present in the original NPSEM that there is no individual level direct effect of Z on Y.

Note that under an NPSEM model the equality (for every unit of the population) of the variables within each of the classes (21) and (22) follows from the equations given in §3.1.1, which were themselves consequences of Definition 1(ii).

3.5 The factorization associated with $\mathcal{G}(\tilde{\mathbf{a}})$

We associate the following factorization criterion with a SWIG:

Definition 10. *A joint distribution* $P(\mathbb{V}(\tilde{\mathbf{a}}))$ factorizes with respect to a SWIG $\mathcal{G}(\tilde{\mathbf{a}})$ *if*

$$P\left(\mathbb{V}(\tilde{\mathbf{a}})\right) = \prod_{Y \in \mathbf{V}} P\left(Y(\tilde{\mathbf{a}}_Y) \,\middle|\, \mathbb{pa}_{\mathcal{G},Y}(\tilde{\mathbf{a}}) \setminus \mathbb{A}(\tilde{\mathbf{a}})\right), \tag{23}$$

whenever the right hand side is well-defined. Here

$$\mathbb{pa}_{\mathcal{G},Y}(\tilde{\mathbf{a}}) \setminus \mathbb{A}(\tilde{\mathbf{a}}) = \{V(\tilde{\mathbf{a}}_V) \mid V \in \mathrm{pa}_{\mathcal{G}}(Y) \setminus \mathbf{A}\}.$$

In words, this definition means that the distribution over the random vertices $\mathbb{V}(\tilde{\mathbf{a}})$ in $\mathcal{G}(\tilde{\mathbf{a}})$ factors into a product of conditional densities, one for each random vertex. The density for $Y(\tilde{\mathbf{a}}_Y)$ conditions on the set of random vertices $V(\tilde{\mathbf{a}}_V)$ that correspond to vertices V in $\mathcal{G}$, that are not in $\mathbf{A}$, but are parents of Y. Observe that in (23) the parents of Y that are in $\mathbf{A}$ (and hence are intervened upon in $\mathcal{G}(\tilde{\mathbf{a}})$) have been removed from the conditioning set for $Y(\tilde{\mathbf{a}}_Y)$ in $\mathcal{G}(\tilde{\mathbf{a}})$.

Note that the factorization (23) corresponds exactly to the factorization associated with an instantiation $\mathcal{G}(\tilde{\mathbf{a}})$ of $\mathcal{G}(\mathbf{a})$, if we were to remove the fixed vertices, but otherwise view it as a conventional DAG model for the counterfactual distribution; see Appendix A for a review. To make this clear, notice that we may re-write (23) as:

$$P\left(\mathbb{V}(\tilde{\mathbf{a}})\right) = \prod_{Y(\tilde{\mathbf{a}}_Y) \in \mathbb{V}(\tilde{\mathbf{a}})} P\left(Y(\tilde{\mathbf{a}}_Y) \,\middle|\, \mathrm{pa}_{\mathcal{G}(\tilde{\mathbf{a}})}(Y(\tilde{\mathbf{a}}_Y)) \setminus \tilde{\mathbf{a}}\right).$$

Also notice that by taking $\mathbf{A} = \emptyset$, Definition 10 includes the usual factorization of $P(\mathbf{V})$ with respect to $\mathcal{G}$ as a special case; see (104).

The relationship between the FFRCISTG model and the factorization (23) is given by the following:

Proposition 11. *Under the FFRCISTG model associated with $\mathcal{G}$ $P(\mathbb{V}(\mathbf{a}^\dagger))$ factorizes according to $\mathcal{G}(\mathbf{a}^\dagger)$.*

We prove this in conjunction with Proposition 16 that establishes the modularity property; see Appendix B.1.

3.5.1 Examples of the Factorization

Instantiations of the templates shown in Figure 9 are associated with the following factorizations (we include the original factorization for comparison):

$$\begin{aligned}
\mathcal{G}\text{:} \quad & P(Z,M,Y) &&= P(Z)P(M \mid Z)P(Y \mid Z,M);\\
\mathcal{G}(\tilde{z})\text{:} \quad & P(Z,M(\tilde{z}),Y(\tilde{z})) &&= P(Z)P(M(\tilde{z}))P(Y(\tilde{z}) \mid M(\tilde{z}));\\
\mathcal{G}(\tilde{m})\text{:} \quad & P(Z,M,Y(\tilde{m})) &&= P(Z)P(M \mid Z)P(Y(\tilde{m}) \mid Z);\\
\mathcal{G}(\tilde{z},\tilde{m})\text{:} \quad & P(Z,M(\tilde{z}),Y(\tilde{z},\tilde{m})) &&= P(Z)P(M(\tilde{z}))P(Y(\tilde{z},\tilde{m})).
\end{aligned}$$

Instantiations of the templates shown in Figure 10 give the following:

$$\begin{aligned}
\mathcal{G}\text{:} \quad & P(Z,M,Y) &=\ & P(Z)P(M \mid Z)P(Y \mid M);\\
\mathcal{G}(\tilde{z})\text{:} \quad & P(Z,M(\tilde{z}),Y(\tilde{z})) &=\ & P(Z)P(M(\tilde{z}))P(Y(\tilde{z}) \mid M(\tilde{z}));\\
\mathcal{G}(\tilde{m})\text{:} \quad & P(Z,M,Y(\tilde{m})) &=\ & P(Z)P(M \mid Z)P(Y(\tilde{m}));\\
\mathcal{G}(\tilde{z},\tilde{m})\text{:} \quad & P(Z,M(\tilde{z}),Y(\tilde{m})) &=\ & P(Z)P(M(\tilde{z}))P(Y(\tilde{m})).
\end{aligned}$$

Notice that in all of these examples the conditioning set in the density associated with $V(\tilde{\mathbf{a}}_V)$ in $\mathcal{G}(\tilde{\mathbf{a}})$ is a set of counterfactual random variables that correspond to a subset of the parents of V in $\mathcal{G}$; further any parent in $\mathcal{G}$ for which a corresponding variable is not present in the conditioning set, has been intervened upon and thus is present in $\tilde{\mathbf{a}}_V$.

3.5.2 The Markov property for $\mathcal{G}(\tilde{\mathbf{a}})$

It immediately follows from standard results in graphical models (e.g. Lauritzen, 1996, Thm. 3.27) that if a distribution factorizes according to $\mathcal{G}(\tilde{\mathbf{a}})$ then the distribution $P(\mathbb{V}(\tilde{\mathbf{a}}))$ also obeys the following Markov property:

Given disjoint subsets $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ of $\mathbf{V}$ ($\mathbf{D}$ may be empty), let $\mathbb{B}(\tilde{\mathbf{a}}) \equiv \{B(\tilde{\mathbf{a}}_B), B \in \mathbf{B}\}$, $\mathbb{C}(\tilde{\mathbf{a}}) \equiv \{C(\tilde{\mathbf{a}}_C), C \in \mathbf{C}\}$, $\mathbb{D}(\tilde{\mathbf{a}}) \equiv \{D(\tilde{\mathbf{a}}_D), D \in \mathbf{D}\}$, be the corresponding sets in $\mathcal{G}(\tilde{\mathbf{a}})$,

$$\begin{aligned}
&\text{if } \mathbb{B}(\tilde{\mathbf{a}}) \text{ is d-separated from } \mathbb{C}(\tilde{\mathbf{a}}) \text{ given } \mathbb{D}(\tilde{\mathbf{a}}) \cup \tilde{\mathbf{a}} \text{ in } \mathcal{G}(\tilde{\mathbf{a}}) \qquad (24)\\
&\qquad \text{then} \quad \mathbb{B}(\tilde{\mathbf{a}}) \perp\!\!\!\perp \mathbb{C}(\tilde{\mathbf{a}}) \mid \mathbb{D}(\tilde{\mathbf{a}}) \quad [P(\mathbb{V}(\tilde{\mathbf{a}}))].
\end{aligned}$$

In words, this states that if in $\mathcal{G}(\tilde{\mathbf{a}})$ two 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 $P(\mathbb{V}(\tilde{\mathbf{a}}))$. Thus we have the following:

Theorem 12. *If $P(\mathbb{V}(\tilde{\mathbf{a}}))$ factors according to $\mathcal{G}(\tilde{\mathbf{a}})$ then $P(\mathbb{V}(\tilde{\mathbf{a}}))$ obeys (24).*

The requirement to also condition on the vertices in $\tilde{\mathbf{a}}$ is intuitive when it is considered that, in a given instantiation, these are fixed constants. Note that since, by construction, vertices in $\tilde{\mathbf{a}}$ have no parents, no additional paths become d-connecting due to conditioning on the nodes in $\tilde{\mathbf{a}}$. Consequently we have the following:

Proposition 13. *If $\mathbb{B}(\tilde{\mathbf{a}})$ and $\mathbb{C}(\tilde{\mathbf{a}})$ are d-separated given $\mathbb{D}(\tilde{\mathbf{a}})$ in $\mathcal{G}(\tilde{\mathbf{a}})$ then $\mathbb{B}(\tilde{\mathbf{a}})$ and $\mathbb{C}(\tilde{\mathbf{a}})$ are d-separated given $\mathbb{D}(\tilde{\mathbf{a}}) \cup \tilde{\mathbf{a}}$ in $\mathcal{G}(\tilde{\mathbf{a}})$.*

It also follows from the above observation that:

Proposition 14. *$\mathbb{B}(\tilde{\mathbf{a}})$ is d-separated from $\mathbb{C}(\tilde{\mathbf{a}})$ given $\mathbb{D}(\tilde{\mathbf{a}})\cup\tilde{\mathbf{a}}$ in $\mathcal{G}(\tilde{\mathbf{a}})$ if and only if $\mathbb{B}(\tilde{\mathbf{a}})$ is d-separated from $\mathbb{C}(\tilde{\mathbf{a}})$ given $\mathbb{D}(\tilde{\mathbf{a}})$ in $(\mathcal{G}(\tilde{\mathbf{a}}))_{\mathbb{V}(\tilde{\mathbf{a}})}$, the induced subgraph of $\mathcal{G}(\tilde{\mathbf{a}})$ on $\mathbb{V}(\tilde{\mathbf{a}})$, obtained by removing all fixed nodes $\tilde{\mathbf{a}}$.*

Thus an alternative approach to implicitly conditioning on the fixed nodes $\tilde{\mathbf{a}}$ is to simply remove the fixed vertices from $\mathcal{G}(\tilde{\mathbf{a}})$ before checking d-separation relations. Notwithstanding this, the fixed nodes play an important role in that they make it possible to recover the original DAG from $\mathcal{G}(\tilde{\mathbf{a}})$ alone, see §3.6.3; they are also instrumental in studying dynamic regimes, see §5.

3.5.3 Examples of the Markov property

Consider first the Markov properties associated with the SWITs shown in Figure 9.

$\mathcal{G}(\tilde{z})$: Applying d-separation we see that for each value of $\tilde{z}$,

$$Z \perp\!\!\!\perp M(\tilde{z}), Y(\tilde{z}); \tag{25}$$

which may be readily seen from the fact that there is no path joining Z and $M(\tilde{z})$, $Y(\tilde{z})$. We stress that it does not follow from this template that Z is jointly independent of $Y(\tilde{z})$ and $Y(z^*)$, nor that Z is jointly independent of $Y(\tilde{z})$ and $M(z^*)$, for distinct values $\tilde{z}, z^*$. These 'cross-world' independence relations do not logically follow from $Z \perp\!\!\!\perp M(z^*), Y(z^*)$ and $Z \perp\!\!\!\perp Y(\tilde{z})$.

$\mathcal{G}(\tilde{m})$: This template implies that for each $\tilde{m}$,

$$M \perp\!\!\!\perp Y(\tilde{m}) \mid Z; \tag{26}$$

which may be seen from the fact that Z is a non-collider on the path from M to $Y(\tilde{m})$.

$\mathcal{G}(\tilde{z}, \tilde{m})$: In this example we see that for each $\tilde{m}$ and $\tilde{z}$,

$$Z \perp\!\!\!\perp M(\tilde{z}), Y(\tilde{m}, \tilde{z}) \quad \text{and} \quad M(\tilde{z}) \perp\!\!\!\perp Y(\tilde{m}, \tilde{z}) \mid Z. \tag{27}$$

The first of these independence relations follows, as before, from the fact that there are no paths from Z to $M(\tilde{z})$ and $Y(\tilde{m}, \tilde{z})$. The latter conditional independence follows when it is recalled that when judging independence we always condition on fixed nodes, such as $\tilde{z}$, so that there is no path d-connecting $M(\tilde{z})$ and $Y(\tilde{m}, \tilde{z})$ given Z.

We now consider the templates shown in Figure 10 where there is no 'direct effect' of Z on Y. These lead to the following independence relations:

$\mathcal{G}(\tilde{z})$: For each value of $\tilde{z}$,

$$Z \perp\!\!\!\perp M(\tilde{z}), Y(\tilde{z}). \tag{28}$$

Note that this is the same independence constraint (25) implied by the template shown in Figure 9(ii). However, we will see that in conjunction with the modularity property introduced in Section 3.6, the absence of the edge $z \to Y(z)$ in the template implies an additional constraint relating the distribution $P(Z, M(\tilde{z}), Y(\tilde{z}))$ to $P(Z, M(z'), Y(z'))$ for $\tilde{z} \neq z'$.

$\mathcal{G}(\tilde{m})$: This template implies that for each $\tilde{m}$,

$$Z, M \perp\!\!\!\perp Y(\tilde{m}); \tag{29}$$

which may be seen from the fact that there are no paths from Z and M to $Y(\tilde{m})$.

$\mathcal{G}(\tilde{z}, \tilde{m})$: It follows from this template that the variables Z, $M(\tilde{z})$ and $Y(\tilde{m})$ are mutually independent.

3.6 Modularity

In general the densities occurring on the RHS of (23) involve potential outcome variables (the exception being those variables for which $\mathbf{a}_V = \emptyset$, which includes any variable that has no parents in $\mathcal{G}$). The modularity property links these conditional densities to the conditional densities $P(V \mid \mathrm{pa}(V))$ associated with the factual variables present in the original DAG.

Definition 15 (Modularity)**.**
Pairs $(\mathcal{G}, P(\mathbf{V}))$ *and* $(\mathcal{G}(\tilde{\mathbf{a}}), P(\mathbb{V}(\tilde{\mathbf{a}})))$ *are said to satisfy the* modularity property *if for every* $Y \in \mathbf{V}$*,*

$$\begin{aligned} P\left(Y(\tilde{\mathbf{a}}_Y) = y \,\middle|\, \left(\mathrm{pa}_{\mathcal{G}(\tilde{\mathbf{a}})}(Y(\tilde{\mathbf{a}}_Y)) \setminus \tilde{\mathbf{a}}\right) = \mathbf{q}\right) \\ = P\Big(Y = y \,\Big|\, \left(\mathrm{pa}_{\mathcal{G}}(Y) \setminus \mathbf{A}\right) = \mathbf{q},\ \left(\mathrm{pa}_{\mathcal{G}}(Y) \cap \mathbf{A}\right) = \tilde{\mathbf{a}}_{\mathrm{pa}_{\mathcal{G}}(Y) \cap \mathbf{A}}\Big), \end{aligned} \tag{30}$$

whenever both sides of (30) *are well-defined.*

In words, this states that the conditional density associated with the random vertex $Y(\tilde{\mathbf{a}}_Y)$ in $\mathcal{G}(\tilde{\mathbf{a}})$ corresponds to the conditional density associated with Y in $\mathcal{G}$ after substituting in the value $\tilde{a}_i$ for any variable $A_i \in \mathbf{A}$

that is a parent of Y. Recall from Proposition 7 that $\mathrm{pa}_{\mathcal{G}(\tilde{\mathbf{a}})}(Y(\tilde{\mathbf{a}}_Y)) \setminus \tilde{\mathbf{a}}$ is simply the set of random nodes in $\mathcal{G}(\tilde{\mathbf{a}})$ that are parents of $Y(\tilde{\mathbf{a}}_Y)$.

Thus formally the modularity property for $\mathcal{G}(\tilde{\mathbf{a}})$ is a relation between two pairs $(P(\mathbf{V}), \mathcal{G})$ and $(P(\mathbb{V}(\tilde{\mathbf{a}})), \mathcal{G}(\tilde{\mathbf{a}}))$, where in each pair the distribution factors according to the graph.

Proposition 16. *Under the FFRCISTG model associated with $\mathcal{G}$, the pairs $(P(\mathbf{V}), \mathcal{G})$ and $\left(P(\mathbb{V}(\mathbf{a}^\dagger)), \mathcal{G}(\mathbf{a}^\dagger)\right)$ obey modularity.*

We prove this in Appendix B.1.

3.6.1 Examples of the Modularity Property

We illustrate the modularity property for the examples considered previously. For the SWITs in Figure 9 we have the following equations:

$\mathcal{G}(\tilde{z})$: Modularity implies that for all values $\tilde{z}, m, y$:

$$P(M(\tilde{z}) = m) = P(M = m \mid Z = \tilde{z}),$$
$$P(Y(\tilde{z}) = y \mid M(\tilde{z}) = m) = P(Y = y \mid M = m, Z = \tilde{z}).$$

$\mathcal{G}(\tilde{m})$: Here modularity implies that for all $z, \tilde{m}, y$

$$P(Y(\tilde{m}) = y \mid Z = z) = P(Y = y \mid M = \tilde{m}, Z = z).$$

$\mathcal{G}(\tilde{z}, \tilde{m})$: Modularity implies that for all values $\tilde{z}, m, \tilde{m}, y$:

$$P(M(\tilde{z}) = m) = P(M = m \mid Z = \tilde{z}),$$
$$P(Y(\tilde{z}, \tilde{m}) = y) = P(Y = y \mid M = \tilde{m}, Z = \tilde{z}).$$

Turning to the templates shown in Figure 10, we have the following:

$\mathcal{G}(\tilde{z})$: Under modularity the template here implies for all values $\tilde{z}, m, y$:

$$P(M(\tilde{z}) = m) = P(M = m \mid Z = \tilde{z}),$$
$$P(Y(\tilde{z}) = y \mid M(\tilde{z}) = m) = P(Y = y \mid M = m).$$

Notice that the second equation here implies the restriction that $P(Y(\tilde{z}) = y \mid M(\tilde{z}) = m)$ does not depend on the value of $\tilde{z}$, hence we have that for $z' \neq \tilde{z}$,

$$P(Y(\tilde{z}) = y \mid M(\tilde{z}) = m) = P(Y(z') = y \mid M(z') = m).$$

Notice that this is a restriction relating the distributions $P(Z, M(\tilde{z}), Y(\tilde{z}))$ and $P(Z, M(z'), Y(z'))$ that are associated with two *different* instantiations of the template: $\mathcal{G}(\tilde{z})$ and $\mathcal{G}(z')$.

$\mathcal{G}(\tilde{m})$: For this template, modularity implies that for all $\tilde{m}, y$

$$P(Y(\tilde{m})=y) = P(Y=y \mid M=\tilde{m}).$$

$\mathcal{G}(\tilde{z},\tilde{m})$: In this case, modularity implies that for all values $\tilde{z}, m, \tilde{m}, y$:

$$P(M(\tilde{z})=m) = P(M=m \mid Z=\tilde{z}),$$
$$P(Y(\tilde{m})=y) = P(Y=y \mid M=\tilde{m}).$$

3.6.2 How the FFRCISTG implies modularity and factorization

Given the FFRCISTG model associated with a graph $\mathcal{G}$ a simple argument establishes that, for a given SWIG $\mathcal{G}(\tilde{\mathbf{a}})$, the distribution of the counterfactuals $P(\mathbb{V}(\tilde{\mathbf{a}}))$ factorizes with respect to $\mathcal{G}(\tilde{\mathbf{a}})$. We illustrate this for the FFRCISTG model associated with the DAG $\mathcal{G}$ in Figure 10(i), showing the three distributions: $P(Z, M(\tilde{z}), Y(\tilde{m}))$, $P(Z, M, Y(\tilde{m}))$ and $P(Z, M, Y)$ factorize with respect to $\mathcal{G}(\tilde{z},\tilde{m})$, $\mathcal{G}(\tilde{m})$ and $\mathcal{G}$ respectively; see Figure 10.

First observe that the FFRCISTG assumption (17) implies that the counterfactuals Z, $M(\tilde{z})$ and $Y(\tilde{m})$ are jointly independent, so:

$$P(Z=z, M(\tilde{z})=m, Y(\tilde{m})=y) = P(Z=z)P(M(\tilde{z})=m)P(Y(\tilde{m})=y). \quad (31)$$

This establishes the factorization property for $\mathcal{G}(\tilde{z},\tilde{m})$.

Next we turn to the factorization property for $\mathcal{G}$. Observe the following:

$$P(M(\tilde{z})=m) = P(M(\tilde{z})=m \mid Z=\tilde{z}) \;=\; P(M=m \mid Z=\tilde{z}), \quad (32)$$

where the first equality follows from factorization (31) and the second uses consistency (14). Substituting into the RHS of (31) we obtain:

$$P(Z=z, M(\tilde{z})=m, Y(\tilde{m})=y) = P(Z=z)P(M=m \mid Z=\tilde{z})P(Y(\tilde{m})=y). \quad (33)$$

Evaluating (33) with $z = \tilde{z}$ implies:

$$P(Z=z, M=m, Y(\tilde{m})=y) = P(Z=z)P(M=m \mid Z=z)P(Y(\tilde{m})=y), \quad (34)$$

where we have used recursive substitution on the LHS since $Z=z$ implies $M(z) = M$. This is the factorization property for $\mathcal{G}(\tilde{m})$.

Lastly, by the same argument used to establish (32):

$$P(Y(\tilde{m})=y) = P(Y(\tilde{m})=y \mid M=\tilde{m}) \;=\; P(Y=y \mid M=\tilde{m}). \quad (35)$$

Substituting into the RHS of (34) and evaluating at $m = \tilde{m}$ implies:

$$P(Z{=}z, M{=}m, Y{=}y) = P(Z{=}z)P(M{=}m \mid Z{=}z)P(Y{=}y \mid M{=}m), \tag{36}$$

so that $P(Z, M, Y)$ factors according to $\mathcal{G}$.

Inspection of equations (32) and (35) reveals that we have also established that $(P(Z, M(\tilde{z}), Y(\tilde{m})), \mathcal{G}(\tilde{z}, \tilde{m}))$ and $(P(Z, M, Y(\tilde{m})), \mathcal{G}(\tilde{m}))$ both obey modularity with respect to $(P(Z, M, Y), \mathcal{G})$.

A proof is given in Appendix B.1, showing that the above argument extends to an arbitrary $\mathcal{G}$ thus proving Propositions 11 and 16.

3.6.3 The rationale for including fixed nodes

The reader may be curious as to the motivation for including fixed vertices on the SWIG. As Proposition 14 shows, the d-separation relations holding among the variables in $P(\mathbb{V}(\tilde{\mathbf{a}}))$ are given by the induced subgraph of $\mathcal{G}(\tilde{\mathbf{a}})$ obtained by removing fixed nodes.

However, if we compare the two templates $\mathcal{G}(\tilde{z})$ in Figures 9(ii) and 10(ii) we see that if the fixed node z were omitted the two graphs would be identical. Thus if we wish to be able to infer the original DAG $\mathcal{G}$ from $\mathcal{G}(\tilde{\mathbf{a}})$ alone, then it is necessary to include fixed nodes.

The ability to discern the DAG $\mathcal{G}$ is important since the modularity condition relates $\mathcal{G}$ and $\mathcal{G}(\tilde{\mathbf{a}})$. Returning to the example above, the conditional densities associated with $Y(\tilde{z})$ under modularity are different in Figures 9(ii) and 10(ii): if the original DAG contains the $Z \to Y$ edge, then we have:

$$P(Y(\tilde{z}){=}y \mid M(\tilde{z}){=}m) = P(Y{=}y \mid M{=}m, Z{=}\tilde{z})$$

while if the $Z \to Y$ edge is absent then we have:

$$P(Y(\tilde{z}){=}y \mid M(\tilde{z}){=}m) = P(Y{=}y \mid M{=}m).$$

This point is of particular relevance in the analysis of dynamic regimes; see §5.

3.6.4 Implications of Modularity

The following is a direct consequence of the factorization (23) and modularity (30) properties.

Proposition 17. *If $P(\mathbf{V})$ factorizes according to $\mathcal{G}$ then $P(\mathbb{V}(\tilde{\mathbf{a}}))$ factorizes according to $\mathcal{G}(\tilde{\mathbf{a}})$ and obeys (30) if and only if*

$$P\left(\mathbb{V}(\tilde{\mathbf{a}}) = \mathbf{v}\right) \tag{37}$$
$$= \prod_{V_i \in \mathbf{V}} P\Big(V_i = v_i \;\Big|\; \{\mathrm{pa}_{\mathcal{G}}(V_i) \setminus \mathbf{A}\} = \mathbf{v}_{\mathrm{pa}_{\mathcal{G}}(V_i)\setminus\mathbf{A}},\ \{\mathrm{pa}_{\mathcal{G}}(V_i) \cap \mathbf{A}\} = \tilde{\mathbf{a}}_{\mathrm{pa}_{\mathcal{G}}(V_i)\cap\mathbf{A}}\Big),$$

whenever both sides of the equation are well-defined.[30]

Though immediate from the definitions, this Proposition is important because (37) links the counterfactual distribution $P(\mathbb{V}(\tilde{\mathbf{a}}))$ to the observed distribution $P(\mathbf{V})$. Under positivity, the right-hand side of (37) is the extended g-formula density of (Robins et al., 2004, p.2222) associated with $(\mathbf{A}, \mathbf{V})$ in the context of graph $\mathcal{G}$.[31] Under the conditions of Proposition 17, equation (37) identifies the effect of $\mathbf{A}$ on $\mathbf{V}$ whenever the RHS is a well-defined functional of the observed distribution $P(\mathbf{V})$. We again emphasize here that $\mathbf{A}$ is a subset of $\mathbf{V}$: in addition to determining the effects of $\mathbf{A}$ on the remaining variables $\mathbf{V} \setminus \mathbf{A}$ we are also determining the effect that earlier interventions $A_i = \tilde{a}_i$ have on the subsequent level of treatment $A_j(\tilde{\mathbf{a}})$ that a subject would receive, were it not for the intervention setting A_j to $\tilde{a}_j$.

Corollary 18. *Let $\mathbf{B} = \mathbf{V} \setminus \mathbf{A}$, be the set of variables in $\mathcal{G}$ that are not intervened on, and let $\mathbb{B}(\tilde{\mathbf{a}})$ be the corresponding variables in $\mathcal{G}(\tilde{\mathbf{a}})$. Then*

$$P\left(\mathbb{B}(\tilde{\mathbf{a}}) = \mathbf{b}\right) \tag{38}$$
$$= \prod_{B_i \in \mathbf{B}} P\Big(B_i = b_i \;\Big|\; \{\mathrm{pa}_{\mathcal{G}}(B_i) \setminus \mathbf{A}\} = \mathbf{b}_{\mathrm{pa}_{\mathcal{G}}(B_i)\setminus\mathbf{A}},\ \{\mathrm{pa}_{\mathcal{G}}(B_i) \cap \mathbf{A}\} = \tilde{\mathbf{a}}_{\mathrm{pa}_{\mathcal{G}}(B_i)\cap\mathbf{A}}\Big).$$

Under positivity, (38) is the (unextended) g-formula of (Robins, 1986, p.1423) for the effect on $\mathbf{B}$ of intervening on $\mathbf{A}$ in the context of graph $\mathcal{G}$. Equation (38) is also known as the 'truncated factorization formula'; see Davis (1984); Robins (1984, 1985, 1986); Spirtes et al. (1993) for some early applications. Note that (38) is the distribution of $\mathbb{B}(\tilde{\mathbf{a}})$ obtained by marginalizing over $\mathbb{A}(\tilde{\mathbf{a}})$ in (37).

Proof: Suppose $A_i(\tilde{\mathbf{a}}_{A_i})$ is the random variable corresponding in $\mathcal{G}(\tilde{\mathbf{a}})$ to some $A_i \in \mathbf{A}$ in $\mathcal{G}$. By the construction of $\mathcal{G}(\tilde{\mathbf{a}})$, $A_i(\tilde{\mathbf{a}}_{A_i})$ has no children in

[30] A product of terms that includes an undefined term, is itself well-defined if and only if at least one of the (other) terms in the product is zero.

[31] More specifically, (37) corresponds to the special case where the ordering of the variables in the g-formula is consistent with the DAG $\mathcal{G}$ and the intervention density for A_i is a point-mass on $\tilde{a}_i$; see also (60).

$\mathcal{G}(\tilde{\mathbf{a}})$, hence may be marginalized without affecting any other terms in the factorization. □

3.6.5 Completeness of the Markov property for $\mathcal{G}(\tilde{\mathbf{a}})$

We have seen in Theorem 12 that for any distribution $P(\mathbb{V}(\tilde{\mathbf{a}}))$ that factorizes according to a SWIG $\mathcal{G}(\tilde{\mathbf{a}})$, d-separation (with $\tilde{\mathbf{a}}$ included implicitly in the separating set) implies conditional independence. The following provides a converse.

Theorem 19. *If $\mathbb{B}(\tilde{\mathbf{a}})$ and $\mathbb{C}(\tilde{\mathbf{a}})$ are d-connected given $\mathbb{D}(\tilde{\mathbf{a}}) \cup \tilde{\mathbf{a}}$ in $\mathcal{G}(\tilde{\mathbf{a}})$, then there exist distributions $P(\mathbf{V})$ and $P(\mathbb{V}(\tilde{\mathbf{a}}))$ that obey modularity and factorize according to $\mathcal{G}$ and $\mathcal{G}(\tilde{\mathbf{a}})$ respectively, such that*

$$\mathbb{B}(\tilde{\mathbf{a}}) \not\perp\!\!\!\perp \mathbb{C}(\tilde{\mathbf{a}}) \mid \mathbb{D}(\tilde{\mathbf{a}}) \quad [P(\mathbb{V}(\tilde{\mathbf{a}}))].$$

Proof: It follows directly from the corresponding results for DAGs (Geiger and Pearl, 1993; Meek, 1995, see) that we may construct a distribution $P(\mathbb{V}(\tilde{\mathbf{a}}))$ in which the independence corresponding to the d-connection does not hold. Under modularity this distribution $P(\mathbb{V}(\tilde{\mathbf{a}}))$ is a partial specification of the conditional densities in the original DAG $\mathcal{G}$. We are then free to choose the remaining pieces of the conditional densities, to specify $P(\mathbf{V})$, to build a distribution factorizing according to $\mathcal{G}$. □

We note briefly that the corresponding result does not hold for 'twin networks' Pearl (2000, 2009) as these involve variables that are deterministically related.[32]

Corollary 20. *If $\mathbb{B}(\tilde{\mathbf{a}})$ and $\mathbb{C}(\tilde{\mathbf{a}})$ are d-connected given $\mathbb{D}(\tilde{\mathbf{a}}) \cup \tilde{\mathbf{a}}$ in $\mathcal{G}(\tilde{\mathbf{a}})$, then there exists a counterfactual distribution in the NPSEM-IE model (and hence also in the FFRCISTG) associated with $\mathcal{G}$, in which $\mathbb{B}(\tilde{\mathbf{a}}) \not\perp\!\!\!\perp \mathbb{C}(\tilde{\mathbf{a}}) \mid \mathbb{D}(\tilde{\mathbf{a}})$.*

Proof: By Theorem 19 there exist distributions $P^*(\mathbb{V}(\tilde{\mathbf{a}}))$ and $P^*(\mathbf{V})$ obeying factorization and modularity such that the dependence holds under $P^*(\mathbb{V}(\tilde{\mathbf{a}}))$. We construct a counterfactual distribution under which the variables $\{V(\mathbf{v}^\dagger_{\text{pa}(V)}) \mid V \in \mathbf{V}, \mathbf{v}^\dagger \in \mathfrak{V}\}$[33] are mutually independent,[34] such that:

$$P(V(\mathbf{v}^\dagger_{\text{pa}(V)}) = v) \equiv P^*(V = v \mid \text{pa}(V) = \mathbf{v}^\dagger_{\text{pa}(V)}).$$

[32] See footnote 3.

[33] Here $\mathfrak{V}$ is the state-space of $\mathbf{V}$.

[34] Observe that this independence assumption is (even) stronger than the usual NPSEM-IE assumption.

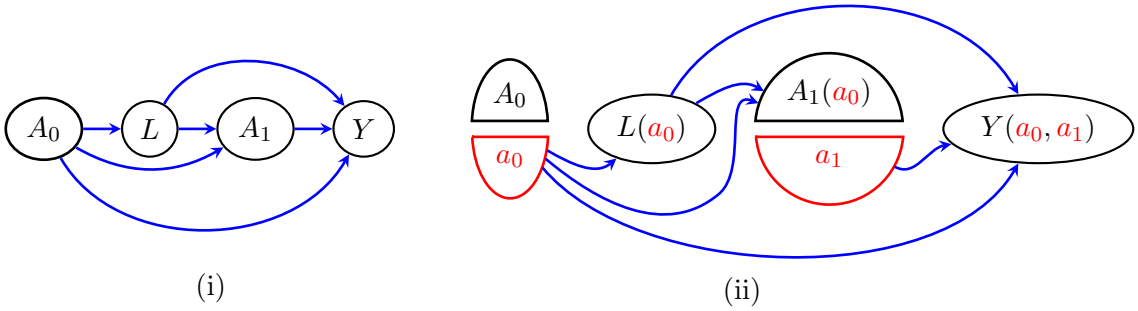


Figure 11: (i) A complete DAG $\mathcal{G}$; (ii) the template $\mathcal{G}(a_0, a_1)$.

This is sufficient to specify a distribution in the NPSEM-IE associated with $\mathcal{G}$. It then follows from Propositions 11 and 16 that $P(\mathbb{V}(\tilde{\mathbf{a}})) = P^*(\mathbb{V}(\tilde{\mathbf{a}}))$ and $P(\mathbf{V}) = P^*(\mathbf{V})$ as required. □

4 Robins' criterion for applying the g-formula and Pearl's erroneous critique

In Section 11.3.7 of his second edition Pearl offers three critiques of the g-formula based methodology developed in Robins (1986, 1987) for estimation of the causal effects of time-varying treatments. In Section 4.2 we demonstrate that all of Pearl's claims are either erroneous or based on misconceptions. We show that two of the false claims are the direct result of mathematical errors made by Pearl; the third is the result of Pearl's misreading or lack of awareness of an example in (Robins, 1986). In Section 4.1 we provide the necessary background by reviewing Robins' g-formula methodology with the aid of SWIGs. In Section 4.2 we show that Pearl's claims are erroneous.

4.1 The g-formula for a sequence of treatments and a single response

Given a causal DAG $\mathcal{G}$, interest often focuses on the distribution of a final response (Y) under a sequence of interventions that assign specific values to earlier variables, e.g. A_0 and A_1, i.e. $P(Y(a_0, a_1))$. We illustrate the simplest such situation in Figure 11(i) where the joint effect of A_0 and A_1

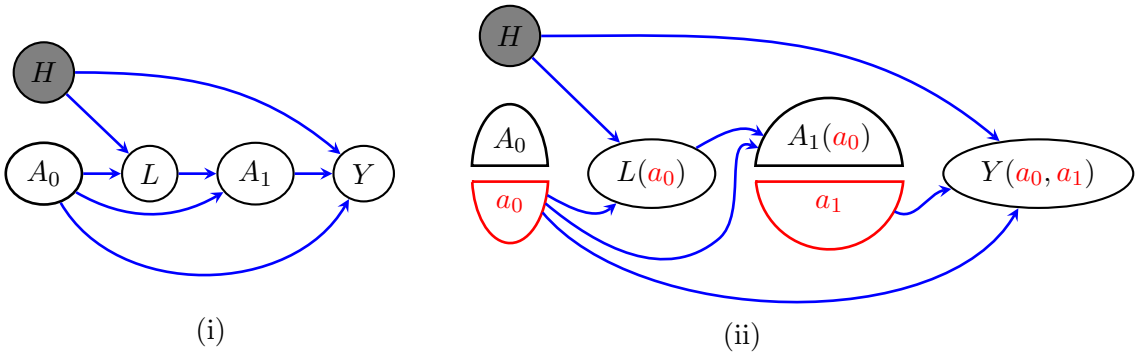


Figure 12: (i) A DAG $\mathcal{G}$ describing a sequentially randomized trial; (ii) the template $\mathcal{G}(a_0, a_1)$.

on L and Y is identified. The marginal distribution of $Y(a_0, a_1)$ is given by:

$$\begin{aligned} P(Y(a_0,a_1)=y) &= \sum_l P(L(a_0)=l, Y(a_0,a_1)=y) \\ &= \sum_l P(L=l \mid A_0=a_0) P(Y=y \mid A_0=a_0, L=l, A_1=a_1). \end{aligned} \tag{39}$$

Here the second equality follows by Corollary 18 with $\mathbf{B} = (L, Y)$. The right-hand side of (39) is precisely the marginal distribution of Y obtained by summing the g-formula density in (38) associated with $\mathbf{A} = \{A\}, \mathbf{V} = \{Y, L, A\}$ over a and l.

A key insight in Robins (1986) was the observation that the same formula (39) could be applied in situations in which there are unmeasured confounding variables. Consider, for example, the graph shown in Figure 12(i), in which H is unobserved. To see that (39) still holds, first observe the following:

$$\begin{aligned} p(l \mid h, a_0) p(h) &= p(l|h,a_0) p(h|a_0) \quad \text{since } H \perp\!\!\!\perp A_0 \\ &= p(h|l,a_0) p(l|a_0) \\ &= p(h|a_1,l,a_0) p(l|a_0) \quad \text{since } H \perp\!\!\!\perp A_1 \mid L, A_0. \end{aligned} \tag{40}$$

We now apply Corollary 18 again to obtain:

$$\begin{aligned} P(Y(a_0,a_1)=y) &= \sum_{l,h} p(l \mid h,a_0)p(y \mid a_1,l,a_0,h)p(h) \\ &= \sum_{l} p(l|a_0) \sum_{h} p(h|a_1,l,a_0)p(y \mid a_1,l,a_0,h) \\ &= \sum_{l} p(l|a_0)p(y \mid a_1,l,a_0), \end{aligned}$$

here the second equality follows from (40).

A similar derivation shows that under the graph in Figure 13(i), $P(Y(a_0,a_1)=y)$ is also identified by the same formula. Note that in this case, unlike the previous cases, both factors, $p(l|a_0)$ and $p(y \mid a_1,l,a_0)$, have no causal interpretation, since neither $P(L(a_0))$ nor $P(Y(a_0,a_1)|L(a_0))$ is identified under this graph. (Robins, 1987, §AD.3) describes a substantive setting in which the underlying FFRCISTG would be that associated with Figure 13(i).

Reading from the associated templates we see that the FFRCISTG models associated with the DAGs in Figures 11, 12, and 13 have the following independencies in common:

$$Y(a_0,a_1) \perp\!\!\!\perp A_1(a_0) \mid L(a_0), A_0 \quad \text{and} \quad Y(a_0,a_1) \perp\!\!\!\perp A_0.$$

We shall see that for all three DAGs those independencies imply identification of the distribution of $Y(a_0,a_1)$ by the marginal of Y under the g-formula density (38) arising from the DAG in Figure 11 (that does not include hidden variables).

4.1.1 The general setting

We now extend the above development to the general context. Suppose that there is a topological ordering of our graph $\mathcal{G}$ under which we have a sequence of observed variables:

$$\mathbf{O} \equiv \langle L_1, A_1, \ldots, L_K, A_K, Y \rangle,$$

where $\overline{\mathbf{A}} = (A_1, \ldots, A_K)$ are the treatments, $\overline{\mathbf{L}} = (L_1, \ldots, L_{K-1})$ are covariates and Y is the final response.[35] We note that this ordering need **not**

[35] In full generality, each of these variables could be a vector, however for convenience we will treat them as singletons in our notation and development.

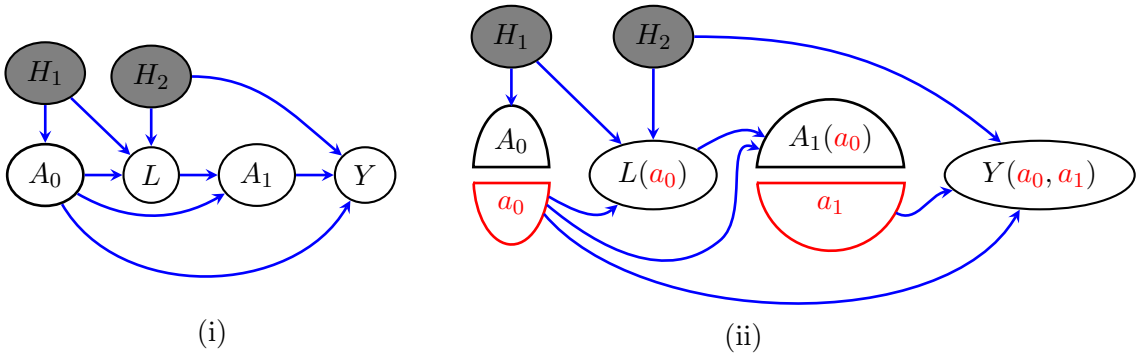


Figure 13: (i) A DAG $\mathcal{G}$ in which initial treatment is confounded, while the second treatment is sequentially randomized; (ii) the SWIT $\mathcal{G}(a_0, a_1)$. L is known to have no direct effect on Y, except indirectly via the effect on A_1.

be temporal.[36] In general there may, in addition, be hidden variables $\mathbf{H}$ in the graph.

Our primary focus here will be on conditions under which we can identify the effect of a given assignment of treatments conditional on treatment and covariate history. To be more precise, given $P(\mathbf{O})$ we wish to identify:

$$P(Y(\mathbf{a}^\dagger_Y) = y \,|\, \overline{\mathbb{L}}_j(\mathbf{a}^\dagger) = \bar{\mathbf{l}}_j, \overline{\mathbb{A}}_{j-1}(\mathbf{a}^\dagger) = \bar{\mathbf{a}}^\dagger_{j-1}) \quad \text{for } j = 0, \ldots, K, \tag{41}$$

here $\bar{\mathbf{a}}_k \equiv (a_1, \ldots, a_k)$, and $\bar{\mathbf{l}}_k \equiv (l_1, \ldots, l_k)$, $\bar{\mathbf{a}}_{-1}$, $\bar{\mathbf{a}}_0$ and $\bar{\mathbf{l}}_0$ are empty vectors and similarly $\overline{\mathbb{L}}_0(\mathbf{a}^\dagger)$, $\overline{\mathbb{A}}_0(\mathbf{a}^\dagger)$ are empty sets of variables.

One motivation for considering identification not merely of the marginal potential outcome distribution $P(Y(\mathbf{a}^\dagger_Y))$ but conditional on $\overline{\mathbb{L}}_j(\mathbf{a}^\dagger)$ and $\overline{\mathbb{A}}_{j-1}(\mathbf{a}^\dagger)$ is that in the case where the induced subgraph of $\mathcal{G}$ on the observed variables is complete,[37] identification of these conditional intervention distributions is necessary and sufficient in order to test the null hypothesis that the distribution of the outcome $Y(g)$[38] is the same for all regimes g, includ-

[36] Since we do not believe the future may cause the past, we will assume that there is a temporal ordering of all variables, which gives a topological ordering of the graph. However, when we consider incomplete graphs there are situations in which we will use non-temporal (but still topological) orderings; see the Healthy Worker Survival Effect example in §4.2.4 and §7 below.

[37] Observe that if the induced subgraph $\mathcal{G}$ on the observed variables is complete then any topological order on $\mathcal{G}$ induces the same ordering on these variables, which will thus coincide with the temporal ordering.

[38] $Y(g)$ is defined formally in Section 5 below.

ing dynamic treatment regimes[39] where treatment levels at subsequent times are determined by the subject's response to earlier treatment; see Corollary 23 and §5.

Definition 21. *The* g-formula for the marginal of Y conditional on the values of an initial sequence of observed variables $\bar{\mathbf{l}}_m$, under treatment assignment $\mathbf{a}^\dagger = \bar{\mathbf{a}}^\dagger_K$, *is defined to be:*

$$b_{\mathbf{a}^\dagger}(y \mid \bar{\mathbf{l}}_m) \equiv \sum_{l_{m+1},\ldots,l_K} p(y|\bar{\mathbf{l}}_K, \bar{\mathbf{a}}^\dagger_K) \prod_{j=m+1}^{K} p(l_j|\bar{\mathbf{l}}_{j-1}, \bar{\mathbf{a}}^\dagger_{j-1}), \tag{42}$$

for $m = 0, \ldots, K$. *We will define* $b_{\mathbf{a}^\dagger}(y) \equiv b_{\mathbf{a}^\dagger}(y \mid \bar{\mathbf{l}}_0)$.

Throughout the remainder of this section and the next (§4-4.2) we will assume that for $i = 1, \ldots, K$:

$$P[\overline{\mathbf{L}}_{i-1} = \bar{\mathbf{l}}_{i-1}, \overline{\mathbf{A}}_{i-1} = \bar{\mathbf{a}}^\dagger_{i-1}] > 0 \;\Rightarrow\; P\left[A_i = a_i^\dagger \mid \overline{\mathbf{L}}_{i-1} = \bar{l}_{i-1}, \overline{\mathbf{A}}_{i-1} = \bar{\mathbf{a}}^\dagger_{i-1}\right] > 0, \tag{43}$$

so that the RHS of (42) is always well-defined.

Note that although the formula (42) uses the specified ordering of the variables, it makes no explicit reference to a graph. Consider a DAG $\mathcal{G}$ with variables $\mathbf{V} = \langle L_1, A_1, \ldots, L_K, A_K, Y \rangle$, under which this ordering is topological. Under positivity the summand in (42) is equivalent to the g-formula density on the right hand side of (38).[40]

By Corollary 18 the formula (42) would arise if we had started with the FFRCISTG model associated with a complete DAG (consistent with this ordering) from which we constructed the distribution $P(\mathbb{V}(\mathbf{a}^\dagger))$ resulting from intervening to set $\mathbf{A} = \{A_1, \ldots, A_K\}$ to $\mathbf{a}^\dagger$. Computing the conditional distribution $P(Y(\mathbf{a}^\dagger_Y) \,|\, \overline{\mathbb{L}}_m(\mathbf{a}^\dagger), \overline{\mathbb{A}}_{m-1}(\mathbf{a}^\dagger))$ from such a complete graph leads to (42).

However, in Theorem 22 we will see that there are many more DAGs for which this is true: if certain counterfactual independence conditions hold then the same formula may be used to identify the same conditional counterfactual distribution in graphs which may include hidden variables. The analysis of Figures 12 and 13 above constitute examples.

[39] Note that treatment regimes that we, following (Robins, 1986), term 'dynamic' are called 'conditional plans' in (Pearl, 2000, 2009, p.121); in that work the term 'dynamic plan' refers to any intervention on multiple variables.

[40] Consequently, $b_{\mathbf{a}^\dagger}(y \mid \bar{\mathbf{l}}_m)$ is the conditional distribution of Y given $\overline{\mathbf{L}}_m$ under the g-formula density (42).

Now consider a graph $\mathcal{G}$ with vertex set $\mathbf{V} = \mathbf{O} \cup \mathbf{H}$. We require that the ordering of $\mathbf{O} = \langle L_1, A_1, \ldots, L_K, A_K, Y \rangle$ may be extended to a topological ordering for $\mathbf{V}$ relative to $\mathcal{G}$. We will let $Y(\mathbf{a}^\dagger)$, $\mathbb{A}(\mathbf{a}^\dagger)$, $\mathbb{L}(\mathbf{a}^\dagger)$, and $\mathbb{H}(\mathbf{a}^\dagger)$ be the corresponding (sets of) variables in $\mathcal{G}(\mathbf{a}^\dagger)$. Likewise, to simplify notation we will also use the following shorthands: $L_k(\mathbf{a}^\dagger) = L_k(\mathbf{a}^\dagger_{L_k})$, $A_k(\mathbf{a}^\dagger) = A_k(\mathbf{a}^\dagger_{A_k})$ and $Y(\mathbf{a}^\dagger) = Y(\mathbf{a}^\dagger_Y)$.

Theorem 22. *Let $\mathbf{a}^\dagger$ be an instantiation for $\mathbf{A}$. If $P(\mathbf{V})$ and $P(\mathbb{V}(\mathbf{a}^\dagger))$ factor according to $\mathcal{G}$ and $\mathcal{G}(\mathbf{a}^\dagger)$ respectively, and obey modularity (30) then for all $j = 0, \ldots, K$, and all $\bar{\mathbf{l}}_j$,*

$$b_{\mathbf{a}^\dagger}(y \mid \bar{\mathbf{l}}_j) = P(Y(\mathbf{a}^\dagger) = y \mid \overline{\mathbb{L}}_j(\mathbf{a}^\dagger) = \bar{\mathbf{l}}_j, \overline{\mathbb{A}}_{j-1}(\mathbf{a}^\dagger) = \bar{\mathbf{a}}^\dagger_{j-1}) \tag{44}$$

if and only if for $k = 1, \ldots, K$:

$$Y(\mathbf{a}^\dagger) \perp\!\!\!\perp I(A_k(\mathbf{a}^\dagger) = a_k^\dagger) \mid \overline{\mathbb{L}}_k(\mathbf{a}^\dagger), \overline{\mathbb{A}}_{k-1}(\mathbf{a}^\dagger) = \bar{\mathbf{a}}^\dagger_{k-1}; \tag{45}$$

here $I(\cdot)$ is the indicator function.

It thus follows that under (45), the counterfactual distribution on the RHS of (44) is identified whenever $b_{\mathbf{a}^\dagger}(y \mid \bar{\mathbf{l}}_j)$ is a well-defined functional of the observed marginal distribution $P(\mathbf{O})$ as would be the case if (43) held. Unless explicitly stated otherwise, in this section we will always assume that this is the case.

As the graphs $\mathcal{G}$ and $\mathcal{G}(\mathbf{a}^\dagger)$ may include hidden variables, the Theorem gives conditions under which causal effects are identified even in the presence of hidden variables. However, whether (45) holds is not empirically testable from $P(\mathbf{O})$ without further assumptions.[41]

In fact, in prior work Robins[42] showed that under consistency (and thus under the FFRCISTG model) the following condition was necessary and sufficient for (44): for $k = 1, \ldots, K$,

$$Y(\mathbf{a}^\dagger) \perp\!\!\!\perp I(A_k = a_k^\dagger) \mid L_1, \ldots, L_{k-1},\ A_1 = a_1^\dagger, A_2 = a_2^\dagger, \ldots, A_{k-1} = a_{k-1}^\dagger.[43] \tag{46}$$

[41] The assumption (45) can be ensured to hold in a sequentially randomized experiment, where treatment A_k is assigned randomly given past treatment and covariate history.

[42] Corollary to Theorem AD.1 on p.930 of (Robins, 1987), with necessity following from reversibility of each step in the proof. Theorem 56 in Appendix B reproduces this result in the notation of the current paper.

[43] (Pearl, 2000, p.103, fn 15) states: "we note that the class of semi-Markovian models satisfying assumption [(46)] corresponds to complete DAGs in which all arrowheads point to X_k originate from observed variables". The graphs in Figure 12 and 13 show that though sufficient, this is very far from a full characterization of the graphs satisfying (46).

This is because under the assumption of consistency, conditions (45) and (46) are equivalent. However, though necessary and sufficient, neither of these conditions may be checked (graphically) from $\mathcal{G}(\mathbf{a}^\dagger)$ since applying d-separation to $\mathcal{G}(\mathbf{a}^\dagger)$ does not lead to context-specific independence statements such as (46), nor to independence from indicator functions as in (45) and (46).

However, we may verify the following stronger counterfactual independence condition: for $k = 1, \ldots, K$,

$$Y(\mathbf{a}^\dagger_Y) \perp\!\!\!\perp A_k(\mathbf{a}^\dagger_{A_k}) \mid \overline{\mathbb{L}}_k(\mathbf{a}^\dagger), \overline{\mathbb{A}}_{k-1}(\mathbf{a}^\dagger). \tag{47}$$

This condition may be checked directly on an instantiated template $\mathcal{G}(\mathbf{a}^\dagger)$. This condition is sufficient, but not necessary.[44] We note that although the condition (47) is not the weakest such condition, it is nonetheless weaker than the (also) sufficient under consistency, but not necessary condition: for $k = 1, \ldots, K$,

$$Y(\mathbf{a}^\dagger) \perp\!\!\!\perp A_k \mid L_1, L_2, \ldots, L_{k-1},\ A_1, A_2, \ldots, A_{k-1}, \tag{48}$$

obtained by dropping the specific values from the conditioning event in (46).

We note that (Pearl and Robins, 1995) also present a different, but logically equivalent, graphical method for checking the condition (47). However, as noted earlier, their algorithm requires the user to construct a separate graph for each $k = 1, \ldots, K$.

As remarked above, identification of all of the conditional distributions (41) given above is necessary and sufficient in order to test the null hypothesis that for all regimes g the distribution of the outcome $Y(g)$ is the same. Thus we have the following result:

Corollary 23 (Robins (1986), Theorem 6.1)**.** *Consider a graph $\mathcal{G}$ in which the induced subgraph on the observed variables is complete. Let $Y(g)$ denote the counterfactual outcome resulting from a dynamic treatment regime in which A_i is set to $g_i(\bar{\mathbf{l}}_i, \bar{\mathbf{a}}_{i-1})$, for $i = 1, \ldots K$. Condition (45) holding for $k = 1, \ldots, K$ is a necessary and sufficient condition for the existence of a consistent test that $P(Y(g)=y) = P(Y=y)$ for all dynamic regimes g.*

Proof: See Robins (1986) Appendix E. □

[44] Although in many situations we will derive (45) by applying d-separation to $\mathcal{G}(\mathbf{a}^\dagger)$, Theorem 22 does not require this. For example, it is conceivable that $\mathcal{G}$ is complete, yet (45) holds 'unfaithfully'.

4.1.2 A complete algorithm for intervention distributions

There are intervention distributions that are identified, but are not identified via any g-formula; see the graph in Figure 14(b), discussed in §4.2 below. Tian and Pearl (2002) presented an algorithm that was shown to be complete for the identification of intervention distributions; see Huang and Valtorta (2006); Shpitser and Pearl (2006b); the algorithm was extended to cover conditional intervention distributions in Shpitser and Pearl (2006a). However the g-formula continues to play a role. Specifically, it follows from (Tian and Pearl, 2002, Lemma 1) that every identifiable intervention distribution may be obtained as the sum over a product of expressions obtained by interleaving applications of the g-formula and marginalization steps; see also (Shpitser et al., 2011).

4.2 Pearl's Erroneous Critique of Robins' criterion

With Section 4.1 as background we now turn to Pearl's critique. Pearl (2009), Section 11.3.7, claims to improve and generalize on Robins' results concerning identification with the g-formula in the previous section. Specifically, Pearl claims that (46) is "over restrictive".[45] This is surprising given that, as stated above, the condition is necessary and sufficient.

4.2.1 Pearl's proposed criterion

Pearl proposes instead the following criterion, which we state here in full:

> **(3.62*) General Condition for g-Estimation [sic]**[46]
> $P(y \mid g = x)$ is identifiable and is given by [(42) above[47]] if every action avoiding back-door path from X_k to Y is blocked by some subset L_k of non-descendants of X_k. (By "action avoiding" we mean a path containing no arrows entering an X variable later than X_k.)

(Note that here the variables X_i are the variables intervened upon, i.e. the A_i's in this paper.)

[45] Pearl (2000), footnote 15 on p.103, states that (46) is the condition that is the "weakest assumption needed for identifying the causal effect". He appears to have changed his mind subsequently as the footnote has been removed in (Pearl, 2009).

[46] Though Pearl (2000, 2009) refers to "g-estimation", the criterion he is extending is usually referred to as the 'g-formula' or the 'g-computation formula'. In contrast, 'g-estimation' (Robins et al., 1992) is a method for estimating the parameters of semi-parametric structural nested models via estimating functions. Thus, g-estimation and g-computation are entirely distinct. Note also a (new) typographical error has been introduced in (Pearl, 2009) in the g-formula (3.63): the sum should be over $\bar{l}_K$ not $\bar{l}_k$.

[47] Specifically, with $m = 0$ and $\mathbf{a}^\dagger = x$.

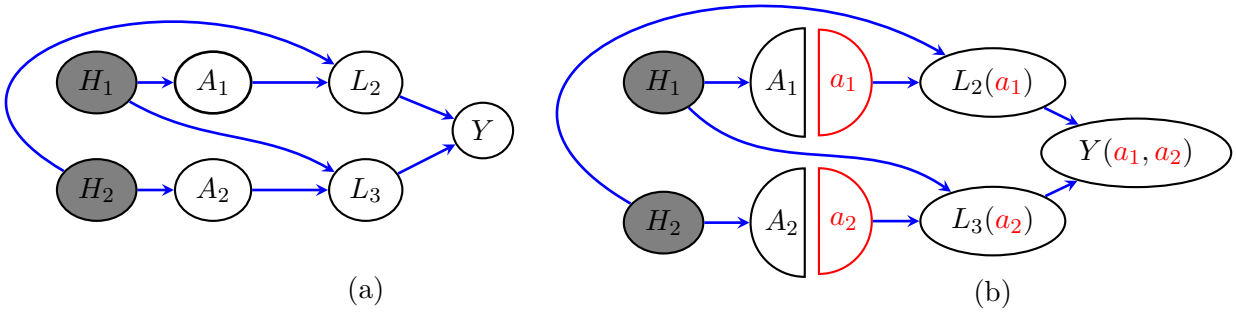


Figure 14: Failure of the extended criterion (3.62*) proposed in Pearl (2009). (a) A DAG $\mathcal{G}$, with hidden variables H_1, H_2, in which A_1 and A_2 obey the extended criterion (3.62*), yet the effect of A_1 and A_2 on Y is not given by any g-formula; (b) the template $\mathcal{G}(a_1, a_1)$.

This extended criterion is not valid. This may be seen by considering the graph shown in Figure 14, in which H_1 and H_2 are unobserved. The graph satisfies condition (3.62*):

A_1: The action-avoiding backdoor path $A_1 \leftarrow H_1 \rightarrow L_3 \rightarrow Y$ is blocked by L_3 which is not a descendant of A_1.

A_2: Similarly, the action avoiding backdoor path $A_2 \leftarrow H_2 \rightarrow L_2 \rightarrow Y$ is blocked by L_2 which is not a descendant of A_2.

Thus according to the claim in (3.62*) the following equality should hold:

$$P(Y(a_1, a_2) = y) \stackrel{??}{=} \sum_{l_2, l_3} p(y \mid l_2, l_3, a_1, a_2) p(l_3 \mid l_2, a_1, a_2) p(l_2 \mid a_1). \tag{49}$$

Unfortunately, the equality (49) is false: the expression above does *not* correspond to $p(Y(a_1, a_2) = y)$ as we now show. First note that the g-formula on the RHS of (49) is equivalent to:

$$\sum_{l_2, l_3} p(y \mid l_2, l_3) p(l_2 \mid a_1) p(l_3 \mid a_1, a_2), \tag{50}$$

where we have used the independence facts: $Y \perp\!\!\!\perp A_1, A_2 \mid L_2, L_3$ and $L_3 \perp\!\!\!\perp L_1 \mid A_1, A_2$. From the template shown in Figure 14(b) we have the following:

$$
\begin{aligned}
&P(Y(a_1,a_2)=y) \\
&\quad= \sum_{h_1,h_2,l_2,l_3} p(h_1)p(h_2)p(L_2(a_1)=l_2 \mid h_2)p(L_3(a_2)=l_3 \mid h_1) \\
&\qquad\qquad \times p(Y(a_1,a_2)=y \mid L_2(a_1)=l_2, L_3(a_2)=l_3) \\
&\quad= \sum_{l_2,l_3} \left(\sum_{h_1} p(h_1)p(l_3 \mid a_2, h_1) \right) \left(\sum_{h_2} p(h_2)p(l_2 \mid a_1, h_2) \right) p(y \mid l_2, l_3) \\
&\quad= \sum_{l_2,l_3} \left(\sum_{h_1} p(h_1 \mid a_2)p(l_3 \mid a_2, h_1) \right) \left(\sum_{h_2} p(h_2 \mid a_1)p(l_2 \mid a_1, h_2) \right) p(y \mid l_2, l_3) \\
&\quad= \sum_{l_2,l_3} p(y \mid l_2, l_3)p(l_2 \mid a_1)p(l_3 \mid a_2). \qquad (51)
\end{aligned}
$$

Here: the first equality follows from truncated factorization (38); the second from modularity; the third since $H_1 \perp\!\!\!\perp A_2$ and $H_2 \perp\!\!\!\perp A_1$. The result (51) could also have been obtained by applying the algorithm given by Tian and Pearl (2002); see also Shpitser et al. (2011).

Though (51) and (50) have two terms in common, they differ on the third. Thus, they will not be the same, so long as $p(l_3 \mid a_2) \neq p(l_3 \mid a_1, a_2)$, or equivalently $L_3 \not\perp\!\!\!\perp A_1 \mid A_2$. Since L_3 and A_1 are d-connected given A_2 on the graph, this independence will not hold in general.

Notice that the identifying expression (51) is *not* an instance of the g-formula because the conditioning set for L_3 is not a superset of that for L_2. Furthermore there is no way to use conditional independence to make this the case. There are six orderings of the observed variables that are compatible with the graph:

$$
\begin{aligned}
&(A_1, A_2, L_2, L_3, Y), \quad (A_1, A_2, L_3, L_2, Y), \quad (A_1, L_2, A_2, L_3, Y), \\
&(A_2, A_1, L_2, L_3, Y), \quad (A_2, A_1, L_3, L_2, Y), \quad (A_2, L_3, A_1, L_2, Y).
\end{aligned}
$$

None lead to a valid g-formula for $P(Y(a_1, a_2))$. Given Theorem 22 above it could not be otherwise: In every topological ordering there is at least one A_i vertex that is not preceded by either L_2 or L_3, hence as may be seen from the template in Figure 14(b), this variable will be d-connected to $Y(a_1, a_2)$ (given $\emptyset$). Since by Theorem 19, d-separation is complete, it follows that under any ordering, there exist distributions $P(\mathbf{V})$ and $P(\mathbb{V}(a_1, a_2))$ that factor and obey modularity with respect to $\mathcal{G}$ and $\mathcal{G}(a_1, a_2)$, yet the condition (45) fails to hold.

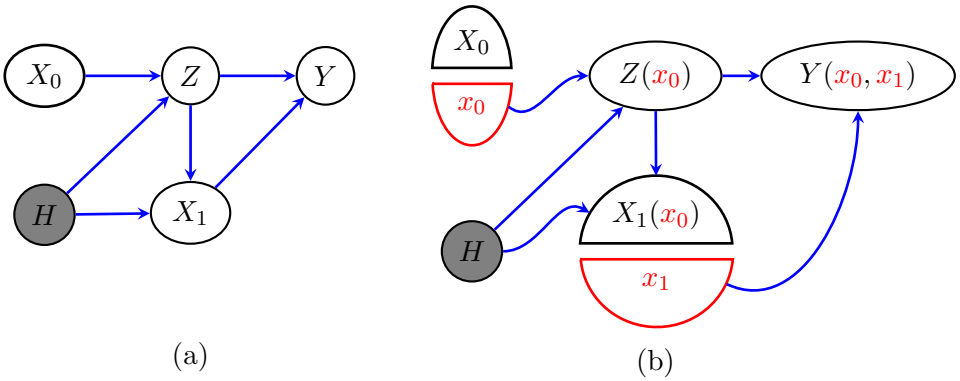


Figure 15: (a) A DAG $\mathcal{G}$ in which Robins' condition (46) is falsely claimed not to hold; see Ex. 11.3.3, Fig. 11.12 in Pearl (2009, p.353); H is unobserved; (b) the template $\mathcal{G}(x_0, x_1)$.

Pearl claims on p.352 that he "derived" the sufficiency of condition (3.62*) for the g-formula by applying Theorem 4.4.1 in (Pearl, 2009). This cannot be the case since the latter, which appears in (Pearl and Robins, 1995), is true, while the former is false.

4.2.2 Examples claiming to show extensions of the g-formula

After presenting the (flawed) criterion (3.62*) Pearl also presents examples, which he incorrectly claims show instances in which a g-formula may be applied even though Robins' criterion fails to hold. One example is based on a mathematical error, the others on misconceptions concerning prior work.

4.2.3 Example: Failure to correctly assess context specific independence in twin networks

Example 11.3.3 in Pearl (2009) is the graph shown[48] in Figure 15(a). We reproduce Pearl's discussion regarding this example:

> Figure [15(a)] demonstrates a case where [(46)] is not satisfied [...] but the graphical condition (3.62*) is. It is easy to see that (3.62*) is satisfied; all back-door action-avoiding paths from X_1 to Y are blocked by $\{X_0, Z\}$. At the same time, it is possible to show (using the Twin

[48] Where we have a hidden variable $Z \leftarrow H \rightarrow X_2$, Pearl uses a bi-directed edge $Z \leftrightarrow X_2$.

> Network Method [...]) that $Y(x_0, x_1)$ is not independent of X_1, given Z and X_0. [...] Therefore, [(46)] is not satisfied for $Y(x_0, x_1)$ and X_1. [...]
>
> This example is another demonstration of the weakness of the potential-outcome language [...]

In this example, as Pearl recognizes, the g-formula is applicable with the ordering $\langle X_0, Z, X_1, Y\rangle$. Further, condition (46) holds, yet Pearl falsely claims it does not. Pearl *is* correct in his assertion that

$$Y(x_0, x_1) \not\perp\!\!\!\perp X_1 \mid Z, X_0. \tag{52}$$

However, careful readers of the material in the last section[49] will observe that the lack of independence (52) does not imply that (46) is false. The latter is a context specific independence. Consequently Pearl's conclusion here is erroneous. To apply the g-formula it is sufficient by (46) that:

$$Y(x_0, x_1) \perp\!\!\!\perp X_0, \quad \text{and} \quad Y(x_0, x_1) \perp\!\!\!\perp X_1 \mid Z, X_0 = x_0. \tag{53}$$

Under the FFRCISTG model associated with Figure 15(a) we can apply the consistency assumption, Def. 1 (ii), to (53) to see that the latter condition is equivalent to:

$$Y(x_0, x_1) \perp\!\!\!\perp X_1(x_0) \mid Z(x_0), X_0 = x_0. \tag{54}$$

We see by examining the template $\mathcal{G}(x_0, x_1)$ shown in Figure 15(b), that:

$$Y(x_0, x_1) \perp\!\!\!\perp X_1(x_0) \mid Z(x_0), X_0, \tag{55}$$

since there is no path d-connecting $Y(x_0, x_1)$ and $X_1(x_0)$ given $\{X_0, Z(x_0)\}$ in $\mathcal{G}(x_0, x_1)$. It then follows immediately (by conditioning on $X_0 = x_0$ in (55)) that (54) holds. We conclude that the independences (53) hold under the FFRCISTG model and thus under the NPSEM-IE sub-model assumed by Pearl. Our approach based on SWIGS (combined with consistency) prevents the user from committing the error made by Pearl.

The independence relations (53) may also be obtained by applying (an extension of) the twin-network method (Pearl, 2000, 2009, §7.1.4), though this requires some care (see Appendix D.4). It is quite clear from Pearl's discussion of this example (Pearl, 2009, p.353; see also p.395) that either

[49] Also those who look up condition (3.62) in (Pearl, 2000, 2009).

he failed to apply this algorithm correctly, or he failed to observe that $Y(x_0, x_1) \not\perp\!\!\!\perp X_1 \mid Z, X_0$ does not imply $Y(x_0, x_1) \not\perp\!\!\!\perp X_1 \mid Z, X_0 = x_0$.[50]

We see some irony here: in (Pearl, 2009, Ex. 11.3.3, p. 353) this example was supposed to illustrate an unnoticed shortcoming of the condition given by Robins for the application of the g-formula. However, in fact Pearl's faulty analysis demonstrates the difficulty he encountered in applying his twin-network method to accurately evaluate the above counterfactual independence statements. Our approach based on SWIGs immunizes the user against drawing these erroneous conclusions.

4.2.4 Examples in which the g-formula may be applied but under a non-temporal order

Example 11.3.1 and Example 11.3.2 in Pearl (2009) make the same point: identification by the g-formula may be possible, but only under a non-temporal order. Specifically Example 11.3.1 in Pearl (2009), shown in Figure 16, is a familiar example of 'm-bias' (Greenland, 2003) where $Y(a) \perp\!\!\!\perp A$, yet $Y(a) \not\perp\!\!\!\perp A \mid Z$. Pearl argues correctly that in this instance, if the ordering of the observed variables was (Z, A, Y), and if we were to apply the g-formula using this ordering then the resulting g-formula would not correspond to $P(Y(a))$. However, under the ordering (A, Z, Y) which is compatible with the DAG (but not the original time-order) we would obtain the correct g-formula (here, simply $P(Y(a)=y) = P(y \mid a)$).

Example 11.3.2 in Pearl (2009), shown in Figure 17, makes the same point. In this instance, the temporal ordering given is (A, S, Y). It is easy to see that because $Y(a) \perp\!\!\!\perp A \mid S$, yet $Y(a) \not\perp\!\!\!\perp A$, we may apply the g-formula but only under a non-temporal ordering.

Both of these examples do not contradict Theorem 22 since that result, though it uses an ordering, does not require that it be temporal. The examples do show that in certain cases it may be beneficial to consider non-temporal orderings. Though interesting, this is not a new point, as it is discussed at length in (Robins, 1986, Section 11), specifically with regard to the discussion of the minimum latent period in the Healthy Worker Survival Effect. We briefly review this analysis from the perspective of SWIGs.

[50] Specifically, he (incorrectly) argues that because $Y(x_0, x_1)$ and X_1 are d-connected (given Z and X_0) in the 'twin network' by the path $X_1 \leftrightarrow Z \leftrightarrow Z(x_0, x_1) \rightarrow Y(x_0, x_1)$, "therefore" (46) does not hold. However (under consistency), when $X_0 = x_0$, $Z = Z(x_0) = Z(x_0, x_1)$, so that in spite of the presence of this path, $Y(x_0, x_1) \perp\!\!\!\perp X_1 \mid Z, X_0 = x_0$; see Appendix D.

The Healthy Worker Survival Effect

Consider a longitudinal cohort study of workers exposed to an industrial lung carcinogen with time to clinically diagnosed lung cancer as outcome. Robins showed that when the minimal latent period (MLP) exceeds x years and the duration of the *healthy worker survivor effect* (HWSE) is less than x years, one may be able to obtain control of confounding due to the HWSE by coding the indicator function for survival at $t+x$ as occurring *prior* to the exposure received at time t: a rather dramatic example of the phenomenon in Example 11.3.2.

Here are the (somewhat simplified) details formulated in graphical terms. We say there is a MLP of at least x years if an exposure received today cannot result in a clinically diagnosed lung cancer within x years. We say that the HWSE lasts no more than z years if, whenever a subject has undiagnosed lung cancer that is sufficiently advanced to cause the subject to quit work (or change exposure in some other way), the cancer will be clinically diagnosed within the next z years. The situation is represented by the DAG in Figure 18(a) where A_t, H_t, and D_t are indicators of carcinogen exposure and of undiagnosed and diagnosed lung cancer at year t. H_t is, of course, unobserved. A_t is a parent of H_{t+1} and D_{t+2}, but not of D_{t+1}, indicating an MLP of between 1 and 2 years. H_t has A_t and D_{t+1} but not D_{t+2} as children, indicating the HWSE has duration no greater than 1 year. The DAG in Figure 18 (b) reorders the data; the survival indicator at $t+1$, D_{t+1}, is now earlier than A_t in the ordering.

The DAG in Figure 18(c) is our template. By d-separation $A_0 \perp\!\!\!\perp D_2\,(a_0) \mid D_1$ where $D_t = 0$ if a subject has survived to t without a lung cancer diagnosis. Hence by Eq. (45) we can apply the g-formula with the ordering $\langle D_1, A_0, D_2 \rangle$ to obtain

$$p[D_2(a_0) = 0] = p[D_2 = 0 | A_0 = a_0, D_1 = 0] p[D_1 = 0]$$

where the summation over D_1 is absent since $D_2 = 0$ implies $D_1 = 0$.

5 SWIGs for dynamic regimes

In this section we show that we may analyze identification conditions for dynamic regimes via a simple extension of our graphical formalism. A dynamic regime g is a policy that assigns treatment on the basis of past history; following Robins (1986) we shall use the letter g to denote a dynamic regime. Robins et al. (2004) proposed allowing this history to include the natural

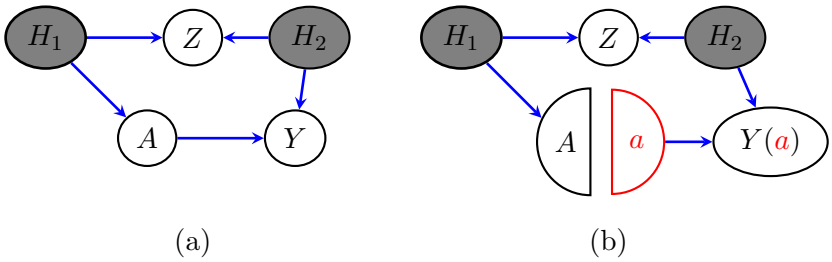


Figure 16: (a) A DAG with two unobserved variables (H_1, H_2) depicting the first example presented by Pearl claiming to extend Robins' criterion for applying the g-formula; see Ex. 11.3.1, Fig. 11.10 in Pearl (2009, p.352); (b) the corresponding template.

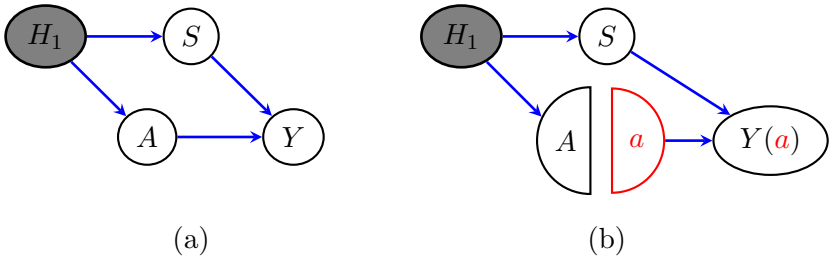


Figure 17: (a) A DAG with a single unobserved variable (H_1) depicting the second example presented by Pearl claiming to extend Robins' criterion for applying the g-formula; see Ex. 11.3.2, Fig. 11.11 in Pearl (2009, p.353); (b) the corresponding template.

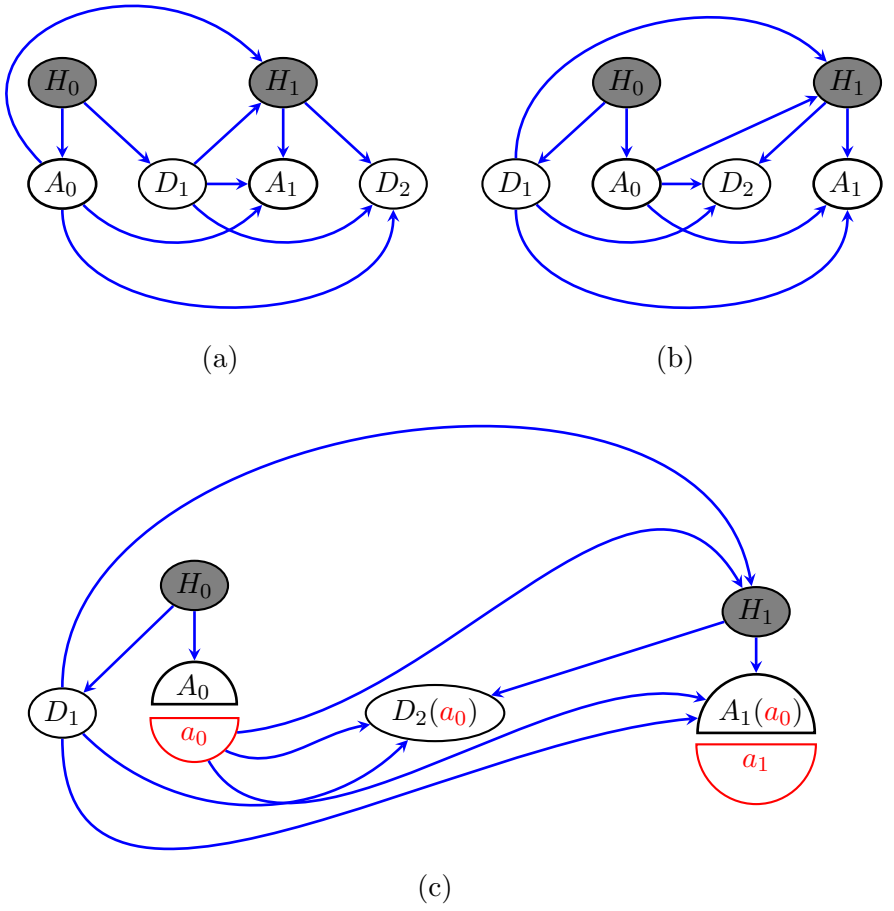


Figure 18: (a) A DAG $\mathcal{G}$ with two unobserved variables (H_1, H_2) depicting the minimum latent period scenario; (b) the same DAG with the variables re-ordered; (c) the corresponding template $\mathcal{G}(a_0, a_1)$.

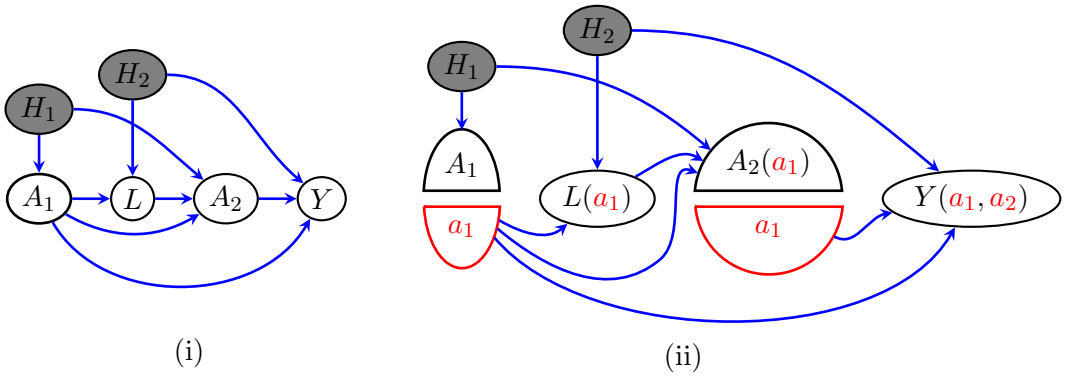


Figure 19: (i) A DAG $\mathcal{G}$ describing an observational study in which the effect of A_1 on A_2 is confounded; (ii) the template $\mathcal{G}(a_1, a_2)$.

values that a variable would take in the absence of its value being specified by the regime. These regimes have also been considered by Danaei et al. (2012); Haneuse and Rotnitzky (2013); Lajous et al. (2012); Muñoz and van der Laan (2011); Taubman et al. (2008, 2009); Young et al. (2012). Due to its novelty we will focus on this type of regime in the introduction to this section. Identification of the distribution of a response Y under a dynamic regime when natural values are *not* used to assign treatment has been considered by Dawid and Didelez (2008, 2010); Pearl (2000); Robins (1986, 1987, 1989a, 1997); Tian (2008). In particular, Tian proposed the first complete algorithm. Identification of the distribution of a response Y under a dynamic regime when natural values are used to assign treatment was discussed briefly in Robins et al. (2004); Haneuse and Rotnitzky (2013); Muñoz and van der Laan (2011, 2012); Shpitser and Pearl (2012); Young et al. (2012).

5.1 Dynamic regimes depending on the natural level of treatment

To consider a specific example, suppose that we have data on minutes exercised (A_1) and subsequent blood pressure (Y) collected from a large HMO. Let $P(A_1, Y)$ be the observed (population) distribution. Now consider the following regime:

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

This context may be conceptualized in terms of there being *two* variables associated with treatment: the 'natural' level of treatment A_1 which the patient performs in the absence of an intervention, and the level of treatment prescribed by the regime g which we denote $A_1^+(g)$. In an example such as this the regime-prescribed treatment level $A_1^+(g)$ is a function of the level (A_1) that the patient would perform (in the absence of an intervention). The exercise regime described would correspond to:

$$A_1^+(g) = g(A_1) \quad \text{where} \quad g(x) \equiv \max\{x, 20\}.$$

We denote the treatment assigned by g via an upper case letter $A_1^+(g)$ since it is random.

Suppose that we are interested in the effect of the regime on the patient's blood pressure (Y). We define the counterfactual response under regime g to be:

$$Y(g) \equiv Y(a_1)|_{a_1 = A_1^+(g) = g(A_1)}. \tag{56}$$

Following Young et al. (2012) we consider the question of whether this regime could be implemented in an actual randomized experiment. Now in place of observational HMO data, consider an experimenter who wishes to implement this regime. Note that in order for this regime to be implemented, data on the natural value of A_1 must be collected.[51] Thus, for example, the experimenter might watch each subject exercise, say in a gymnasium, and if the subject goes to get dressed after A_1 minutes with $A_1 < 20$, the experimenter obliges them to continuing exercising until they have exercised for 20 minutes. We will see below that under the FFRCISTG model associated with the DAG $\mathcal{G}$ in Figure 20(a), the distribution $P(Y(g))$ of $Y(g)$ under regime g, is identified by the distribution $P(A_1, Y)$ of HMO data. We will represent this dynamic regime via the template shown in Figure 20(c).

5.2 Treatment at multiple time-points

If there are two temporally ordered treatments then we can consider more complex regimes. Consider the regime where:

$$A_1^+(g) = g_1(A_1) \quad \text{with} \quad g_1(x) \equiv \max\{x, 20\};$$
$$A_2^+(g) = g_2(A_2(g), A_1^+(g), A_1), \quad \text{where} \quad g_2(a_2, a_1^+, a_1) \equiv a_2 + (a_1^+ - a_1).$$

[51] There exist other regimes that depend on the natural value which could not be implemented. Young et al. (2012) discuss some of the philosophical and conceptual difficulties that then arise. We also discuss both the mathematical and philosophical implications of this in §6.1.4 but we do so in the context of an FFRCISTG model and obtain some counter-intuitive results.

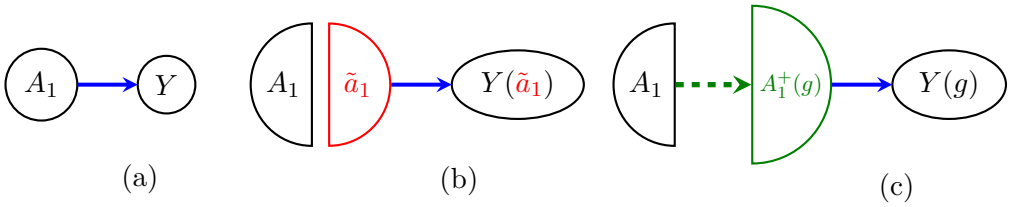


Figure 20: (a) a causal DAG $\mathcal{G}$; (b) the (non-dynamic) SWIG $\mathcal{G}(\tilde{a}_1)$; (c) the dynamic SWIG $\mathcal{G}(g)$ representing a dynamic regime g under which $A_1^+(g)$, the level of treatment prescribed by the regime, depends on the level of treatment A_1 that the patient *would* choose in the absence of the regime. The edge from A_1 to $A_1^+(g)$ is dashed to show that it is specified by the regime g and not the natural data-generating process.

$A_2(g)$ is the level of exercise that the individual would perform in the absence of an intervention on Day 2 having followed the regime g_1 and thus having exercised for time $A_1^+(g)$ on Day 1. In general this will not be the same as A_2, which is the number of minutes of exercise the individual would have performed on the second day, having *also* chosen their own level of exercise on Day 1.

Note that in order to avoid further complicating the notation we write $A_2^+(g)$ rather than $A_2^+(g_2(A_2(g), A_1^+(g), A_1))$ or $A_2^+(g_2)$ though one could choose to do so.[52] Such information is present in and simpler to read from the graph $\mathcal{G}(g)$ that we now construct. However, though we will write $A_t^+(g)$ we will require that the function g_t^+ specifying the prescribed assignment at stage t only takes as input variables that are temporally prior to $A_t^+(g)$.

We will allow dynamic regimes under which both the prescribed levels $A_s^+(g)$ for $s < t$ and the 'natural' levels of treatment $A_s(g)$ for $s \leq t$ may be used in determining the current assigned treatment $A_t^+(g)$.[53] We will assume as in (56) that under the dynamic regime g *only* the treatment

[52] For example, if two treatments A_1 and A_2 are dynamically assigned via functions g_1 and g_2, but only A_2 is an ancestor of a response Y, we might wish to make this clear by writing $Y(g_2)$. Further, if g_2 took as input L_0 and A_1 (but not $A_1^+(g_1(A_1))$ or $L_1(g_1(A_1))$) we could write $Y(g_2(A_1))$ to make this clear. However, it is not hard to see that such an explicit notation will quickly become burdensome when variables influenced by earlier dynamic treatments are used as inputs for later dynamic treatments e.g. if written explicitly the final response $Y(g)$ in Figure 21(ii) would be: $Y(a_1, a_2)|_{a_1=g_1(A_1), a_2=g_2\{A_2(g_1(A_1)), L(g_1(A_1)), A_1^+(g_1(A_1)), A_1\}}$ (!)

[53] Robins et al. (2004) and Young et al. (2012) considered the special case in which the only natural level determining $A_t^+(g)$ was $A_t(g)$.

received $A_t^+(g)$ would affect subsequent responses, and not the 'natural' level $A_t(g)$.

Note that in our examples, g_t had only A_t and prior history as arguments, but we can easily imagine settings where g_t might also include a randomizer δ_t, generated by a random number generator, as an additional argument.[54] Note that δ_t unlike the error terms ε_j occurring in an NPSEM are simply determined by the (randomization device of the) person implementing the dynamic regime. They do not constitute an intrinsic feature of the subject. It is for this reason that we assume that the new error terms (δ_t) are jointly independent and independent of the ε_j. (Recall that under the FFRCISTG model, in contrast to the NPSEM-IE, we do not assume that the ε_j are jointly independent.)

5.3 Creating a dynamic regime-specific structural equation model

Formally, given an NPSEM specifying equations for the variables $\{V_m(\widetilde{\mathbf{pa}}_m)\}$, the imposition of the dynamic regime g supplements and modifies this set of equations in the following three steps:

(i) For every variable $V_m \in \mathbf{V}$ replace every occurrence of a_t as an argument on the RHS of the structural equation for $V_m(\mathbf{pa}_m)$ by a_t^+;[55]

$$\begin{aligned} &V_m(\ldots, a_t, \ldots) = f_m(\ldots, a_t, \ldots, \boldsymbol{\epsilon}_m) \\ &\quad \Rightarrow \quad V_m(\ldots, a_t^+, \ldots) = f_m(\ldots, a_t^+, \ldots, \boldsymbol{\epsilon}_m). \end{aligned}$$

(ii) For every variable $A_t \in \mathbf{A} \subseteq \mathbf{V}$, create a new variable A_t^+ and structural equation determined by the regime g:

$$A_t^+(\mathbf{pa}_t^+) = g_t(\mathbf{pa}_t^+, \delta_t). \tag{57}$$

The set $\mathbf{pa}_t^+$ of input variables for g_t are prescribed by the regime g and may include variables a_t, a_s^+, a_s, with $s \neq t$, and v_m.

We require that there is a total ordering of all of the variables in the new system such that the variables that are inputs to a given equation occur earlier in the ordering.[56]

[54] For example, suppose that in a particular implementation the additional exercise time requires the presence of a coach, and thus cannot be offered to all patients. Thus subjects are randomized to have the additional training.

[55] Note that since $\mathbf{A} \subseteq \mathbf{V}$, Step (i) includes the structural equations for treatment variables $A_t(\mathbf{pa}_t)$.

[56] Such an ordering exists if we make the assumption that the function g_t^+ only takes as input variables that are temporally prior to A_t^+.

(iii) The terms δ_t in the new structural equations are generated independently of any error terms ε_j or other terms δ_s.[57]

We then perform recursive substitution using the new NPSEM to express each of the variables V_m or A_t^+ as a function of the error terms $\{\varepsilon_s\}$ and the randomization terms $\{\delta_s\}$, via function $\{f_s\}$ and $\{g_s\}$. We write each of the resulting variables as $V_m(g)$ or $A_t^+(g)$ unless it *solely* determined by functions $\{f_s\}$; note that by definition, $A_t^+(g)$ is determined by g_t.

We will use $\mathbb{A}^+(g)$ to denote the set of variables $\{A_t^+(g)\}$; with slight abuse of notation we will use $\mathbb{V}(g)$ to denote all other counterfactual variables (i.e. those without a '+' superscript) even if they are not labelled with g; finally

$$\mathbb{W}(g) \equiv \mathbb{V}(g) \cup \mathbb{A}^+(g).$$

5.3.1 Example

To take the example of the FFRCISTG model corresponding to the graph $\mathcal{G}$ in Figure 20, we might start with the NPSEM:

$$A_1 = \varepsilon_1, \qquad\qquad Y(a_1) = f(a_1, \varepsilon_Y);$$

under the dynamic regime this is then transformed into:

$$A_1 = \varepsilon_1, \qquad A_1^+(a_1) = g_1(a_1, \delta_1), \qquad Y(a_1^+) = f(a_1^+, \epsilon_Y),$$

from which we obtain by recursive substitution:

$$\begin{aligned} A_1 &= \varepsilon_1, \\ A_1^+(g) &= A_1^+(A_1) = g_1(A_1, \delta_1) = g_1(\varepsilon_1, \delta_1), \\ Y(g) &= f(A_1^+(A_1), \epsilon_Y) = f(g_1(\varepsilon_1, \delta_1), \epsilon_Y). \end{aligned}$$

5.3.2 Special Cases

Notice the following two special cases:

- If for all t, $A_t^+(a_t) = g_t(a_t) = a_t$ then $V_m(g) = V_m$ for every m and every unit in the population. Thus $P(\mathbf{V}) = P(\mathbb{V}(g))$.
- If for all t, $A_t^+ = g_t = \tilde{a}_t$ for some fixed constant $\tilde{a}_t$, so the function g_t takes no arguments as input, then $V_m(g) = V_m(\tilde{\mathbf{a}})$, and $P(\mathbf{V}(\tilde{\mathbf{a}})) = P(\mathbb{V}(g))$.

[57] We say that the regime g is *deterministic* at s if, $g_s(\mathbf{pa}_s^+, \delta_s)$ does not depend on δ_s. The latter is equivalent to having $\delta_s = 0$ (or some other fixed constant) with probability one. We say the regime g is deterministic if this holds for each s.

5.4 Construction of a dynamic Single World Intervention Graph (d-SWIG) $\mathcal{G}(g)$

Our approach is to first construct a SWIG $\mathcal{G}(\mathbf{a})$, but to then add extra edges from the variables that are inputs into the function g_t to the fixed node a_t; we will then relabel each fixed node a by $A^+(g)$.

Definition 24 (The d-SWIG $\mathcal{G}(g)$)**.** *Given a template $\mathcal{G}(\mathbf{a})$ and a dynamic regime g for $\mathbf{A}$, the d-SWIG $\mathcal{G}(g)$ is defined by applying the following transformation:*

(i) *replacing each fixed vertex a_i with a random vertex $A_i^+(g)$ that inherits children from a_i; we will add dashed directed edges (⇢) from every variable that is an input to the function g_i to the variable $A_i^+(g)$;*[58]

(ii) *Each random vertex $V(\mathbf{a}_V)$ that is a descendant of at least one variable $A_i^+(g)$ is relabeled as $V(g)$.*

Note that $\mathcal{G}(g)$ will be a DAG since in clause (ii) of the definition of a dynamic regime-specific structural equation model we require that there is a total ordering of the variables such that the variables that are inputs to a given equation occur earlier in the ordering.

For a given dynamic SWIG $\mathcal{G}(g)$, we will refer to any dynamic regime g' that would have resulted in the same dynamic SWIG as being *compatible* with $\mathcal{G}(g)$. In words, these regimes share in common the variables that are used as input in determining (the distribution of the) the treatment $A_t^+(g)$ prescribed by the regime; however, they may differ on the specific functional forms. For a given d-SWIG $\mathcal{G}(g)$ we will refer to the *class of dynamic regimes g' that are compatible with $\mathcal{G}(g)$*.

The following is an immediate consequence of the construction above.

Proposition 25. *Under the FFRCISTG model associated with $\mathcal{G}$, the dynamic regime-specific NPSEM constructed above exists and will be an FFRCISTG associated with $\mathcal{G}(g)$.*

Note that the vertex set of $\mathcal{G}(g)$ is $\mathbb{W}(g)$. Observe that under the regime where $A_t^+(a_t) = g_t(a_t) = a_t$, then for all t $A_t^+(g) = A_t$. If we re-merge all split nodes and drop all (g) labels we recover the original graph $\mathcal{G}$. Similarly, under the regime where for all t, $g_t = \tilde{a}_t$, the $A_t^+(g)$ nodes become fixed constants, and the dynamic SWIG reduces to a static SWIG $\mathcal{G}(\tilde{\mathbf{a}})$.

[58] We used dashed edges (⇢) here to indicate that these are given by the function g_i

5.5 Factorization and Modularity

Proposition 26. *Under the FFRCISTG model associated with $\mathcal{G}$, for a given dynamic regime g, the distribution $P(\mathbb{W}(g))$ factorizes according to $\mathcal{G}(g)$.*

Proof: By Proposition 25 the dynamic regime-specific NPSEM is an FFRCISTG associated with $\mathcal{G}(g)$ with the 'observed' distribution $P(\mathbb{W}(g))$. The conclusion then follows from Proposition 11.[59] □

For a given dynamic regime g, let $\mathfrak{Q}(g) \equiv \{q_t^g(a_t^+ \mid \mathrm{pa}_t^+),\ \text{for } A_t \in \mathbf{A}\}$ be the set of conditional densities implied by the structural equations $A_t^+(\mathbf{pa}_t^+)$ given by (57). The following is then the natural extension of modularity to dynamic SWIGs:

Definition 27 (Dynamic SWIG Modularity)**.**

The triple $(\mathcal{G}, P(\mathbf{V}), \mathfrak{Q}(g))$ and the pair $(\mathcal{G}(g), P(\mathbb{W}(g)))$ are said to satisfy the dynamic modularity property *if*

(i) *for every $Y(g) \in \mathbb{V}(g)$,*

$$P\Big(Y(g)=y \;\Big|\; \{\mathrm{pa}_{\mathcal{G}(g)}(Y(g)) \setminus \mathbb{A}^+(g)\}=\mathbf{r},\ \{\mathrm{pa}_{\mathcal{G}(g)}(Y(g)) \cap \mathbb{A}^+(g)\}=\mathbf{a}^+\Big) \\ = P\big(Y=y \;\big|\; \{\mathrm{pa}_{\mathcal{G}}(Y) \setminus \mathbf{A}\}=\mathbf{r},\ \{\mathrm{pa}_{\mathcal{G}}(Y) \cap \mathbf{A}\}=\mathbf{a}^+\big)\,, \tag{58}$$

provided that both sides of (58) are well-defined, and

(ii) *for every $A_t^+(g) \in \mathbb{A}^+(g)$,*

$$P\Big(A_t^+(g)=a^+ \;\Big|\; \mathrm{pa}_{\mathcal{G}(g)}(A_t^+(g))=\mathbf{pa}_t^+\Big) = P_{q_t^g}\big(A_t^+=a^+ \;\big|\; \mathbf{PA}_t^+=\mathbf{pa}_t^+\big)\,. \tag{59}$$

In words, this simply states that for any variable in $\mathbb{V}(g)$ the conditional distribution for the variable given its parents in $\mathcal{G}(g)$ under $P(\mathbb{W}(g))$ is equal to that for V given its parents in $\mathcal{G}$ under $P(\mathbf{V})$, while for variables $A_t^+(g)$, the distribution is specified by the regime g itself, so $q_t^g \in \mathfrak{Q}(g)$. We then have:

Proposition 28. *Under the FFRCISTG model associated with $\mathcal{G}$, for a given dynamic regime g, $(\mathcal{G}, P(\mathbf{V}), \mathfrak{Q}(g))$ and $(\mathcal{G}(g), P(\mathbb{W}(g)))$ obey dynamic modularity.*

[59] If g is non-deterministic for each s, d-separation is complete but not otherwise. However no results in this paper depend on whether or not g is deterministic for some s.

Proof: Follows directly from the construction of the regime specific NPSEM, the definition of dynamic modularity and Proposition 16. □

Proposition 29. *If $P(\mathbf{V})$ factorizes according to $\mathcal{G}$ then $P(\mathbb{V}(g), \mathbb{A}^+(g))$ factorizes according to $\mathcal{G}(g)$ and obeys dynamic modularity with respect to $(\mathcal{G}, P(V), \mathfrak{Q}(g))$ if and only if*

$$
\begin{aligned}
&P\left(\mathbb{V}(g) = \mathbf{v}, \mathbb{A}^+(g) = \mathbf{a}^+\right) \qquad (60)\\
&= \prod_{V_i \in \mathbf{V}} P\Big(V_i = v_i \;\Big|\; \{\mathrm{pa}_{\mathcal{G}}(V_i) \setminus \mathbf{A}\} = \mathbf{v}_{\mathrm{pa}_{\mathcal{G}}(V_i)\setminus\mathbf{A}},\ \{\mathrm{pa}_{\mathcal{G}}(V_i) \cap \mathbf{A}\} = \mathbf{a}^+_{\mathrm{pa}_{\mathcal{G}}(V_i)\cap\mathbf{A}}\Big)\\
&\quad \times \prod_{A_t^+(g) \in \mathbb{A}^+(g)} P_{q_t^g}\Big(A_t^+(g) = a_t^+ \;\Big|\; \{\mathrm{pa}_{\mathcal{G}(g)}(A_t^+(g)) \setminus \mathbb{A}^+(g)\} = \mathbf{v}_{\mathrm{pa}_{\mathcal{G}(g)}(A_t^+(g))\setminus\mathbb{A}^+(g)},\\
&\qquad\qquad \{\mathrm{pa}_{\mathcal{G}(g)}(A_t^+(g)) \cap \mathbb{A}^+(g)\} = \mathbf{a}^+_{\mathrm{pa}_{\mathcal{G}(g)}(A_t^+(g))\cap\mathbb{A}^+(g)}\Big).
\end{aligned}
$$

[60] Under the conditions of Proposition 29, equation (60) identifies the distribution resulting from the dynamic regime whenever the RHS is a well-defined functional of the observed distribution $P(\mathbf{V})$.

5.5.1 Identification via the extended g-formula

Now consider a DAG $\mathcal{G}$ with vertex set $\mathbf{V} = \mathbf{O} \cup \mathbf{H}$ where $\mathbf{O}$ and $\mathbf{H}$ are, respectively, subsets of observed and hidden variables.

Definition 30 (The dynamic extended g-formula)**.** *Consider a DAG $\mathcal{G}$ with node set $\mathbf{V} = \mathbf{O} \cup \mathbf{H}$, regime g, and dSWIG $\mathcal{G}(g)$ under which there exists a topological ordering:*[61]

$$\langle L_1(g), A_1(g), A_1^+(g), \ldots, L_K(g), A_K(g), A_K^+(g), Y(g)\rangle,$$

[60] As noted earlier, (37) corresponded to the special case of (60) where the intervention density q_i for A_i is a point-mass on $\tilde{a}_i$, i.e. a static regime.

[61] Here, with slight abuse of notation, we use $A_j(g)$ as a short hand for '$A_j(g)$ or A_j depending on whether A_j depends on some function g_t in the dynamic regime specific NPSEM corresponding to g'. Likewise for $L_i(g)$. Here $L_1(g)$ and $A_1(g)$ would simply be L_1 and A_1 in $\mathcal{G}(g)$.

of the set $\mathbb{O}(g) \cup \mathbb{A}^+(g)$ *in the dSWIG* $\mathcal{G}(g)$.[62] *Define:*

$$f^g(y, \bar{\mathbf{l}}_K, \bar{\mathbf{a}}_K, \bar{\mathbf{a}}_K^+) \equiv p(y|\bar{\mathbf{l}}_K, \bar{\mathbf{a}}_K^+) \prod_{j=1}^{K} p(l_j, a_j|\bar{\mathbf{l}}_{j-1}, \bar{\mathbf{a}}_{j-1}^+) \prod_{t=1}^{K} q_t^g(a_t^+ \mid \mathbf{pa}_t^+), \tag{61}$$

here $q_t^g \in \mathfrak{Q}(g)$ *is the density corresponding to* g_t*;* $\mathbf{pa}_t^+$ *is the subvector of* $(\bar{\mathbf{a}}_{t-1}^+, \bar{\mathbf{l}}_t, \bar{\mathbf{a}}_t)$ *of inputs to* g_t*.*

The right-hand side of (61) is the dynamic extended g-formula density of Robins et al. (2004, p.2222) associated with $(\mathbf{A}, \mathbf{O})$ and the densities $\mathfrak{Q}(g)$ in the context of graph $\mathcal{G}$. In the absence of hidden variables, so that $\mathbf{V} = \mathbf{O}$, and (61) being a well-defined function of the distribution of $P(\mathbf{O})$ then (61) and the right-hand side of (60) are equal. Young et al. (2012) gave a positivity condition that guarantees (61) is well-defined. To describe that condition note the distribution obtained from (61) by marginalizing over $\bar{\mathbf{a}}_K$ is:

$$\begin{aligned}
&f^g(y, \bar{\mathbf{l}}_K, \bar{\mathbf{a}}_K^+) \\
&\quad = p(y|\bar{\mathbf{l}}_K, \bar{\mathbf{a}}_K^+) \left(\prod_{j=1}^{K} p(l_j|\bar{\mathbf{l}}_{j-1}, \bar{a}_{j-1}^+) \right) \sum_{\bar{\mathbf{a}}_K} \prod_{j=1}^{K} q_j^g(a_j^+ \mid \mathbf{pa}_j^+) p(a_j|\bar{\mathbf{l}}_j, \bar{\mathbf{a}}_{j-1}^+) \\
&\quad = p(y|\bar{\mathbf{l}}_K, \bar{\mathbf{a}}_K^+) \prod_{j=1}^{K} p(l_j|\bar{\mathbf{l}}_{j-1}, \bar{\mathbf{a}}_{j-1}^+) \left\{ \prod_{j=1}^{K} \widetilde{q}_j^{g,p}(a_j^+|\bar{\mathbf{a}}_{j-1}^+, \bar{\mathbf{l}}_j) \right\},
\end{aligned}$$

where, for $m = 1, \ldots, K$, $\widetilde{q}_j^{g,p}$ is defined recursively in terms of q_j^g and $p(a_j|\bar{\mathbf{l}}_j, \bar{\mathbf{a}}_{j-1}^+)$ by

$$\begin{aligned}
&\widetilde{q}_m^{g,p}\left(a_m^+|\bar{\mathbf{a}}_{m-1}^+, \bar{\mathbf{l}}_m\right) \\
&\quad = \left\{ \prod_{j=1}^{m-1} \widetilde{q}_j^{g,p}\left(a_j^+|\bar{\mathbf{a}}_{j-1}^+, \bar{\mathbf{l}}_j\right) \right\}^{-1} \left\{ \sum_{\bar{\mathbf{a}}_m} \prod_{j=1}^{m} q_j^g\left(a_j^+|\mathbf{pa}_j^+\right) p\left(a_j|\bar{\mathbf{l}}_j, \bar{\mathbf{a}}_{j-1}^+\right) \right\}.
\end{aligned}$$

[62] Note that here we implicitly restrict attention to orderings of the variables under which A_t directly precedes A_t^+. In general one might wish to consider more general orderings of the set of variables $\mathbb{V}(g), \mathbb{A}^+(g)$ as there are distributions $P(Y(g))$ that are only identified via g-formulae under these more general orderings. Since, as already noted in §4.1.2, the g-formula is not a complete algorithm for identification even for static regimes, we choose to defer the consideration of these more general g-formulae to the broader problem of constructing a complete algorithm for the dynamic regimes that we consider.

Then for any $\mathbf{Z} \subseteq \mathbf{O} \cup \overline{\mathbf{A}}_K^+$, $f^g(\mathbf{z})$ is well-defined if for all $m, \overline{\mathbf{a}}_m^+, \overline{\mathbf{l}}_m$,

$$\widetilde{q}_m^{g,p}\left(a_m^+ | \overline{\mathbf{a}}_{m-1}^+, \overline{\mathbf{l}}_m\right) > 0 \;\Rightarrow\; p\left(a_m^+ | \overline{\mathbf{a}}_{m-1}^+, \overline{\mathbf{l}}_m\right) > 0, \tag{62}$$

(Young et al., 2012).[63] Equation (62) is equivalent to the statement that

$$f^g\left(\overline{\mathbf{a}}_m^+, \overline{\mathbf{l}}_m\right) > 0 \;\Rightarrow\; p\left(\overline{\mathbf{a}}_m^+, \overline{\mathbf{l}}_m\right) > 0.$$

Here

$$p\left(a_m^+ \mid \overline{\mathbf{a}}_{m-1}^+, \overline{\mathbf{l}}_m\right) = P(A_m = a_m^+ \mid \overline{\mathbf{A}}_{m-1} = \overline{\mathbf{a}}_{m-1}^+, \overline{\mathbf{L}}_m = \overline{\mathbf{l}}_m).$$

Note that (62) suffices to make $f^g(\mathbf{z})$ well defined even if $\mathbf{Z} \cap \overline{\mathbf{A}}_K \neq \emptyset$. As in our discussion of g-formulae for static regimes, there is no requirement that the ordering of the sequence of variables be temporal. Again, by analogy with our earlier discussion we suppose that the ordering of the variables given by time constitutes one topological ordering of the graph $\mathcal{G}(g)$, though, as before, there may exist other topological orderings that are non-temporal.

5.6 Conditions for identification of a dynamic regime via a dynamic extended g-formula

Young et al. (2012) noted that when a regime g depends on the natural value of treatment the identification of the marginal distribution of $Y(g)$ by the marginal distribution of Y under the dynamic extended g-formula (61) requires an additional condition that is not needed when the regime only depends on past values of $\mathbf{L}$ and $\mathbf{A}^+$. Specifically they noted that the natural value of treatment should not itself be a confounder. We now provide explicit conditions for identification and confirm their insight. However, as with our earlier discussion of static regimes, in the presence of unobserved variables, our conditions are not complete.

Suppose that for all $\mathbf{a}^*$, $P(\mathbf{V})$ and $P(\mathbb{V}(\mathbf{a}^*))$ factor according to $\mathcal{G}$ and $\mathcal{G}(\mathbf{a}^*)$ respectively, and obey modularity. Consider a topological ordering of the graph $\mathcal{G}(g)$. Let $\overline{\mathbb{L}}_j(g)$, $\overline{\mathbb{A}}_j(g)$ and $\overline{\mathbb{A}}_j^+(g)$ indicate initial sequences under this ordering of the variables in the sets $\mathbb{L}(g)$, $\mathbb{A}(g)$ and $\mathbb{A}^+(g)$ respectively.

[63] For a single time-independent treatment Muñoz and van der Laan (2011) and Haneuse and Rotnitzky (2013) obtained similar results. Muñoz and van der Laan (2012) consider longitudinal data but appear to analyze it as a collection of single time-independent treatments, as opposed to the truly time-dependent analysis given in Young et al. (2012).

We first define the following sets for $k = 1, \ldots, K$:[64]

$$\mathbb{Z}_k(g) \equiv \left(\text{an}_{\mathcal{G}(g)}(Y(g))\right) \setminus \left(\overline{\mathbb{L}}_k(g) \cup \overline{\mathbb{A}}_k(g) \cup \mathbb{A}^+(g)\right) \tag{63}$$

and $\mathbb{Z}_k(\mathbf{a}^*) \equiv \{V(\mathbf{a}^*) \mid V(g) \in \mathbb{Z}_k(g)\}$, the subset of (random) vertices in $\mathcal{G}(\mathbf{a}^*)$ that correspond to the vertices in $\mathbb{Z}_k(g)$. Finally, let

$$\mathbb{Z}(g) \equiv \left(\text{an}_{\mathcal{G}(g)}(Y(g))\right) \setminus \mathbb{A}^+(g),$$

and

$$f^g(\mathbf{z}) \equiv \sum_{\overline{\mathbf{a}}_K^+, \mathbf{s}} f^g(y, \bar{\mathbf{l}}_K, \overline{\mathbf{a}}_K, \overline{\mathbf{a}}_K^+), \tag{64}$$

where $\mathbf{S} = \mathbf{O} \setminus \mathbf{Z}$. We then have the following:

Theorem 31. *Assuming* (62)*, for any dynamic regime g compatible with $\mathcal{G}(g)$, if for all $\mathbf{a}^*$ and all $k = 1, \ldots, K$,*

$$\mathbb{Z}_k(\mathbf{a}^*) \perp\!\!\!\perp I(A_k(\mathbf{a}^*) = a_k^*) \mid \overline{\mathbb{L}}_k(\mathbf{a}^*), \overline{\mathbb{A}}_{k-1}(\mathbf{a}^*) = \overline{\mathbf{a}}_{k-1}^*, \tag{65}$$

then

$$f^g(\mathbf{z}) = P(\mathbb{Z}(g) = \mathbf{z}). \tag{66}$$

In particular,

$$P(Y(g) = y) = f^g(y),$$

where

$$f^g(y) = \sum_{\bar{\mathbf{l}}_K, \overline{\mathbf{a}}_K, \overline{\mathbf{a}}_K^+} f^g(y, \bar{\mathbf{l}}_K, \overline{\mathbf{a}}_K, \overline{\mathbf{a}}_K^+). \tag{67}$$

Arguing as in Young et al. (2012) we have the following:

Corollary 32. [65] *Assuming* (62)*, and the premise of* Theorem 31*,*

$$pr[Y(g) = y] = E\left[I\left\{Y(g) = y\right\} W^g\right] \tag{68}$$

where

$$W^g = \frac{\prod_{j=1}^{K} \tilde{q}_j^{g,p}(A_j | \overline{\mathbf{A}}_{j-1}, \overline{\mathbf{L}}_j)}{\prod_{j=1}^{K} p(A_j | \overline{\mathbf{A}}_{j-1}, \overline{\mathbf{L}}_j)}.$$

[64] Note that by definition of 'ancestor', $Y(g) \in \text{an}_{\mathcal{G}(g)}(Y(g))$.

[65] Analogous results were obtained by Haneuse and Rotnitzky (2013) and Muñoz and van der Laan (2011) for the case of a single time independent treatment.

Proof: This follows from the Radon-Nikodym Theorem since $W(g)$ is the likelihood ratio:

$$\left(f^g(Y, \overline{\mathbf{L}}_K, \overline{\mathbf{a}}_K^+)\big|_{\overline{\mathbf{a}}_K^+ = \overline{\mathbf{A}}_k} \right) \Big/ p(\mathbf{O}),$$

which by (62) is finite with probability one under $P(\mathbf{O})$. □

Here (68) is the Inverse Probability Weighted version of the dynamic extended g-formula for the distribution of $Y(g)$. See Young et al. (2012) for further development.

According to Theorem 31, to check identification of $Y(g)$ we proceed in two steps using two different graphs. For each k:

$\mathcal{G}(g)$: We find the set $\mathbb{Z}_k(g)$ of ancestors of $Y(g)$ in $\mathcal{G}(g)$ that are not in $\overline{\mathbb{L}}_k(g) \cup \overline{\mathbb{A}}_k(g) \cup \mathbb{A}^+(g)$;

$\mathcal{G}(\mathbf{a}^*)$: We find the corresponding set $\mathbb{Z}_k(\mathbf{a}^*)$ and test whether $A_k(\mathbf{a}^*)$ is d-separated from $\mathbb{Z}_k(\mathbf{a}^*)$ given $\overline{\mathbb{L}}_k(\mathbf{a}^*) \cup \overline{\mathbb{A}}_{k-1}(\mathbf{a}^*) \cup \mathbf{a}^*$ in $\mathcal{G}(\mathbf{a}^*)$.

In fact, this test may be performed directly on the graph $\mathcal{G}(g)$ if we treat dashed edges as not being present for the purpose of testing the d-separation relation.[66] The proof of Theorem 31 is given in Appendix B.3.

5.6.1 Examples

Consider the DAG $\mathcal{G}$ in Figure 19(i) and the dynamic regime g represented by the dSWIG $\mathcal{G}(g)$ shown in Figure 21(i). In this example the first treatment $A_1^+(g)$ depends on the natural level A_1, while the second treatment $A_1^+(g)$ depends on both the natural and assigned levels of treatment at time 1, and on the intermediate $L(g)$. However, $A_2^+(g)$ does not depend on the natural level $A_2(g)$. We thus find:

$$\mathbb{Z}(g) = \{A_1, L(g), Y(g)\}, \quad \mathbb{Z}_1(g) = \{Y(g), L(g)\} \text{ and } \mathbb{Z}_2(g) = \{Y(g)\}$$

for the g depicted in Figure 21(i). Examining the template $\mathcal{G}(a_1, a_2)$ shown in Figure 21(iii) shows that:

$$\{Y(\mathbf{a}^\dagger), L(\mathbf{a}^\dagger)\} \perp\!\!\!\perp A_1 \quad \text{and} \quad Y(\mathbf{a}^\dagger) \perp\!\!\!\perp A_2(\mathbf{a}^\dagger) \mid L(\mathbf{a}^\dagger), A_1; \tag{69}$$

[66] Formally an edge (⇢) may go from a temporally later variable to an earlier A^+. In that case it then follows from Theorem 31 that if (65) holds the distribution of $Y(g)$ is identified even though the dynamic regime g cannot be implemented in practice since it would imply that randomization probabilities at time t depend on variables that are only observed subsequent to time t. Note that this cannot arise in the absence of individual level exclusion restrictions: in this case $\mathcal{G}$ will be complete and the temporal ordering will be the only topological ordering of the graph $\mathcal{G}$; see also §7.

recall that in order to check this on the graph $\mathcal{G}(a_1, a_2)$ we also condition on a_1 and a_2. Consequently, it follows from Theorem 31 that we have the following identifying formula for $P(\mathbb{Z}(g))$:

$$\begin{aligned} &P(A_1 = a_1, L(g) = l, Y(g) = y) \\ &\quad = \sum_{a_2, a_1^+, a_2^+} p(y|a_2^+, l, a_1^+) q(a_2^+|l, a_1^+, a_1) p(l|a_1^+) q(a_1^+|a_1) p(a_1), \end{aligned}$$

from which we may obtain the following:

$$\begin{aligned} &P(Y(g) = y) \\ &\quad = \sum_{a_1, a_1^+, l, a_2, a_2^+} p(y|a_2^+, l, a_1^+) q(a_2^+|l, a_1^+, a_1) p(l|a_1^+) q(a_1^+|a_1) p(a_1), \\ &P(Y(g) = y \mid A_1 = a_1) \\ &\quad = \sum_{a_1^+, l, a_2^+} p(y|a_2^+, l, a_1^+) q(a_2^+|l, a_1^+, a_1) p(l|a_1^+) q(a_1^+|a_1), \\ &P(Y(g) = y \mid A_1 = a_1, A_1^+(g) = a_1^+, L(g) = l) \\ &\quad = \sum_{a_2^+} p(y|a_2^+, l, a_1^+) q(a_2^+|l, a_1^+, a_1). \end{aligned}$$

Note that these identification results hold even though $Y(g) \not\perp\!\!\!\perp A_1$, as shown by the fact that there is a directed path from A_1 to $Y(g)$ in $\mathcal{G}(g)$ shown in Figure 21(i). Inspection of the dSWIG in Figure 21(i) shows that

$$A_2(g) \perp\!\!\!\perp Y(g) \mid A_1, A_1^+(g), L(g)$$

hence, further we obtain:

$$\begin{aligned} &P(Y(g) = y \mid A_1 = a_1, A_1^+(g) = a_1^+, L(g) = l, A_2(g) = a_2) \\ &\quad = P(Y(g) = y \mid A_1 = a_1, A_1^+(g) = a_1^+, L(g) = l). \end{aligned}$$

Now consider the dynamic regime represented by the graph $\mathcal{G}(g)$ shown in Figure 21(ii). In this example the regime for the second treatment $A_2^+(g)$ depends on the natural level of the second treatment $A_2(g)$ after having already implemented the regime at the first time point. Consequently, we now have:

$$\mathbb{Z}(g) = \{A_1, L(g), A_2(g), Y(g)\}, \;\; \mathbb{Z}_1(g) = \{Y(g), L(g), A_2(g)\}, \;\; \mathbb{Z}_2(g) = \{Y(g)\}.$$

However, A_1 is not d-separated from $A_2(a_1)$ in the graph $\mathcal{G}(a_1, a_2)$ shown in Figure 21(iii), since the path

$$A_1 \leftarrow H_1 \rightarrow A_2(a_1)$$

d-connects (given a_1 and a_2). Due to the presence of the unmeasured confounder H_1, in general the distribution $P(A_2(a_1))$ is not identified. Consequently there are dynamic regimes g compatible with $\mathcal{G}(g)$ under which $P(Y(g))$ is not identified. This last conclusion follows since d-separation is a complete criterion for independence on a DAG, hence d-separation is also complete for independence among the counterfactual variables in a dSWIGs $\mathcal{G}(g)$, letting the dynamic regime g vary over all compatible regimes. Following Young et al. (2012), we say the lack of identification is due to 'confounding by the natural value of treatment'.

Consider now the random dynamic regime g_{rand} in which for $m = 1, 2$, A_m^+ is drawn from $\widetilde{q}_m^{g,p}(a_m^+ | \bar{\mathbf{a}}_{m-1}^+, \bar{\mathbf{l}}_m)$; see (Young et al., 2012). Note that the regime g_{rand} does not depend on a subject's 'natural' values $A_m(g_{\text{rand}})$, so that there can no longer be confounding by the natural value of treatment and the distribution of $Y(g_{\text{rand}})$ should be identified. Indeed it follows from Figure 19 and Corollary 34 (below) that the distribution of $Y(g_{\text{rand}})$ is identified by the dynamic extended g-formula (67), even though, as we have just seen, the distribution of $Y(g)$ is not identified.

5.7 Reformulation of the identification condition via perturbed regimes

The identification condition (65) given in Theorem 31 requires one to check d-separation between each $A_k(\mathbf{a}^*)$ and the set of vertices $\mathbb{Z}_k(\mathbf{a}^*)$. The next result shows that by considering a set of modified regimes we can re-formulate the condition in terms of the existence of paths between (the vertices corresponding to) A_k and Y.

Specifically, given a dynamic regime g we will define for $j = 1, \ldots, K$, the *perturbed regime* g_{-j} to be the regime obtained from g by replacing a_j by a fixed number $\tilde{a}_j$ whenever it is present in any equation:

$$A_k^+(\ldots, a_j, \ldots) = g_k(\ldots, a_j, \ldots) \quad \Rightarrow \quad A_k^+(\ldots, \tilde{a}_j, \ldots) = g_k(\ldots, \tilde{a}_j, \ldots). \tag{70}$$

Note that it follows from the construction of the dynamic regime specific NPSEM that a_j is only present in structural equations for other variables A_k^+ (since in the construction a_j is replaced by a_j^+ for all other variables). It follows that the corresponding dynamic SWIG $\mathcal{G}(g_{-j})$ is obtained from $\mathcal{G}(g)$ by simply removing any (dashed) edges that are directed out of $A_j(g)$.

Lemma 33. *The following conditions are equivalent:*

(i) *In* $\mathcal{G}(\mathbf{a}^\dagger)$, $A_k(a^\dagger)$ *is d-separated from* $\mathbb{Z}_k(\mathbf{a}^\dagger)$ *by* $\overline{\mathbb{A}}_{k-1}(\mathbf{a}^\dagger) \cup \overline{\mathbb{L}}_k(\mathbf{a}^\dagger) \cup \mathbf{a}^\dagger$;

(ii) *In* $\mathcal{G}(g_{-k})$, $A_k(g_{-k})$ *is d-separated from* $Y(g_{-k})$ *given* $\overline{\mathbb{A}}_{k-1}(g_{-k}) \cup \overline{\mathbb{L}}_k(g_{-k}) \cup \overline{\mathbb{A}}^+_{k-1}(g_{-k})$.

Note that condition (ii) is equivalent to checking in $\mathcal{G}(g)$ for the absence of a d-connecting path between $A_k(g)$ and $Y(g)$ that begins $A_k(g) \leftarrow \cdots$, sometimes referred to as a 'backdoor path'. See Appendix B.4 for the proof.

5.7.1 Special case: Regimes in which treatment assignment does not depend on natural treatment levels

In the special case where treatment at time k, $A^+_k(g)$, only depends on $\overline{\mathbb{L}}_k(g) \cup \overline{\mathbb{A}}^+_{k-1}(g)$, but does *not* depend on any of the 'natural' treatment levels ($\overline{\mathbb{A}}_k(g)$) then there is a simple identification condition that solely requires testing d-separation in $\mathcal{G}(g)$:

Corollary 34. *In a graph* $\mathcal{G}(g)$, *if there are no edges from any* $A^-_k(g)$ *to any* $A^+_j(g)$, *then the following are equivalent:*

(i) *In* $\mathcal{G}(\mathbf{a})$, $A_k(a)$ *is d-separated from* $\mathbb{Z}_k(\mathbf{a})$ *given* $\overline{\mathbb{A}}_{k-1}(\mathbf{a}) \cup \overline{\mathbb{L}}_k(\mathbf{a}) \cup \mathbf{a}$;

(ii) *In* $\mathcal{G}(g)$, $A_k(g)$ *is d-separated from* $Y(g)$ *given* $\overline{\mathbb{A}}_{k-1}(g) \cup \overline{\mathbb{L}}_k(g) \cup \overline{\mathbb{A}}^+_{k-1}(g)$.

Thus, if condition (ii) *holds then the identifying formula* (66) *for the distribution of* $Y(g)$ *holds.*[67]

Proof: This follows immediately from Lemma 33 and the observation that if there are no edges from $A^-_k(g)$ to $A^+_j(g)$ then $\mathcal{G}(g_{-k}) = \mathcal{G}(g)$. □

6 NPSEM-IE: Doubly-exponentially many additional untestable assumptions

To bring the distinction between the FFRCISTG and NPSEM-IE models into sharper focus we consider the particular case of a complete graph $\mathcal{G}$

[67] Robins (1997, Theorem 4.2) and Dawid and Didelez (2008) also gave a graphical condition that implied the identifying formula (66). However, unlike the graphical condition (ii) given here, their condition required K graphs to be consulted.

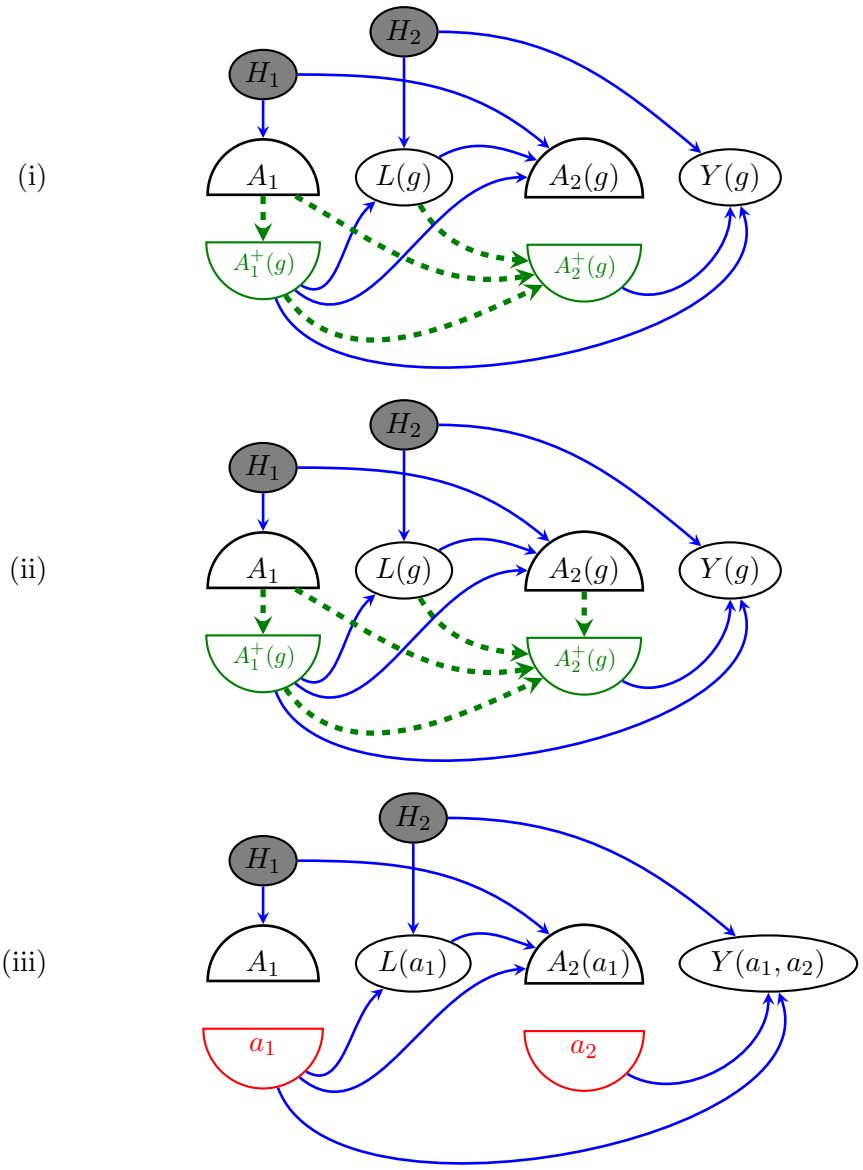


Figure 21: (i) and (ii) show graphs $\mathcal{G}(g)$ derived from the template in Figure 19(ii). In (i) $A_2^+(g)$ is not a function of $A_2(g)$, and $P(Y(g))$ is identified. This identification fails in (ii) since $A_2^+(g)$ is a function of $A_2(g)$, thus $A_2(g) \in \mathbb{Z}_0(g)$. (iii) shows the corresponding template $\mathcal{G}(a_1, a_2)$ used to test the d-separation relation (65) in Theorem 31.

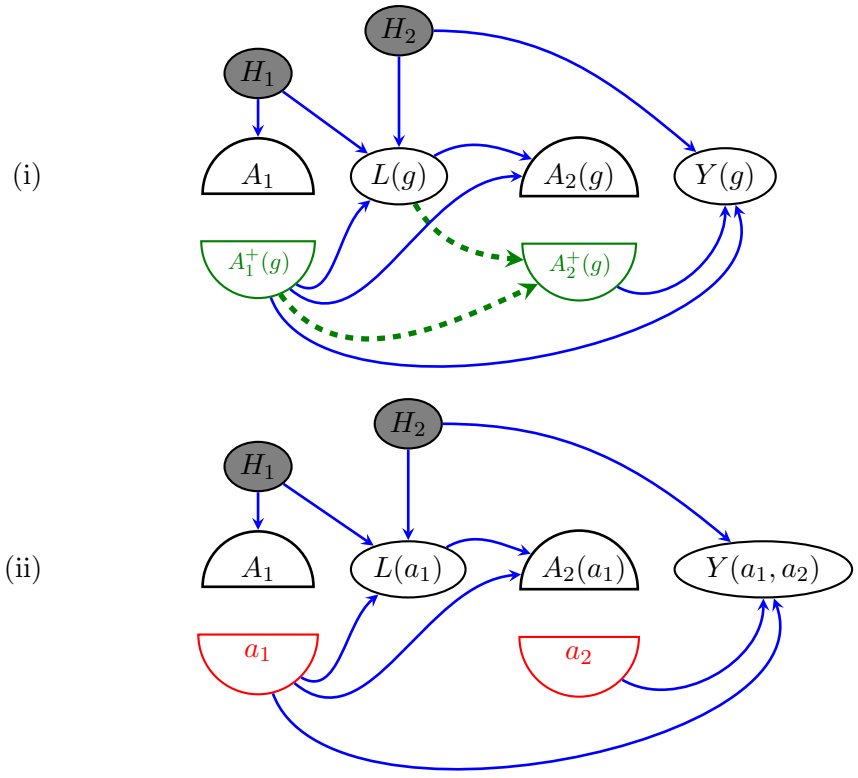


Figure 22: (i) the d-SWIG $\mathcal{G}(g)$ derived from the template in Figure 13(ii) under a dynamic regime g for which treatment at the second time $A_2^+(g)$ depends on past treatment and covariate history ($A_1^+(g)$ and $L(g)$). $P(Y(g))$ is not identified since $A_2^+(g)$ is a function of $L(g)$ and hence $L(g) \in \mathbb{Z}_0(g)$, but there is a d-connecting path from A_1 to $L(a_1)$ in $\mathcal{G}(a_1, a_2)$, as shown in (ii). Thus the d-separation relation (65) in Theorem 31 does not hold. Since g here does not depend on A_1 or $A_2(g)$ we may also apply Corollary 34 to $\mathcal{G}(g)$ directly: Since there is a path d-connecting A_1 and $Y(g)$ (given $\emptyset$) condition (iii) fails to hold for $k = 1$.

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

Table 2: Dimensions of counterfactual models associated with complete graphs with binary variables.

with all variables binary. In this case $\text{pa}_{\mathcal{G}}(V_k) = \{V_1, \ldots, V_{k-1}\}$ hence there are 2^{k-1} counterfactual variables associated with V_k:

$$V_k(0, \ldots, 0), V_k(0, \ldots, 0, 1), \ldots, V_k(1, \ldots, 1).$$

Thus in total there are $2^K - 1$ primitive counterfactual variables associated with the NPSEM for $\mathcal{G}$. It follows that in the absence of any restrictions, the set of all joint distributions:

$$P(\{V_j(\overline{\mathbf{a}}^{\dagger}_{j-1}) ; 1 \leq j \leq k, \overline{\mathbf{a}}^{\dagger}_{j-1} \in \overline{\mathfrak{A}}_{j-1}\})$$

is of dimension $2^{2^K-1} - 1$.

Recall that the defining independence (17) for an FFRCISTG simply requires that the set of counterfactual variables that are compatible with the same intervention:

$$\{V_1, V_2(a_1^{\dagger}), \ldots, V_K(\overline{\mathbf{a}}^{\dagger}_{K-1})\}$$

are jointly independent. Notice that this is exactly the set of independence relations represented in the graphs $\mathcal{G}(\mathbf{a}^{\dagger})$. For the case where $K = 3$, the independence structure is illustrated in Figure 23. The graphs shown in Figure 23(c),(d),(e) use bi-directed edges ($\leftrightarrow$) to represent marginal independence.[68] The independence restrictions present in the FFRCISTG model

[68]The independence structure may be read from these graphs by simply applying the natural extension of the d-separation whereby every (non-endpoint) vertex on a path is a collider, and there are no non-trivial ancestors, so $\text{an}(V) = \{V\}$. Thus, in particular, if two vertices are not joined by a bi-directed edge then they are marginally independent; see Drton and Richardson (2008) for more details.

(with $K = 3$) are shown via the five graphs in Figure 23(c). It is impossible to represent the structure using a single graph because, although the model requires $M(z_0) \perp\!\!\!\perp Y(z_0, m_0)$ and $M(z_0) \perp\!\!\!\perp Y(z_0, m_1)$ it does not imply $M(z_0) \perp\!\!\!\perp Y(z_0, m_0), Y(z_0, m_1)$.[69] Thus in particular, the independence structure does not obey the composition axiom:

$$A \perp\!\!\!\perp B \quad \text{and} \quad A \perp\!\!\!\perp C \quad \Rightarrow \quad A \perp\!\!\!\perp B, C.$$

All existing approaches to representing conditional independence structure via pathwise Markov properties obey this axiom.[70]

We finesse this problem here by indicating the independence structure on the specific margins associated with variables that are observed (the four graphs containing three variables). We also include a single complete graph including all variables in order to emphasize that joint independence relations such as $M(z_0) \perp\!\!\!\perp Y(z_0, m_0), Y(z_0, m_1)$, do *not* hold even though the marginal independencies $M(z_0) \perp\!\!\!\perp Y(z_0, m_i)$ for $i = 0, 1$ hold under the FFRCISTG.[71]

In Figure 23(d) we show the smallest bi-directed graphical model that contains the FFRCISTG. This example is important because it shows that there are data-generating processes that do not involve determinism or parametric cancellation in which cross-world independence relations such as $M(z_0) \perp\!\!\!\perp Y(z_1, m_0)$ do not hold.

It follows from the Möbius parametrization given in (Drton and Richardson, 2008) that such a distribution is defined by a set of equations of the form:

$$P(V_k(\overline{\mathbf{a}}^{\dagger}_{k-1}) = 0, \mathbb{W}(\overline{\mathbf{a}}^{\dagger}_{k-2}) = \mathbf{0}) = P(V_k(\overline{\mathbf{a}}^{\dagger}_{k-1}) = 0) P(\mathbb{W}(\overline{\mathbf{a}}^{\dagger}_{k-2}) = \mathbf{0}) \quad (71)$$

where $\mathbb{W}(\overline{\mathbf{a}}^{\dagger}_{k-2})$ is a non-empty subset of $\{V_1, V_2(a_1^{\dagger}), \ldots, V_{k-1}(\overline{\mathbf{a}}^{\dagger}_{k-2})\}$. Since, choosing $\overline{\mathbf{a}}^{\dagger}_{k-1}$, there are 2^{k-1} random variables $V_k(\overline{\mathbf{a}}^{\dagger}_{k-1})$, and then for fixed $\overline{\mathbf{a}}^{\dagger}_{k-1}$ there are $2^{k-1} - 1$ non-empty subsets $\mathbb{W}(\overline{\mathbf{a}}^{\dagger}_{k-2})$, we see that the set of joint distributions in the FFRCISTG model is of dimension:

$$(2^{(2^K-1)} - 1) - \sum_{j=1}^{K} 2^{j-1}(2^{j-1} - 1) = (2^{(2^K-1)} - 1) - \sum_{j=1}^{K-1} (4^j - 2^j).$$

[69] Here $Y(z_0, m_1)$ is an abbreviation for $Y(z{=}0, m{=}1)$, likewise for $M(z_i)$.

[70] In this respect, the claim of (Pearl, 2009, p.353, fn.12) that "with all due respect to multiculturalism, all approaches to causation are variants or abstractions of the structural theory presented in this book." seems incorrect.

[71] As we have emphasized earlier, elsewhere in this paper we do not construct graphs including potential outcomes corresponding to interventions from different worlds. We construct such graphs here, solely for the purpose of contrasting the FFRCISTG and the NPSEM-IE.

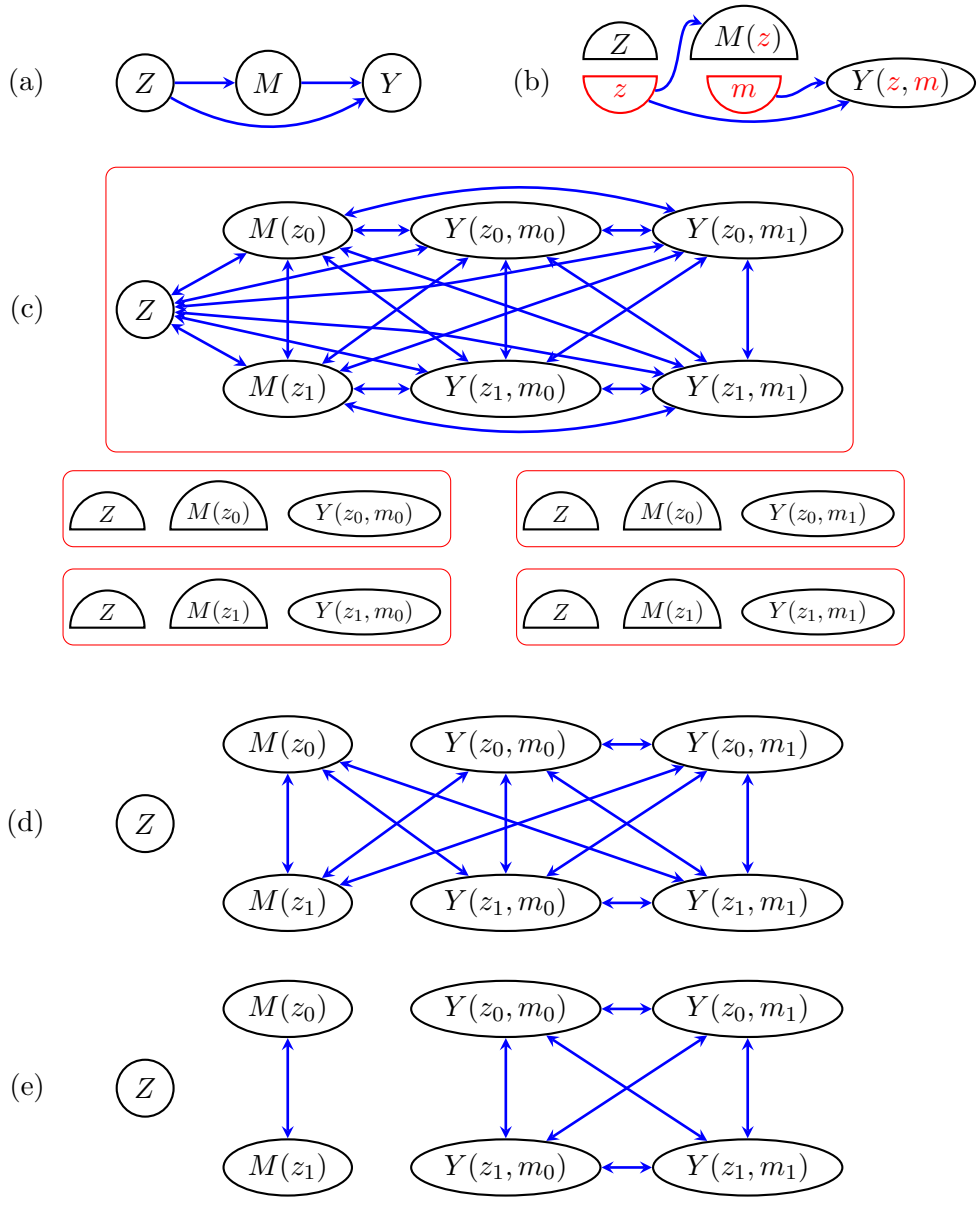


Figure 23: (a) A DAG $\mathcal{G}$; (b) the template $\mathcal{G}(z, m)$; Bi-directed graphs indicating the independence assumptions on the one-step ahead counterfactuals present (with binary variables) in: (c) The FFRCISTG model associated with $\mathcal{G}$: this is a non-graphical Markov model represented here via separate graphs for separate margins (semi-circles indicate SWIGs, with fixed nodes omitted). (d) the smallest graphical marginal independence model containing the FFRCISTG; (e) the NPSEM-IE. Here 'z_i' abbreviates '$z=i$', likewise for m_i.

In contrast, the NPSEM-IE model supposes that sets of counterfactual variables associated with distinct actual variables are independent; see Figure 23(e). As the graph suggests we may group together the 2^{j-1} counterfactual random variables associated with V_j, simply regarding each group as a single single random variable with a state space of cardinality $2^{2^{j-1}}$; the NPSEM-IE requires that these 'grouped variables' be jointly independent. It follows from this directly that the dimension of the NPSEM is:

$$\sum_{j=1}^{K}(2^{2^{j-1}} - 1).$$

Further, we have the following relationship between the dimensions of the models corresponding to complete graphs with K variables in the case where all variables are binary:

$$\dim(\text{FFRCISTG}(K)) - \dim(\text{NPSEM-IE}(K)) \geq 2^{(2^K-2)} \tag{72}$$

for $K > 2$. Thus the number of additional counterfactual independence assumptions present in an NPSEM-IE, but not in the FFRCISTG grows at a doubly-exponential rate. These results, together with those for $K = 2, 3, 4$ are summarized in Table 6.

The doubly exponential growth in the number of untestable assumptions is of great significance given that, as we have discussed, these additional restrictions are not experimentally testable.

6.1 Differences in identification results between FFRCISTGs and NPSEM-IEs

In this section we further examine differences between the FFRCISTG model and the NPSEM-IE by studying causal effects that are identified and counterfactual independencies that hold under the latter but not the former model.

6.1.1 Cross-world independence

Consider the independence:

$$Y(z{=}1, m{=}0) \perp\!\!\!\perp M(z{=}0) \mid Z, \tag{73}$$

in the context of the DAG shown in Figure 9(i). (73) is a 'cross-world independence' and thus is not implied by an FFRCISTG model as there is no SWIG that contains both $Y(z = 1, m = 0)$ and $M(z = 0)$.

The following is an example of a distribution contained in the FFRCISTG model (and thus obeying factorization and modularity), represented as an NPSEM (with dependent errors), that violates (73).

$$\begin{array}{lll} Z = \varepsilon_Z; & M(z) = (1-z)\cdot \mathrm{I}(\varepsilon_M > 1/2); & Y(z,m) = z \cdot \mathrm{I}\left(\varepsilon_Y > 1/2\right); \\ \varepsilon_Z \perp\!\!\!\perp \{\varepsilon_M, \varepsilon_Y\}; & \varepsilon_Z = \mathrm{Ber}(1/2); & \varepsilon_M = \varepsilon_Y \sim \mathrm{Unif}(0,1). \end{array} \tag{74}$$

Under an NPSEM, the distribution of the errors $P(\varepsilon_Z, \varepsilon_M, \varepsilon_Y)$ is used to induce a distribution on the actual variables $P(Z, M, Y)$.[72] Note that $M(z=1) = 0$, while $Y(z=0,m) = 1$. Thus $Y(z=1,m) \perp\!\!\!\perp M(z=1)$ and $Y(z=0,m) \perp\!\!\!\perp M(z=0)$ both hold under an FFRCISTG model; therefore both independences must hold on the associated SWIG $\mathcal{G}(z,m)$; see Figure 9(iv). We then have:

$$\begin{aligned} P(Y(z=1,m=0)=1, M(z=0)=1 \mid Z) &= P(Y(z=1,m=0)=1, M(z=0)=1) \\ &= P(\varepsilon_Y > 1/2, \varepsilon_M > 1/2) \\ &= P(\varepsilon_Y > 1/2) = 1/2. \end{aligned}$$

However, $P(Y(z=1,m=0)=1 \mid Z) = P(M(z=0)=1 \mid Z) = 1/2$, so (73) does not hold.[73] The independence (73) is of interest because it is used to identify

$$PDE \equiv E[Y(z=1, M(z=0))] - E[Y(z=0, M(z=0))],$$

the pure direct effect (Robins and Greenland, 1992), later re-named the 'natural direct effect' (Pearl, 2001b). Robins and Richardson (2011) give bounds on the PDE under the independence relations implied by the template $\mathcal{G}(z,m)$.

In contrast, the independence (73) would hold under the NPSEM-IE associated with the DAG in Figure 9(i).[74] This is because the NPSEM-IE model wrongly assumes the data were generated by a distribution with ε_M and ε_Y independent. However even though the errors are actually correlated,

[72] Thus here $M = M(Z)$; $Y = Y(Z, M(Z))$.

[73] The observant reader will note that in NPSEM in (74) M has no effect on Y. This is purely to reduce the number of parameters used to specify the NPSEM and is not required for (73) to fail. Similar comments apply to the determinism in this example. See Figure 23(d) for a general (graphical) generating process that violates (73).

[74] This may be verified by examining Figure 23(e) or via the 'twin-network' method of Balke and Pearl (1994).

the mistaken hypothesis that they are independent cannot be empirically refuted, even were additional data available from the following ideal randomized experiments – one in which Z alone was randomized thereby identifying $P(M(z)=m, Y(z)=y)$ and one in which both Z and M were randomized allowing identification of $P(Y(z,m)=y)$. Since the observed generating process was an FFRCISTG model it follows from the modularity property of $\mathcal{G}(z)$ in Figure 9(ii) that we would find $P(Y=y, M=m \mid Z=z)$ agrees with the distribution $P(M(z)=m, Y(z)=y)$ from the first experiment. Similarly, by modularity for $\mathcal{G}(z,m)$, we would find that $\mathrm{P}(Y=y \mid M=m, Z=z)$ agrees with $P(Y(z,m)=y)$ from the second trial. If such was not the case we could reject the FFRCISTG model (of course still assuming ideal trials).

Presumably Pearl, in insisting on the NPSEM-IE model, would view the DAG in Figure 9(i) as mis-specified since the dependence between the errors ε_M and ε_Y is not represented on the graph. In contrast, since the FFRCISTG model allows for such dependence, there is no need for the graph to represent the dependence, especially when, as discussed above, there is no experimental test (involving the variables Z, M and Y) that could distinguish between the two models. It is of little practical relevance to insist that all NPSEM-IE models should be correctly specified (i.e. with independent errors), if there is no empirical test (even in principle) that can determine whether this is the case.[75]

Nonetheless if there were agreement between experimental distributions and observed conditional distributions, one might often conclude, based on Bayesian inferences with a prior over structures, or simplicity arguments (aka the faithfulness assumption; Spirtes et al. (1993)) that $\varepsilon_M \perp\!\!\!\perp \varepsilon_Y$, so that an NPSEM-IE held.

Our point is simply that such inferences clearly rely on additional external information or principles, which must be argued for in each case. Of particular interest, Robins and Richardson (2011, §4, Fig. 6.3) discuss an expanded graph ($\mathcal{G}^*$) containing more variables than the original graph ($\mathcal{G}$) in Figure 9(i) under which (73) could be tested in a randomized experiment whose treatments correspond to the variables on the expanded graph $\mathcal{G}^*$ [76] that are absent from the original graph $\mathcal{G}$, thereby preserving the link between causal model and experiments introduced by Neyman (1923).

[75] In the absence of any method, even in principle, of testing the underlying assumptions, Pearl's contention that inferences based on an NPSEM-IE are safe provided that the model is correctly specified Pearl (2010) is reminiscent of the reassurance that 'Though it has no seat-belts, this car is safe to drive provided that one avoids being involved in an accident.'

[76] Though we do not do so here, the SWIG framework may be extended to include interventions on expanded graphs of this type.

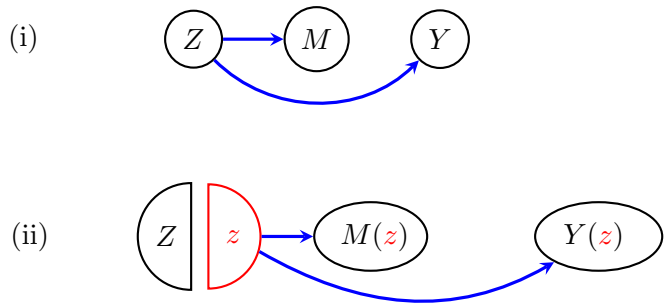


Figure 24: (i) A DAG $\mathcal{G}$: (ii) the SWIT $\mathcal{G}(z)$.

6.1.2 A common cause of two unrelated effects

Next consider the DAG shown Figure 24(i), in which M has no direct effect on Y. Suppose that we wish to test the independence:

$$Y(z{=}1) \perp\!\!\!\perp Z \mid M. \tag{75}$$

An NPSEM with dependent errors that obeys factorization and modularity, but violates (75) is shown here:

$$\begin{aligned}
Z &= \varepsilon_Z;\\
M(z) &= (1-z)\cdot \mathrm{I}(\varepsilon_{M0} < 3/4) + z \cdot \mathrm{I}(\varepsilon_{M1} < 1/4);\\
Y(z) &= (1-z)\cdot \mathrm{I}(\varepsilon_{Y0} < 1/4) + z \cdot \mathrm{I}(\varepsilon_{Y1} < 3/4);\\
\varepsilon_Z &\perp\!\!\!\perp \{\varepsilon_M \equiv (\varepsilon_{M0}, \varepsilon_{M1}),\ \varepsilon_Y \equiv (\varepsilon_{Y0}, \varepsilon_{Y1})\};\\
\varepsilon_{M0} &\perp\!\!\!\perp \varepsilon_{M1}, \qquad \varepsilon_{Mi} = \varepsilon_{Y(1-i)} \quad \text{and} \quad \varepsilon_{Mi} \sim \mathrm{Unif}(0,1) \text{ for } i = 0, 1.
\end{aligned}$$

Notice that here the error terms ε_M and ε_Y are bivariate. The error structure here is such that although $Y(z) \perp\!\!\!\perp M(z)$, we have dependence between $Y(z)$ and $M(1-z)$.

The independence (75) is of interest as it is required in order to identify the intervention distribution $P(Y(\tilde{z}) \,|\, M = m)$ consequently this is not identified under our counterfactual model. However, $P(Y(\tilde{z}) \,|\, M(\tilde{z}) = m)$ is identified, since $Y(\tilde{z}) \perp\!\!\!\perp Z \mid M(\tilde{z})$ is implied by $\mathcal{G}(\tilde{z})$. Once again, the independence (75) *is* implied by the NPSEM-IE associated with the DAG shown Figure 24(a), which may be verified via the 'twin-network' method of Balke and Pearl (1994).

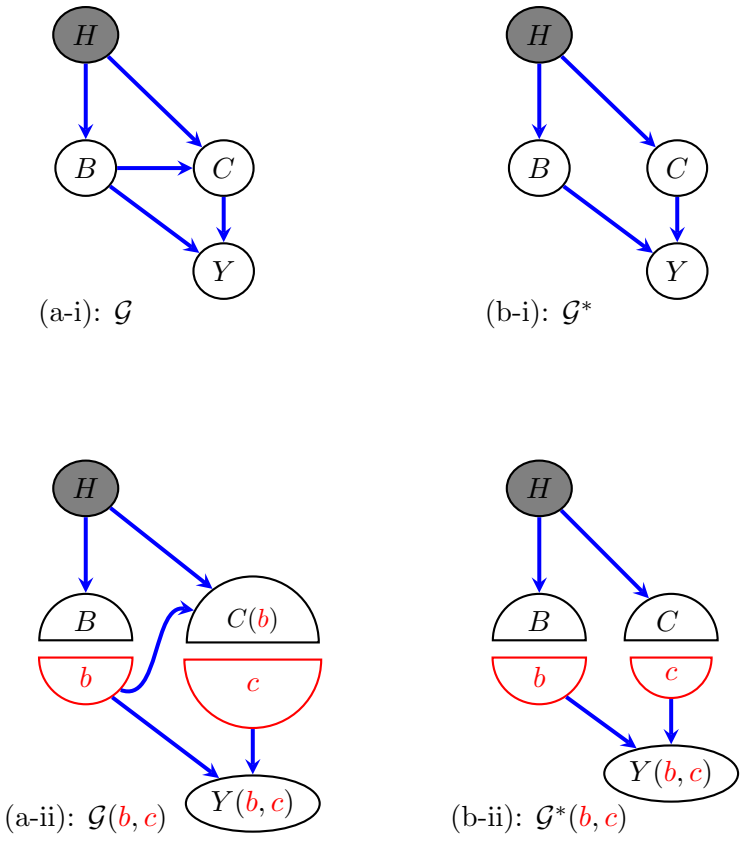


Figure 25: (a-i), (b-i) Two different DAGs $\mathcal{G}$, $\mathcal{G}^*$; H is unobserved, but known to have no (individual-level) effect on Y; in $\mathcal{G}^*$ in addition, B has no (individual level) effect on C; (a-ii), (b-ii) the corresponding templates $\mathcal{G}(b, c)$ and $\mathcal{G}^*(b, c)$. Under the associated FFRCISTG models the 'effect of treatment on the double treated' $E[Y(b_1, c_1) - Y(b_0, c_0)|B = 1, C = 1]$ is identified under (b) but not (a). In contrast, the this quantity is identified under both of the associated NPSEM-IE models.

6.1.3 Effect of treatment on the double treated

For our last example we consider the DAGs $\mathcal{G}$, $\mathcal{G}^*$, shown in Figure 25(a-i) and (b-i). First note that the controlled direct effect (CDE):

$$E[Y(b', c') - Y(b^\dagger, c^\dagger)]$$

is identified under both of these models.

However, now consider the 'effect of treatment on the double treated':

$$E[Y(b{=}1, c{=}1) - Y(b{=}0, c{=}0) \mid B{=}1, C{=}1]. \tag{76}$$

This quantity is identified if

$$Y(b{=}0, c{=}0) \perp\!\!\!\perp C, B. \tag{77}$$

This may be seen as follows:

$$\begin{aligned}
&E[Y(b{=}1, c{=}1) - Y(b{=}0, c{=}0) \mid B{=}1, C{=}1] \\
&\quad = E[Y(b{=}1, c{=}1) \mid B{=}1, C{=}1] - E[Y(b{=}0, c{=}0) \mid B{=}1, C{=}1] \\
&\quad = E[Y(b{=}1, c{=}1) \mid B{=}1, C{=}1] - E[Y(b{=}0, c{=}0) \mid B{=}0, C{=}0] \\
&\quad = E[Y \mid B{=}1, C{=}1] - E[Y \mid B{=}0, C{=}0].
\end{aligned}$$

Here the second equality follows from the independence (77), and the third is by consistency. Inspection of the template $\mathcal{G}^*(b, c)$ shown in Figure 25(b-ii) reveals that this graph does imply (77). Consequently (76) is identified.

However, when the $B \to C$ edge is present, as in $\mathcal{G}$, we are not able to infer (77) from $\mathcal{G}(b^\dagger, c^\dagger)$ as the variable $C(b^\dagger)$ rather than C appears in the SWIG. In Section 7.1.4 we give an example of an NPSEM model that is in the FFRCISTG model associated with $\mathcal{G}$, yet for which (77) does not hold.[77] Consequently (76) is not identified under this model. However, similarly to the analysis in §6.1.2, the independence:

$$Y(b{=}0, c{=}0) \perp\!\!\!\perp C, B(b{=}0). \tag{78}$$

does hold under the FFRCISTG associated with $\mathcal{G}$.

In contrast, the independence (77) does hold under the NPSEM-IE model associated with $\mathcal{G}$, and consequently (76) is identified under this model. This may be established via the counterfactual graph approach of Shpitser and Pearl (2007, 2008).

[77] In fact, the NPSEM given in Section 7.1.4 is also in the FFRCISTG corresponding to the submodel in which the edges $B \to H \leftarrow C$ are not present.

Naively one might imagine that if the FFRCISTG model associated with $\mathcal{G}$ held then this would also imply that the FFRCISTG model $X \to Y$ held where $X = (B, C)$ is a new variable obtained by combining B and C so that X has state space $\mathfrak{X} = \mathfrak{B} \times \mathfrak{C}$ (where $\mathfrak{B}$ and $\mathfrak{C}$ are the state spaces for B and C). However, this implication is false because if the FFRCISTG $X \to Y$ held then it would follow directly that $X \perp\!\!\!\perp Y(x)$ which is equivalent to the independence (77); recall the template Figure 3. Furthermore this implication would also fail even if the edges $B \leftarrow H \to C$ were absent from the DAG $\mathcal{G}$.

6.1.4 Implications for feasibility and identification of dynamic regimes

As discussed in (Robins and Richardson, 2011) the limited counterfactual independence conditions that define an FFRCISTG model were chosen in an attempt to ensure that any intervention distribution identified under the model would be, at least in principle, experimentally testable via a randomized experiment. However, Robins (1986) did not consider intervention regimes in which the treatment $A_1^+(g_1(A))$ was a function of the natural value A_1 of A. In Section 5 we considered the regime with $g(A) = \max(A, 20)$ where A was the natural level of exercise and described how the regime could implemented in an experiment consistent with Robins' intentions.

Now however consider the exercise regime:

$$A_1^+(g') = g'(A_1) \quad \text{where} \quad g'(x) \equiv \max\{0, x - 10\}.$$

under which if in the absence of an intervention one would have exercised for x minutes, then under the intervention one would be required to reduce their exercise time by 10 minutes.[78] Clearly this regime, in contrast to g, could not be experimentally implemented as the act of observing the amount of exercise the patient *would have carried out in the absence of the regime*, necessarily precludes them from exercising for the exact number of minutes that the regime prescribes ![79]

However under the FFRCISTG model for the DAG in $\overline{\mathcal{G}}$ in Figure 20(a) Theorem 31 implies that the distribution of $P(Y(g') = y)$ is identified by

[78] Note that this regime is not the same as the regime: '*Exercise for* 10 *minutes less than you planned to.*' because "The best laid plans of mice and men often go astray". For example, in the absence of a regime, someone might, due to unexpected pain, stop twenty minutes sooner than they had planned to.

[79] Here assuming the non-existence of time travel and that $A_1 > 10$

the dynamic extended g-formula,

$$P(Y(g')=y) = \sum_{a,a^+} P(Y = y \mid A_1 = a^+) I\big(a^+ = \max\{0, a-10\}\big)\, P(A_1 = a)$$

which seemingly contradicts Robins' intention mentioned above that identification results under an FFRCISTG be testable (at least in principle). However this is a contradiction only if, in the context under discussion, the assumption that the DAG in Figure 20(a) represents an FFRCISTG is warranted.

However, this is not the case due to the phenomenon that we observed in §6.1.3. There, we considered a counterfactual distribution

$$P(\{B, C(b^\dagger), Y(b^\dagger, c^\dagger) \text{ such that } b^\dagger \in \mathfrak{B}, c^\dagger \in \mathfrak{C}\})$$

in the FFRCISTG model corresponding to the DAG in Figure 28(a), that includes an edge $B \to C$, so that the first treatment (B) influenced the second (Y).[80] We saw that it does *not* follow that the counterfactual distribution:

$$P(\{X, Y(x^\dagger) \text{ such that } X = (B, C); x^\dagger = (b^\dagger, c^\dagger); b^\dagger \in \mathfrak{B}, c^\dagger \in \mathfrak{C}\})$$

obtained by combining B and C into a single variable, will be in the FFRCISTG model associated with the graph $X \to Y$.

The implication for the infeasible regime g' described above is thus as follows: Let $B_1, \ldots, B_t, \ldots, B_T$ be a sequence of temporally ordered treatments, where B_t indicates that the patient is still exercising at time t, and T is large. Let $\mathcal{G}_T$ be the complete DAG given by the ordering $\langle B_1, \ldots, B_t, \ldots B_T, Y\rangle$. Interventions on the variables on graph $\mathcal{G}_T$ would all be implementable and would include a dynamic regime equivalent to the feasible regime g described above. Since the DAG $\mathcal{G}_T$ represents the actual data generating process it is reasonable to assume that the FFRCISTG assumption holds.

Recall that the variable A_1, encoding the total time exercised, is obtained by combining the treatments $\{B_i\}$. The obvious extension of the discussion in Section 6.1.3 then implies that the fact that the FFRCISTG model associated with $\mathcal{G}_T$ holds, does not imply that the FFRCISTG associated with the graph $A_1 \to Y$ holds. Consequently there is no contradiction.

[80] In Section 6.1.3 we also allowed for an unobserved confounder H of B and C, but as we noted our conclusions are not changed if the edges $B \leftarrow H \to C$ are absent.

7 SWIGs using population-level exclusions

The FFRCISTG approach presumes an underlying NPSEM. In the previous section we built our SWIG under the assumption that we knew the structure of this NPSEM and were given a graph $\mathcal{G}$ in which the absence of an edge $A \to Y$ indicated that A does not appear in the structural equation for Y. In other words, we presuppose that in such a case we have prior knowledge that there is no individual-level effect of A on Y (relative to the other variables in the graph), so that $Y(\tilde{a}, \widetilde{\mathbf{pa}}_Y) = Y(a^*, \widetilde{\mathbf{pa}}_Y)$, for all $\tilde{a}$, a^* and assignments $\widetilde{\mathbf{pa}}_Y$ of values to the parents of Y in $\mathcal{G}$.

Suppose on the other hand we did not know *a priori* whether or not A had a direct effect on Y and therefore we collected observational data to test whether the effect was present. Assume we had unlimited data so sampling variability can be ignored. Suppose positivity holds and thus under our FFRCISTG assumption, $P(Y(\tilde{a}, \widetilde{\mathbf{pa}}_Y) = y)$ is identified by $P(Y = y \mid A = a, PA_Y = \widetilde{\mathbf{pa}}_Y)$ and we found that $P(Y = y \mid A = a, \mathbf{PA}_Y = \widetilde{\mathbf{pa}}_Y)$ did not depend on a for any $\widetilde{\mathbf{pa}}_Y$. We therefore concluded that $P(Y(\tilde{a}, \widetilde{\mathbf{pa}}_Y) = y)$ did not depend on a and thus that A had no population direct effect on Y (relative to the variables in $\mathbf{PA}_Y$). Further assume that were we to perform an idealized randomized experiment on an exchangeable population to check the no confounding assumption of our FFRCISTG we would again find no direct effect. The question then arises whether one should conclude that A had no direct effect for every subject and therefore that A does not occur in the structural equation for Y; alternatively one may choose to remain agnostic as to whether A is present in the structural equation as it is logically possible, and empirically untestable, that harmful and helpful individual level direct effects of A on Y exist but balance out over the population.

Suppose to remain (epistemologically) consistent with the spirit of the FFRCISTG model we choose to make only those assumptions that could be consistently tested were data from an ideal randomized experiment available. In that case, we would not assume that A is absent from the structural equation for Y, as there is no consistent test for the absence of a variable from a structural equation. This is because, as discussed above, it is logically possible that for all values of $\widetilde{\mathbf{pa}}_Y$,

$$P(Y(\tilde{a}, \widetilde{\mathbf{pa}}_Y) = y) = P(Y(a^\dagger, \widetilde{\mathbf{pa}}_Y) = y),$$

and yet $Y(\tilde{a}, \widetilde{\mathbf{pa}}_Y) \neq Y(a^\dagger, \widetilde{\mathbf{pa}}_Y)$ for some units. Thus an approach that requires that *all* assumptions in the model be testable (at least in principle) via an ideal randomized experiment would make exclusion restrictions at the population rather than individual level.

In this section we will describe how to construct a SWIG that allows for some missing arrows to represent no effect at the individual level and others to represent only no causal effects at the population level; thereby allowing one to have different opinions for different variables about absence of direct effects at the individual level. This is a novel innovation supplied by SWIGs. We will see that this may be achieved by first constructing a SWIG from an FFRCISTG, as described in §3.3 and then removing edges from the SWIG when this is permitted by (the Markov structure of) the distribution over the actual variables.

7.1 The SWIG $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$

Our starting point will be to assume an (individual level) FFRCISTG model associated with a DAG $\overline{\mathcal{G}}$ as in Section 3. Again to stay within the sprit of the FFRCISTG model we would usually take $\overline{\mathcal{G}}$ to simply be a complete DAG given by the time-order. However the general theory that we now develop does not require $\overline{\mathcal{G}}$ to be complete. In the following we will suppose that the distribution of the actual variables $P(\mathbf{V})$ factorizes according to a DAG $\mathcal{G}$ that is a subgraph of $\overline{\mathcal{G}}$.

Definition 35. *The* template $\overline{\mathcal{G}}_{\mathcal{G}}(\mathbf{a})$ induced from $\overline{\mathcal{G}}(\mathbf{a})$ by the subgraph $\mathcal{G}$ *is the graph obtained from* $\overline{\mathcal{G}}(\mathbf{a})$ *by removing any edge that is not present in* $\mathcal{G}$ *under the natural correspondence* (19).

Note that the labeling of the counterfactual variables in $\overline{\mathcal{G}}(\mathbf{a})$ is inherited by the variables in $\overline{\mathcal{G}}_{\mathcal{G}}(\mathbf{a})$. Thus if $V(\mathbf{a}_V)$ is a vertex in $\overline{\mathcal{G}}_{\mathcal{G}}(\mathbf{a})$ then $\mathbf{a}_V$ is the set of fixed nodes that are ancestors of V in $\overline{\mathcal{G}}(\mathbf{a})$, which will, in general, be a superset of the set of fixed vertices that are ancestors of $V(\mathbf{a}_V)$ in $\overline{\mathcal{G}}_{\mathcal{G}}(\mathbf{a})$.

7.1.1 Factorization and Modularity

Under positivity, the SWIG $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$ obtained by instantiating $\overline{\mathcal{G}}_{\mathcal{G}}(\mathbf{a})$ captures the structure of the distribution of the counterfactual variables $\mathbb{V}(\tilde{\mathbf{a}})$ under the additional assumption that $P(\mathbf{V})$ factorizes according to $\mathcal{G}$:

Proposition 36. *Under the* FFRCISTG *model associated with the DAG* $\overline{\mathcal{G}}$ *if* $P(\mathbf{V})$ *is positive and factors according to a subgraph* $\mathcal{G}$ *of* $\overline{\mathcal{G}}$ *, then* $P(\mathbb{V}(\tilde{\mathbf{a}}))$ *factorizes with respect to the SWIG* $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$.

Proof: This is a special case of Theorem 40 below. □

In addition, the analogous modularity property holds:

Proposition 37. *Under the* FFRCISTG *model associated with the DAG* $\overline{\mathcal{G}}$ *if* $P(\mathbf{V})$ *is positive and factors according to a subgraph* $\mathcal{G}$ *then* $(P(\mathbb{V}(\tilde{\mathbf{a}})), \overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}}))$ *obeys modularity with respect to* $(P(\mathbf{V}), \mathcal{G})$.

Proof: This is a special case of Theorem 42 below. □

It follows from Propositions 36 and 37 that under positivity conditions the g-formula identification results in Section 4 also apply to $(P(\mathbb{V}(\tilde{\mathbf{a}})), \overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}}))$ and $(P(\mathbf{V}), \mathcal{G})$. This is because Theorem 22 merely required that the properties of factorization and modularity hold, and did not require that the counterfactual distribution be in the FFRCISTG associated with $\mathcal{G}$.[81] In fact, since we do not make all of the individual level exclusion restrictions associated with $\mathcal{G}$, although the counterfactual distributions $P(\mathbb{V}(\tilde{\mathbf{a}}))$ we consider will be in the FFRCISTG model associated with $\overline{\mathcal{G}}$ they may *not* be in the FFRCISTG associated with $\mathcal{G}$. Also note that when applying Theorem 22 we are free to use any topological ordering of of $\mathcal{G}$ even though this may not be a topological ordering of $\overline{\mathcal{G}}$.

The population exclusion property is formalized in the following:

Proposition 38. *Under the* FFRCISTG *model associated with the DAG* $\overline{\mathcal{G}}$ *if* $P(\mathbf{V})$ *is positive and factors according to a subgraph* $\mathcal{G}$*, then for any set* $\mathbf{B}$ *such that* $\text{pa}_{\mathcal{G}}(Y) \subset \mathbf{B}$*, for all assignments* $\mathbf{b}^{\dagger}$ *,* $\mathbf{b}^{\ddagger}$*,*

$$P(Y(\mathbf{b}^{\dagger}) = y) = P(Y(\mathbf{b}^{\ddagger}) = y), \quad \textit{whenever} \quad \mathbf{b}^{\dagger}{}_{\text{pa}_{\mathcal{G}}(Y)} = \mathbf{b}^{\ddagger}{}_{\text{pa}_{\mathcal{G}}(Y)}.$$

Proof: This is an immediate consequence of Proposition 36; see (94) below. □

7.1.2 A two variable example

Consider a distribution in FFRCISTG model associated with the DAG $\overline{\mathcal{G}}$ in Figure 26(a). The SWIG $\overline{\mathcal{G}}(b)$ is shown in Figure 26(c). By Proposition 11 factorization is obeyed so that for all $b, y, b^{\dagger}$:

$$P(B{=}b, Y(b^{\dagger}){=}y) = P(B{=}b)P(Y(b^{\dagger}){=}y). \tag{79}$$

Further, by Proposition 16 modularity holds so we have for all $b^{\dagger}, y$:

$$P(Y(b^{\dagger}){=}y) = P(Y{=}y \mid B{=}b^{\dagger}) \tag{80}$$

[81] Propositions 11 and 16 establish that for any counterfactual distribution in the FFRCISTG model associated with $\mathcal{G}$ factorization and modularity hold with respect to $\mathcal{G}$, but as shown in this section, membership in the FFRCISTG model for $\mathcal{G}$ is not a necessary condition for factorization and modularity with respect to $\mathcal{G}$ to hold.

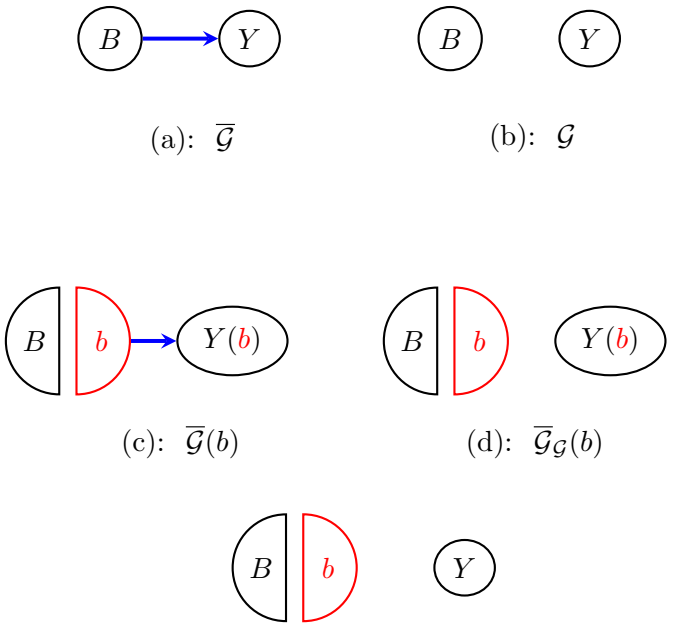


Figure 26: Representing population level exclusion restrictions: (a) The underlying NPSEM $\overline{\mathcal{G}}$; (b) a DAG $\mathcal{G}$ with respect to which we assume $P(B, Y)$ factors; (c) the template $\overline{\mathcal{G}}(b)$; (d) the template $\overline{\mathcal{G}}_{\mathcal{G}}(b)$ induced from $\overline{\mathcal{G}}(b)$ by $\mathcal{G}$. (e) For comparison, the SWIG $\mathcal{G}(b)$ obtained from $\mathcal{G}$ if the individual level exclusions given by the FFRCISTG associated with $\mathcal{G}$ held.

for all values for which both sides of the equality are well-defined.

Now consider the trivial subgraph $\mathcal{G}$ in Figure 26(b).

The induced SWIG $\overline{\mathcal{G}}_{\mathcal{G}}(b^\dagger)$ is shown in Figure 26(d). Notice that the equation (79) establishes that $P(B=b, Y(b^\dagger)=y)$ *also* factorizes according to $\overline{\mathcal{G}}_{\mathcal{G}}(b^\dagger)$. This is because the edge that is 'lost' in $\overline{\mathcal{G}}_{\mathcal{G}}(b^\dagger)$ takes the form $b \to Y(b^\dagger)$, originating from a variable that is fixed in $\overline{\mathcal{G}}_{\mathcal{G}}(b^\dagger)$. Thus $P(B=b, Y(b^\dagger)=y)$ factorizes according to $\overline{\mathcal{G}}_{\mathcal{G}}(b^\dagger)$ regardless of whether $P(B,Y)$ factorizes according to $\mathcal{G}$; see Corollary 41 below.

Now suppose further that the observed distribution $P(B,Y)$ factorizes according to $\mathcal{G}$, so that for all b, y:

$$P(B=b, Y=y) = P(B=b)P(Y=y).$$

Under positivity, it follows that for all $b^\dagger$:

$$P(Y=y) = P(Y=y \mid B=b^\dagger). \tag{81}$$

Together, (80) and (81) establish that

$$P(Y(b^\dagger)=y) = P(Y=y), \tag{82}$$

holds for all $b^\dagger, y$, and thus $(P(B, Y(b^\dagger)), \overline{\mathcal{G}}_{\mathcal{G}}(b^\dagger))$ obeys modularity with respect to $(P(B,Y), \mathcal{G})$. Note that as an immediate implication of (82) it holds that $P(Y(b^\dagger)) = P(Y(b^\ddagger))$, so that a population exclusion restriction obtains even though under the NPSEM associated with $\overline{\mathcal{G}}$, $Y(b^\dagger) = Y(b^\ddagger)$ may *not* hold.

We contrast this interpretation with that of the SWIG $\mathcal{G}(b^\dagger)$ shown in Figure 26(e), that would be obtained under the stronger assumption of an NPSEM obeying the FFRCISTG associated with $\mathcal{G}$ (rather than $\overline{\mathcal{G}}$). This SWIG does imply $Y(b^\dagger) = Y(b^\ddagger)$, as may be inferred by recalling Section §3.4: since in the SWIG $\mathcal{G}(b^\dagger)$, Y is not labeled with $b^\dagger$, the node represents the equivalence class of counterfactual variables $\{Y(b^*)\}$, where b^* is any value of B.

The following is an example of an NPSEM that satisfies the FFRCISTG independence assumption associated with $\overline{\mathcal{G}}$, though not the FFRCISTG assumption for $\mathcal{G}$, yet for which $P(B,Y)$ still factorizes with respect to $\mathcal{G}$:

$$B = \varepsilon_B; \qquad\qquad Y(b) = I(\varepsilon_Y = b);$$

where ε_B, ε_Y are independent Bernoulli(0.5) random variables. It follows that $P(Y(b=0) = y) = P(Y(b=1) = y) = 0.5$, however, we see that $Y(b=1) = 1 - Y(b=0)$; (see also Pearl, 2000, pp.35-36). Consequently, the FFRCISTG model for $\mathcal{G}$ does not hold since it implies no individual-level effect of B on Y, or equivalently $Y(b^\dagger) = Y(b^\ddagger)$.

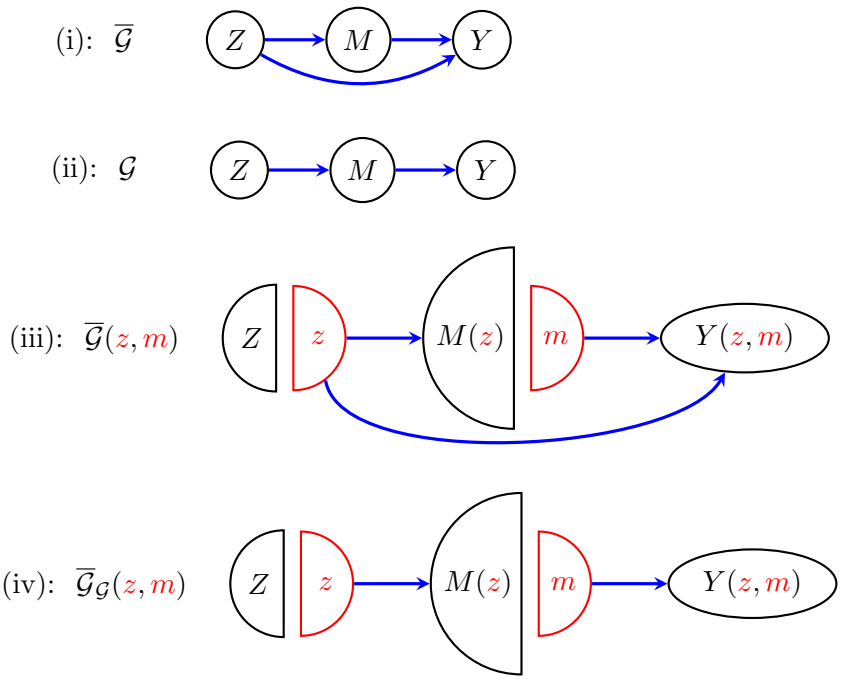


Figure 27: (i) The underlying NPSEM $\overline{\mathcal{G}}$; (ii) the DAG $\mathcal{G}$ according to which $P(Z, M, Y)$ factors; (iii) the template $\overline{\mathcal{G}}(z, m)$; (iv) the template $\overline{\mathcal{G}}_{\mathcal{G}}(z, m)$ induced from $\overline{\mathcal{G}}(z, m)$ by $\mathcal{G}$; $\overline{\mathcal{G}}_{\mathcal{G}}(z, m)$ in contrast to $\mathcal{G}(z, m)$ shown in Figure 9(iv) does not imply $Y(z, m) = Y(z', m)$.

7.1.3 Example: absence of a population direct effect

Consider a distribution in FFRCISTG model associated with the DAG $\overline{\mathcal{G}}$ as shown in Figure 27(i). Since factorization and modularity hold with respect to $\overline{\mathcal{G}}(\tilde{z}, \tilde{m})$, shown in Figure 27(iii) we have for all $\tilde{z}, \tilde{m}$, z, m, y:

$$P(Z=z, M(\tilde{z})=m, Y(\tilde{z},\tilde{m})=y) = P(Z=z)P(M(\tilde{z})=m)P(Y(\tilde{z},\tilde{m})=y), \tag{83}$$

and

$$P(M(\tilde{z})=m) = P(M=m \mid Z=\tilde{z}), \quad P(Y(\tilde{z},\tilde{m})=y) = P(Y=y \mid M=\tilde{m}, Z=\tilde{z}).$$

Now consider the subgraph $\mathcal{G}$ of $\overline{\mathcal{G}}$ that does not include the $Z \to Y$, and the resulting induced SWIG $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{z}, \tilde{m})$, shown in in Figure 27(ii) and (iv) respectively. Observe that (83) immediately implies that $P(Z=z, M(\tilde{z})=m, Y(\tilde{z},\tilde{m})=y)$ factors according to $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{z}, \tilde{m})$. Again this is because the only edge that is in $\overline{\mathcal{G}}(\tilde{z}, \tilde{m})$, but not $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{z}, \tilde{m})$ is $z \to Y(\tilde{z}, \tilde{m})$.

Now suppose in addition that the joint distribution $P(Z, M, Y)$ factors according to $\mathcal{G}$:

$$P(Z=z, M=m, Y=y) = P(Z=z)P(M=m \mid Z=z)P(Y=y \mid M=m). \tag{84}$$

Under positivity we then have that for all $\tilde{z}, \tilde{m}, y$:

$$P(Y(\tilde{z},\tilde{m})=y) = P(Y=y \mid M=\tilde{m}, Z=\tilde{z}) = P(Y=y \mid M=\tilde{m}). \tag{85}$$

Thus $(P(Z, M(\tilde{z}), Y(\tilde{z}, \tilde{m})), \overline{\mathcal{G}}_{\mathcal{G}}(\tilde{z}, \tilde{m}))$ obeys modularity with respect to $(P(Z, M, Y), \mathcal{G})$. Consequently it holds that

$$P(Y(\tilde{z},\tilde{m})=y) = P(Y=y \mid M=\tilde{m}) = P(Y(z^{\dagger},\tilde{m})=y),$$

for all $\tilde{z}$, $z^{\dagger}$ so that the average controlled direct effect of Z on Y, when M is fixed, vanishes.

However, $\overline{\mathcal{G}}_{\mathcal{G}}(z, m)$ does not imply $Y(\tilde{z}, \tilde{m}) = Y(z^{\dagger}, \tilde{m})$. The latter individual level exclusion restriction is implied by the SWIG $\mathcal{G}(z, m)$ shown in Figure 9(iv), that would result from the FFRCISTG model associated with $\mathcal{G}$.

7.1.4 Example: the effect of treatment on the double treated

In the previous subsection we saw that for the controlled direct effect the interpretation of the missing edge did not alter the identification results. We

now present an example in which the distinction leads to different results. First consider the graphs $\overline{\mathcal{G}}$ and $\mathcal{G}$ shown in Figure 28(a) and (b). It follows from the discussion in §6.1.3 that the effect of treatment on the double treated:

$$E[Y(b', c') - Y(b^\dagger, c^\dagger) \mid B = b', C = c']. \tag{86}$$

is identified under the FFRCISTG model associated with $\mathcal{G}$, but not under the FFRCISTG associated with $\overline{\mathcal{G}}$. Thus the effect of treatment on the double treated is identified in the absence of an individual level direct effect of B on C.[82]

We now consider a distribution in the FFRCISTG model associated with $\overline{\mathcal{G}}$ under which, in addition, the observed distribution $P(B, C, Y)$ factors according to $\mathcal{G}$, so that there is no population-level effect of B on C. The template for the SWIG $\overline{\mathcal{G}}_{\mathcal{G}}(b^\dagger, c^\dagger)$ induced from $\overline{\mathcal{G}}(b^\dagger, c^\dagger)$ by $\mathcal{G}$ is shown in Figure 28(d). We see that $\overline{\mathcal{G}}_{\mathcal{G}}(b^\dagger, c^\dagger)$ implies

$$Y(b^\dagger, c^\dagger) \perp\!\!\!\perp B, C(b^\dagger), \tag{87}$$

as does the SWIG $\overline{\mathcal{G}}(b^\dagger, c^\dagger)$. However neither $\overline{\mathcal{G}}(b^\dagger, c^\dagger)$ nor $\overline{\mathcal{G}}_{\mathcal{G}}(b^\dagger, c^\dagger)$ imply:

$$Y(b^\dagger, c^\dagger) \perp\!\!\!\perp B, C. \tag{88}$$

Consequently, we would expect that there exists a counterfactual distribution in the FFRCISTG model associated with $\overline{\mathcal{G}}$, for which $P(B, C, Y)$ factorizes according $\mathcal{G}$, and yet (88) does not hold. We prove this below by constructing such an example; see NPSEM II below.

We also construct a counterfactual distribution in the NPSEM-IE (and hence also the FFRCISTG) associated with $\mathcal{G}$; see NPSEM I below. Both NPSEM I and NPSEM II give rise to the same observed distribution for $P(B, C, Y)$, and both obey modularity so that $P(C(b^\dagger), Y(b^\dagger)) = P(C, Y|B = b)$ and $P(Y(b^\dagger, c^\dagger)) = P(Y|B = b^\dagger, C = c^\dagger)$. Thus there is no way to distinguish between these models on the basis of observational or experimental data.

NPSEM I:

$$\begin{aligned} B &= \varepsilon_B; & C(b) &= \varepsilon_C; \\ Y(b, c) &= I(0.5 - 0.2bc > \varepsilon_Y); \end{aligned}$$

[82] Although there is no consistent test for the absence of an individual level direct effect of B on C, we can reject that null hypothesis by performing a randomized experiment in which B is randomized, and we find an association between B and C conditional on some pre-treatment covariate.

where $\varepsilon_B \sim \text{Ber}(0.5)$; $\varepsilon_C \sim \text{Ber}(0.5)$; $\varepsilon_Y \sim \text{Unif}[0,1]$ and $\varepsilon_B \perp\!\!\!\perp \varepsilon_C \perp\!\!\!\perp \varepsilon_Y$.

Under this model there is no individual level direct effect of B on C, so $B \perp\!\!\!\perp C$ and we obtain:

$$\begin{aligned} E[Y(1,1) - Y(0,0) \mid B=1, C=1] &= E[Y \mid B=1, C=1] - E[Y \mid B=0, C=0] \\ &= 0.3 - 0.5 = -0.2. \end{aligned}$$

Model II:

$$\begin{aligned} B &= \varepsilon_B; \\ C(b) &= I(0.5 + (-1)^{\varepsilon_{C1}}(0.4 + 0.05b) > \varepsilon_{C2}); \\ Y(b,c) &= I(0.1 \times 3^{1+(1-bc)(2\epsilon_{Y1}-1)} > \varepsilon_{Y2}); \end{aligned}$$

where $\varepsilon_B \sim \text{Ber}(0.5)$; $\varepsilon_{C1} \sim \text{Ber}(0.5)$; $\varepsilon_{C2} \sim \text{Unif}[0,1]$; $\varepsilon_{Y1} = \varepsilon_{C1}$; $\varepsilon_{Y2} \sim \text{Unif}[0,1]$ and

$$\varepsilon_B \perp\!\!\!\perp \varepsilon_{C1} \perp\!\!\!\perp \varepsilon_{C2} \perp\!\!\!\perp \varepsilon_{Y2}.$$

Under this model, even though B has an individual level effect on C, $P(C=1 \mid B=b) = 0.5$ so $B \perp\!\!\!\perp C$. Further,

$$\begin{aligned} &E[Y(1,1) - Y(0,0) \mid B=1, C=1] \\ &\quad = E[Y \mid B=1, C=1] - E[Y(0,0) \mid B=1, C=1] \\ &\quad = 0.3 - 0.14 = 0.16. \end{aligned}$$

The effect of treatment on the double treated has the opposite sign in Model I versus Model II. These NPSEMs thus show that the effect of treatment on the double treated is not identified under the FFRCISTG model associated with $\overline{\mathcal{G}}$ *even* in the absence of a population level direct effect of B on C.

7.2 Positivity Conditions

In Section 7.1.1 above we showed that under positivity of the observed distribution $P(\mathbf{V})$ factorization and modularity held for the induced SWIG $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$ for every set $\mathbf{A}$ and every assignment $\tilde{\mathbf{a}}$. We now give weaker positivity conditions that are sufficient for factorization and modularity for a given set $\mathbf{A}$ and assignment $\tilde{\mathbf{a}}$.

Condition 39 (Conditions for factorization and modularity for $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$)**.** *Given a counterfactual distribution $P(\mathbb{V}(\tilde{\mathbf{a}}))$ in the* FFRCISTG *model associated with the DAG $\overline{\mathcal{G}}$, where $P(\mathbf{V})$ factors according to a subgraph $\mathcal{G}$, and a given vertex $V \in \mathbf{V}$, we consider the following conditions:*

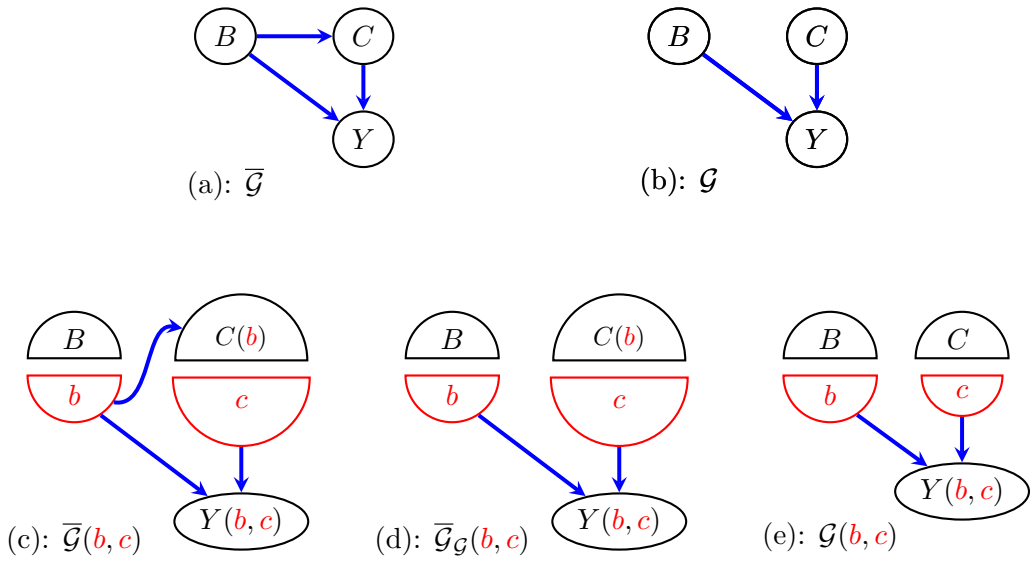


Figure 28: (a) A DAG $\overline{\mathcal{G}}$ representing an FFRCISTG model; (b) the graph $\mathcal{G}$ with respect to which $P(B, C, Y)$ factorizes; (c) the template $\overline{\mathcal{G}}(b, c)$ for the SWIG obtained from $\overline{\mathcal{G}}$; (d) the template $\overline{\mathcal{G}}_{\mathcal{G}}(b, c)$ for the SWIG induced from $\overline{\mathcal{G}}(b, c)$ by $\mathcal{G}$; (e) the template $\mathcal{G}(b, c)$ for the SWIG obtained under the assumption of no individual level effect of B on C.

(I) *If* $\mathrm{pa}_{\overline{\mathcal{G}}}(V) \setminus (\mathrm{pa}_{\mathcal{G}}(V) \cup \mathbf{A}) \neq \emptyset$ *then for all* $\mathbf{q}$*:*

$$P\left((\mathrm{pa}_{\overline{\mathcal{G}}(\tilde{\mathbf{a}})}(V(\tilde{\mathbf{a}})) \setminus \tilde{\mathbf{a}}) = \mathbf{q}\right) > 0$$
$$\Rightarrow P\left((\mathrm{pa}_{\overline{\mathcal{G}}}(V) \setminus \mathbf{A}) = \mathbf{q}, (\mathrm{pa}_{\overline{\mathcal{G}}}(V) \cap \mathbf{A}) = \tilde{\mathbf{a}}_{\mathrm{pa}_{\overline{\mathcal{G}}}(V) \cap \mathbf{A}}\right) > 0. \quad (89)$$

(II) *If* $\mathrm{pa}_{\overline{\mathcal{G}}}(V) \setminus (\mathrm{pa}_{\mathcal{G}}(V) \cup \mathbf{A}) = \emptyset$*, but* $\mathrm{pa}_{\overline{\mathcal{G}}}(V) \setminus \mathrm{pa}_{\mathcal{G}}(V) \neq \emptyset$ *then for all* $\mathbf{q}$*:*

$$P\left((\mathrm{pa}_{\overline{\mathcal{G}}(\tilde{\mathbf{a}})}(V(\tilde{\mathbf{a}})) \setminus \tilde{\mathbf{a}}) = \mathbf{q}\right) > 0 \quad \textit{and}$$
$$P\left((\mathrm{pa}_{\mathcal{G}}(V) \setminus \mathbf{A}) = \mathbf{q}, (\mathrm{pa}_{\mathcal{G}}(V) \cap \mathbf{A}) = \tilde{\mathbf{a}}_{\mathrm{pa}_{\mathcal{G}}(V) \cap \mathbf{A}}\right) > 0$$
$$\Rightarrow P\left((\mathrm{pa}_{\overline{\mathcal{G}}}(V) \setminus \mathbf{A}) = \mathbf{q}, (\mathrm{pa}_{\overline{\mathcal{G}}}(V) \cap \mathbf{A}) = \tilde{\mathbf{a}}_{\mathrm{pa}_{\overline{\mathcal{G}}}(V) \cap \mathbf{A}}\right) > 0. \quad (90)$$

In both of the above conditions we consider any event concerning an empty set of variables to be (vacuously) true and hence to have probability one.

First observe that these conditions are non-trivial only for those variables V for which the set of parents in $\mathcal{G}$ are a strict subset of the set of parents in $\overline{\mathcal{G}}$. Second note that Clause (I) and (II) place restrictions on disjoint subsets of these variables:

- Clause (I) applies to variables V for which the *random* parents of the corresponding counterfactual variable $V(\tilde{\mathbf{a}})$ in $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$ are a strict subset of the *random* parents of $V(\tilde{\mathbf{a}})$ in $\overline{\mathcal{G}}(\tilde{\mathbf{a}})$.
- Clause (II) applies to variables that have the *same random* parents in $\overline{\mathcal{G}}(\tilde{\mathbf{a}})$ and $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$, but the *fixed* parents of $V(\tilde{\mathbf{a}})$ in $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$ are a strict subset of the *fixed* parents of $V(\tilde{\mathbf{a}})$ in $\overline{\mathcal{G}}(\tilde{\mathbf{a}})$.

Clause (I) is sufficient for factorization of $P(\mathbb{V}(\tilde{\mathbf{a}}))$ with respect to $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$, while for modularity of $(P(\mathbb{V}(\tilde{\mathbf{a}})), \overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}}))$ with respect $(P(\mathbf{V}), \mathcal{G})$ both conditions are needed:

Theorem 40. *Given a counterfactual distribution* $P(\mathbb{V}(\tilde{\mathbf{a}}))$ *in the* FFRCISTG *model associated with the DAG* $\overline{\mathcal{G}}$*, if* $P(\mathbf{V})$ *factors according to a subgraph* $\mathcal{G}$ *and* Condition 39(I) *holds for all* V *then* $P(\mathbb{V}(\tilde{\mathbf{a}}))$ *factorizes with respect to* $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$*.*

Proof: By Proposition 11, $P(\mathbb{V}(\tilde{\mathbf{a}}))$ factorizes with respect to $\overline{\mathcal{G}}(\tilde{\mathbf{a}})$. Further by Proposition 16, for every $\mathbf{q}, v$:

$$P(V(\tilde{\mathbf{a}}) = v \mid (\mathrm{pa}_{\overline{\mathcal{G}}(\tilde{\mathbf{a}})}(V) \setminus \tilde{\mathbf{a}}) = \mathbf{q}) \quad (91)$$
$$= P(V = v \mid (\mathrm{pa}_{\overline{\mathcal{G}}}(V) \setminus \mathbf{A}) = \mathbf{q}, (\mathrm{pa}_{\overline{\mathcal{G}}}(V) \cap \mathbf{A}) = \tilde{\mathbf{a}}_{\mathrm{pa}_{\overline{\mathcal{G}}}(V) \cap \mathbf{A}}),$$

whenever both sides are well-defined. Since, by hypothesis, $P(\mathbf{V})$ factorizes according to $\mathcal{G}$, which is a subgraph of $\overline{\mathcal{G}}$, it follows that for every $V \in \mathbf{V}$:

$$\begin{aligned} &P(V = v \mid (\mathrm{pa}_{\overline{\mathcal{G}}}(V) \setminus \mathbf{A}) = \mathbf{q}, (\mathrm{pa}_{\overline{\mathcal{G}}}(V) \cap \mathbf{A}) = \tilde{\mathbf{a}}_{\mathrm{pa}_{\overline{\mathcal{G}}}(V) \cap \mathbf{A}}) \\ &\quad = P(V = v \mid (\mathrm{pa}_{\mathcal{G}}(V) \setminus \mathbf{A}) = \mathbf{q}_{\mathrm{pa}_{\mathcal{G}}(V) \setminus \mathbf{A}}, (\mathrm{pa}_{\mathcal{G}}(V) \cap \mathbf{A}) = \tilde{\mathbf{a}}_{\mathrm{pa}_{\mathcal{G}}(V) \cap \mathbf{A}}) \end{aligned} \tag{92}$$

whenever both sides are well-defined.

If $\mathrm{pa}_{\overline{\mathcal{G}}}(V) \setminus (\mathrm{pa}_{\mathcal{G}}(V) \cup \mathbf{A}) \neq \emptyset$ then Condition 39(I) ensures that whenever the LHS of (91) is well-defined, the RHS is also well-defined. This immediately implies that both sides of (92) are well-defined (since the conditioning event on the right is a subset of that on the left). Hence for all $\mathbf{q}, v$:

$$\begin{aligned} &P(V(\tilde{\mathbf{a}}) = v \mid (\mathrm{pa}_{\overline{\mathcal{G}}(\tilde{\mathbf{a}})}(V) \setminus \tilde{\mathbf{a}}) = \mathbf{q}) \\ &\quad = P(V = v \mid (\mathrm{pa}_{\mathcal{G}}(V) \setminus \mathbf{A}) = \mathbf{q}_{\mathrm{pa}_{\mathcal{G}}(V) \setminus \mathbf{A}}, (\mathrm{pa}_{\mathcal{G}}(V) \cap \mathbf{A}) = \tilde{\mathbf{a}}_{\mathrm{pa}_{\mathcal{G}}(V) \cap \mathbf{A}}) \end{aligned} \tag{93}$$

whenever both sides are well-defined. Thus for all $\mathbf{q}$:

$$\begin{aligned} &P(V(\tilde{\mathbf{a}}) = v \mid (\mathrm{pa}_{\overline{\mathcal{G}}(\tilde{\mathbf{a}})}(V) \setminus \tilde{\mathbf{a}}) = \mathbf{q}) \\ &\quad = P(V(\tilde{\mathbf{a}}) = v \mid (\mathrm{pa}_{\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})}(V) \setminus \tilde{\mathbf{a}}) = \mathbf{q}_{\mathrm{pa}_{\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})}(V)}), \end{aligned} \tag{94}$$

whenever the left hand side is well-defined. Here we have used the fact that by the construction of $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$, $\mathrm{pa}_{\overline{\mathcal{G}}(\tilde{\mathbf{a}})}(V) \setminus \tilde{\mathbf{a}}$ are the set of random vertices in $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$ that correspond to $\mathrm{pa}_{\mathcal{G}}(V) \setminus \mathbf{A}$ in $\mathcal{G}$.

If $\mathrm{pa}_{\overline{\mathcal{G}}}(V) \setminus (\mathrm{pa}_{\mathcal{G}}(V) \cup \mathbf{A}) = \emptyset$ then again by the construction of $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$, $\mathrm{pa}_{\overline{\mathcal{G}}(\tilde{\mathbf{a}})}(V) \setminus \tilde{\mathbf{a}} = \mathrm{pa}_{\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})}(V) \setminus \tilde{\mathbf{a}}$ and there is nothing to prove. This establishes that $P(\mathbb{V}(\tilde{\mathbf{a}}))$ factorizes according to $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$. □

Notice that in the absence of Condition 39(I), it is possible that for some $\mathbf{q}$, the LHS of (91) is well-defined, yet the RHS is not. In this circumstance the fact that $P(\mathbf{V})$ factorizes according to $\mathcal{G}$ places no restriction on the LHS of (91) – since modularity for $\overline{\mathcal{G}}$ only requires equality when both the LHS and RHS of (91) are well-defined. Consequently if Condition 39(I) does not hold for some V, then $P(\mathbb{V}(\tilde{\mathbf{a}}))$ may fail to factorize according to $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$.

Notice that, as noted in the proof of Theorem 40, for vertices V that have the same set of random parents in both $\overline{\mathcal{G}}(\tilde{\mathbf{a}})$ and $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$, there is nothing to prove – we do not need to use the hypothesis that $P(\mathbf{V})$ factorizes according to $\mathcal{G}$, nor Condition 39(I). Thus we have the following result:

Corollary 41. *Given two DAGs $\mathcal{G}$ and $\overline{\mathcal{G}}$, if for every edge $X \to Y$ that is present in $\overline{\mathcal{G}}$, but not in $\mathcal{G}$, $X \in \mathbf{A}$ then every counterfactual distribution $P(\mathbb{V}(\tilde{\mathbf{a}}))$ in the* FFRCISTG *model associated with the DAG $\overline{\mathcal{G}}$ factorizes with respect to $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$.*

We now establish that Condition 39(I)&(II) are sufficient for modularity:

Theorem 42. *Given a counterfactual distribution* $P(\mathbb{V}(\tilde{\mathbf{a}}))$ *in the* FFR-CISTG *model associated with the DAG* $\overline{\mathcal{G}}$*, if* $P(\mathbf{V})$ *factors according to a subgraph* $\mathcal{G}$ *and* Condition 39(I)&(II) *hold for all* V *then* $(P(\mathbb{V}(\tilde{\mathbf{a}})), \overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}}))$ *obeys modularity with respect* $(P(\mathbf{V}), \mathcal{G})$*.*

Proof: Propositions 11 and 16 imply that $(P(\mathbb{V}(\tilde{\mathbf{a}})), \overline{\mathcal{G}}(\tilde{\mathbf{a}}))$ obeys modularity with respect to $(P(\mathbf{V}), \overline{\mathcal{G}})$, so:

$$\begin{aligned} &P\left(V(\tilde{\mathbf{a}}) = v \,\middle|\, \left(\mathrm{pa}_{\overline{\mathcal{G}}(\tilde{\mathbf{a}})}(V(\tilde{\mathbf{a}})) \setminus \tilde{\mathbf{a}}\right) = \mathbf{q}\right) \\ &\qquad = P\Big(V = v \,\Big|\, \left(\mathrm{pa}_{\overline{\mathcal{G}}}(V) \setminus \mathbf{A}\right) = \mathbf{q},\ \ \left(\mathrm{pa}_{\overline{\mathcal{G}}}(V) \cap \mathbf{A}\right) = \tilde{\mathbf{a}}_{\mathrm{pa}_{\overline{\mathcal{G}}}(V) \cap \mathbf{A}}\Big), \end{aligned} \tag{95}$$

whenever both sides are well-defined.

If $\mathrm{pa}_{\overline{\mathcal{G}}}(V) = \mathrm{pa}_{\mathcal{G}}(V)$ then $V(\tilde{\mathbf{a}})$ has the same random and fixed parents in both $\overline{\mathcal{G}}(\tilde{\mathbf{a}})$ and $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$, hence there is nothing to prove. Thus suppose that $\mathrm{pa}_{\overline{\mathcal{G}}}(V) \setminus \mathrm{pa}_{\mathcal{G}}(V) \neq \emptyset$.

Suppose that for some assignment $\mathbf{q}'$ to $\mathrm{pa}_{\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})}(V(\tilde{\mathbf{a}})) \setminus \tilde{\mathbf{a}}$,

$$P\left(V(\tilde{\mathbf{a}}) = v \,\middle|\, \left(\mathrm{pa}_{\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})}(V(\tilde{\mathbf{a}})) \setminus \tilde{\mathbf{a}}\right) = \mathbf{q}'\right)$$

is well-defined, so that

$$P\left(\left(\mathrm{pa}_{\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})}(V(\tilde{\mathbf{a}})) \setminus \tilde{\mathbf{a}}\right) = \mathbf{q}'\right) > 0.$$

Since $(\mathrm{pa}_{\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})}(V(\tilde{\mathbf{a}})) \setminus \tilde{\mathbf{a}}) \subset (\mathrm{pa}_{\overline{\mathcal{G}}(\tilde{\mathbf{a}})}(V(\tilde{\mathbf{a}})) \setminus \tilde{\mathbf{a}})$, it follows that there exists an assignment $\mathbf{q}^*$ to $\mathrm{pa}_{\overline{\mathcal{G}}(\tilde{\mathbf{a}})}(V(\tilde{\mathbf{a}})) \setminus \tilde{\mathbf{a}}$ such that $\mathbf{q}'$ is a sub-vector of $\mathbf{q}^*$, and

$$P\left(\left(\mathrm{pa}_{\overline{\mathcal{G}}(\tilde{\mathbf{a}})}(V(\tilde{\mathbf{a}})) \setminus \tilde{\mathbf{a}}\right) = \mathbf{q}^*\right) > 0. \tag{96}$$

We then argue as follows:

$$\begin{aligned} &P\left(V(\tilde{\mathbf{a}}) = v \,\middle|\, \left(\mathrm{pa}_{\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})}(V(\tilde{\mathbf{a}})) \setminus \tilde{\mathbf{a}}\right) = \mathbf{q}'\right) \\ &\quad = P\left(V(\tilde{\mathbf{a}}) = v \,\middle|\, \left(\mathrm{pa}_{\overline{\mathcal{G}}(\tilde{\mathbf{a}})}(V(\tilde{\mathbf{a}})) \setminus \tilde{\mathbf{a}}\right) = \mathbf{q}^*\right) && (97) \\ &\quad = P\Big(V = v \,\Big|\, \left(\mathrm{pa}_{\overline{\mathcal{G}}}(V) \setminus \mathbf{A}\right) = \mathbf{q}^*,\ \ \left(\mathrm{pa}_{\overline{\mathcal{G}}}(V) \cap \mathbf{A}\right) = \tilde{\mathbf{a}}_{\mathrm{pa}_{\overline{\mathcal{G}}}(V) \cap \mathbf{A}}\Big) && (98) \\ &\quad = P\Big(V = v \,\Big|\, \left(\mathrm{pa}_{\mathcal{G}}(V) \setminus \mathbf{A}\right) = \mathbf{q}',\ \left(\mathrm{pa}_{\mathcal{G}}(V) \cap \mathbf{A}\right) = \tilde{\mathbf{a}}_{\mathrm{pa}_{\mathcal{G}}(V) \cap \mathbf{A}}\Big). && (99) \end{aligned}$$

Here the first equality follows by Theorem 40 since $P(\mathbb{V}(\tilde{\mathbf{a}}))$ obeys the Markov property for $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$; by definition of $\mathbf{q}^*$ both of these conditional distributions are well-defined. If both sides of (95) are well-defined then this implies the second equality. The third equality then follows because $P(\mathbf{V})$ obeys the Markov property for $\mathcal{G}$.

It remains to show that if the conditioning events in (97) and (99) have non-zero probability then so does the conditioning event in (98). There are two cases to consider:

If $\text{pa}_{\overline{\mathcal{G}}}(V) \setminus (\text{pa}_{\mathcal{G}}(V) \cup \mathbf{A}) \neq \emptyset$, so that the random parents of $V(\tilde{\mathbf{a}})$ in $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$ are a strict subset of the random parents of $V(\tilde{\mathbf{a}})$ in $\overline{\mathcal{G}}(\tilde{\mathbf{a}})$ then since (96) holds, Condition 39(I) implies that the conditioning event in (98) will have non-zero probability.[83]

If $V(\tilde{\mathbf{a}})$ has the same random parents in $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$ and $\overline{\mathcal{G}}(\tilde{\mathbf{a}})$ then

$$\text{pa}_{\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})}(V(\tilde{\mathbf{a}})) \setminus \tilde{\mathbf{a}} = \text{pa}_{\overline{\mathcal{G}}(\tilde{\mathbf{a}})}(V(\tilde{\mathbf{a}})) \setminus \tilde{\mathbf{a}}$$

and $\mathbf{q}' = \mathbf{q}^*$. Further $V(\tilde{\mathbf{a}})$ must have different fixed parents in $\overline{\mathcal{G}}(\tilde{\mathbf{a}})$ and $\overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}})$. Condition 39(II) then implies that whenever the conditioning events in both (97) and (99) are assigned non-zero probability then this will also be true of (98).

This establishes that $(P(\mathbb{V}(\tilde{\mathbf{a}})), \overline{\mathcal{G}}_{\mathcal{G}}(\tilde{\mathbf{a}}))$ obeys modularity with respect to $(P(\mathbf{V}), \mathcal{G})$ as required. □

7.2.1 Two variable example re-visited without positivity

We return to the simple SWIG $\overline{\mathcal{G}}_{\mathcal{G}}(b^\dagger)$ shown in Figure 26 discussed earlier in §7.1.2. In general, if, contrary to Condition 39(II), for some value $b^\dagger$ we have $P(B = b^\dagger) = 0$ then $P(Y(b^\dagger) = y)$ is not identified and hence, in general, $P(Y(b^\dagger) = y) \neq P(Y = y)$, so modularity fails. To see this consider the following simple NPSEM, in which B is ternary taking states $0, 1, 2$, and Y is binary:

$$B = \varepsilon_B \qquad\qquad Y(b) = I(b \leq 1)I(\varepsilon_Y = b) + I(b > 1)$$

where ε_Y is a Bernoulli(0.5), $P(\varepsilon_B = 0) = P(\varepsilon_B = 1) = 0.5$, but $P(\varepsilon_B = 2) = 0$, and $\varepsilon_B \perp\!\!\!\perp \varepsilon_Y$. Observe that for all units, $Y(b=0) = 1 - Y(b=1)$, but $Y(b=2) = 1$, however, $P(B=2) = 0$. It follows that for all $y \in \{0, 1\}$,

[83] Though not strictly required for the proof, this further implies that the conditioning (99) will also have non-zero probability.

$b \in \{0,1,2\}$:[84]

$$P(Y{=}y, B{=}b) = P(Y{=}y)P(B{=}b),$$

thus establishing factorization with respect to the edgeless graph in Figure 26(b). At the same time, modularity does *not* hold since:

$$1 = P(Y(b{=}2) = 1) \;\neq\; P(Y{=}1) = 1/2.$$

This is because owing to the violation of Condition 39(II), even though the left hand side of (80) and the right hand side of (81) are well-defined, because the expression on the right of (80) (and the left of (81)) is not well-defined. Thus both sides of (82) are well-defined, yet they are not equal, thereby violating modularity.

In spite of this, as implied by Proposition 16, $(P(B, Y(b^\dagger)), \overline{\mathcal{G}}_{\mathcal{G}}(b^\dagger))$ *does* obey modularity with respect $(P(B,Y), \overline{\mathcal{G}})$ since the modularity condition with respect to $\overline{\mathcal{G}}$ only requires that $P(Y(b^*){=}y) = P(Y{=}y \mid B{=}b^*)$, for all b^* for which both sides are well-defined, i.e. for $b^* = 0, 1$, but not 2.

However if we assume that graph $\mathcal{G}$ in Figure 26(b) was an FFRCISTG, so that there is no individual level effect of B on Y, then we do find that $P(Y(b^\dagger) = y)$ is identified and equal to $P(Y = y)$ even when $P(B = b^\dagger)$ is zero; the corresponding template $\mathcal{G}(b)$ is shown in Figure 26(e).

7.2.2 Population direct effect re-visited without positivity

Here we re-visit the three variable example discussed in 7.1.3.

To see the importance of the positivity condition Condition 39(II) consider the following NPSEM in which all variables are binary:

$$\begin{aligned} Z &= \varepsilon_Z; \\ M(z) &= I(z = 1) + I(z = 0) \cdot \varepsilon_M; \\ Y(z,m) &= I(\varepsilon_Y < (1 + m + 2(1-m)z)/4) \end{aligned}$$

where ε_Z, ε_M and ε_Y are jointly independent, $\varepsilon_Z, \varepsilon_M \sim \text{Ber}(0.5)$ and $\varepsilon_Y \sim \text{Unif}[0,1]$. Notice that $Z = 1$ implies $M = 1$. Further, we have $P(Y{=}1 \mid Z{=}z, M{=}1) = 1/2$, while $P(Y{=}1 \mid Z{=}0, M{=}0) = 1/4 = P(Y{=}1 \mid M{=}0)$. Thus for all pairs of values (z,m) for which $P(Z{=}z, M{=}m) > 0$ we have:

$$P(Y{=}1 \mid Z{=}z, M{=}m) = P(Y{=}1 \mid M{=}m),$$

[84] Here we use the fact that even though $P(Y{=}y|B{=}2)$ is not uniquely defined, since $P(B{=}2) = 0$, the product is defined.

so that (84) holds for all z, m, y, and thus $P(Z, M, Y)$ factors with respect to $\mathcal{G}$. However, since $P(M=0) > 0$, but $P(Z=1, M=0) = 0$, the condition (90) is violated. We see that

$$3/4 = P(Y(z=1, m=0) = 1) \neq P(Y=1 \mid M=0) = 1/4,$$

so that $(P(Z, M(\tilde{z}), Y(\tilde{z}, \tilde{m}), \overline{\mathcal{G}}_{\mathcal{G}}(z, m))$ does not obey modularity with respect to $(P(Z, M, Y), \mathcal{G})$. However, as in the previous example modularity is obeyed with respect to $(P(Z, M, Y), \overline{\mathcal{G}})$. If the counterfactual distribution was in the FFRCISTG associated with $\mathcal{G}$ (rather than $\overline{\mathcal{G}}$) then we would have modularity with respect to $(P(Z, M, Y), \mathcal{G})$.

8 SWIGs in the presence of restricted interventions

In §3 we based our theory on a NPSEM, under which counterfactuals are assumed well-defined for interventions on any variable: Formally a counterfactual distribution in the FFRCISTG model associated with $\mathcal{G}$ with vertex set $\mathbf{V}$ is a distribution over the set of counterfactual random variables

$$\{V(\tilde{\mathbf{c}}) \mid V \in \mathbf{V}, \tilde{\mathbf{c}} \in \mathfrak{C}, \mathbf{C} \subseteq \mathbf{V}\} \tag{100}$$

where $\mathfrak{C}$ is the state space of the set of variables $\mathbf{C}$.

However, there are many substantive contexts in which a process may not be sufficiently well understood for interventions on all variables to be regarded as well-defined.

Our SWIG theory may still be applied in this setting if, following (Robins, 1986)[85] we take as our starting point a counterfactual distribution for a smaller set of counterfactual variables:

$$\{V(\tilde{\mathbf{c}}) \mid V \in \mathbf{V}, \tilde{\mathbf{c}} \in \mathfrak{C}, \mathbf{C} \subseteq \mathbf{A}\}. \tag{101}$$

The set (101) differs from (100) in that we now only consider counterfactuals defined for intervention on subsets of the set of variables in $\mathbf{A}$. Thus we introduce the following:

Definition 43 (Counterfactual Existence Assumption Restricted to $\mathbf{A}$)**.**

(i) *The counterfactuals $V(\mathbf{a}^{\dagger}_{V})$ appearing in the SWIG $\mathcal{G}(\mathbf{a}^{\dagger})$ exist for any setting of $\mathbf{a}^{\dagger}$ in the state space $\mathfrak{A}$ of $\mathbf{A}$; recall that $\mathbf{A}_V \equiv \{A_j \mid A_j(\mathbf{a}^{\dagger}) \in \text{an}_{\mathcal{G}(\mathbf{a}^{\dagger})}(V(\mathbf{a}^{\dagger}_{\mathbf{V}})), A_j \in \mathbf{A}\}$.*

[85] See Appendix C.2

(ii) *Both the factual variables V and the counterfactuals $V(\mathbf{r})$ for any $\mathbf{R} \subseteq \mathbf{A}$ exist and are obtained recursively from the counterfactuals given in* (i) *as follows:*

$$V(\mathbf{r}^\dagger) = V\left(\mathbf{r}^\dagger_{\mathbf{A}_V \cap \mathbf{R}},\ \mathbf{B}(\mathbf{r}^\dagger)\right), \tag{102}$$

where $\mathbf{B}(\mathbf{r}^\dagger) \equiv \{A_j(\mathbf{r}^\dagger) \mid A_j(\mathbf{a}^\dagger_{A_j}) \in \text{an}_{\mathcal{G}(\mathbf{a}^\dagger)}(V(\mathbf{a}^\dagger_{\mathbf{V}})), A_j \in \mathbf{A} \setminus \mathbf{R}\}$.

Here the set $\mathbf{B}(\mathbf{r}^\dagger)$ consists of the counterfactual variables corresponding to variables in $\mathbb{A}(\mathbf{a}^\dagger)$ that are ancestors of $V(\mathbf{a}^\dagger)$ in $\mathcal{G}(\mathbf{a}^\dagger)$, but are not in $\mathbf{R}$, evaluated under the intervention $\mathbf{r}^\dagger$.

Our model is then defined via the local Markov property (Lauritzen, 1996) associated with the set of SWIGs $\{\mathcal{G}(\mathbf{a}^\dagger) \mid \mathbf{a}^\dagger \in \mathfrak{A}\}$ associated with the template $\mathcal{G}(\mathbf{a})$:

Definition 44 (FFRCISTG Indep. Assumption for $\mathcal{G}$ restricted to $\mathbf{A}$)**.**
For every $\mathbf{a}^\dagger \in \mathfrak{A}$, and every $V \in \mathbf{V}$, we assume the following:

$$V(\mathbf{a}^\dagger_V) \perp\!\!\!\perp \text{npnd}_{\mathcal{G}(\mathbf{a}^\dagger)}(V(\mathbf{a}^\dagger_V)) \mid \text{pa}_{\mathcal{G}(\mathbf{a}^\dagger)}(V(\mathbf{a}^\dagger_V)) \setminus \mathbf{a}^\dagger \tag{103}$$

where $\text{npnd}_{\mathcal{G}(\mathbf{a}^\dagger)}(V(\mathbf{a}^\dagger_V)$ *is the set of* non-parental non-descendants *of* $V(\mathbf{a}^\dagger_V)$ *in* $\mathcal{G}(\mathbf{a}^\dagger)$.

Proposition 45. *Under the FFRCISTG model associated with $\mathcal{G}$ restricted to $\mathbf{A}$, for any $\mathbf{C} \subseteq \mathbf{A}$ and assignment $\tilde{\mathbf{c}}$, $P(\mathbb{V}(\mathbf{c}^\dagger))$ factorizes according to $\mathcal{G}(\mathbf{c}^\dagger)$.*

Proposition 46. *Under the FFRCISTG model associated with $\mathcal{G}$ restricted to $\mathbf{A}$, for any $\mathbf{C} \subseteq \mathbf{A}$ and assignment $\tilde{\mathbf{c}}$, the pairs $(P(\mathbf{V}), \mathcal{G})$ and $\big(P(\mathbb{V}(\mathbf{c}^\dagger)), \mathcal{G}(\mathbf{c}^\dagger)\big)$ obey modularity.*

Proof: The local Markov property given Definition 44 implies that the distribution over the variables $P(\mathbb{V}(\mathbf{a}^\dagger))$ factorizes according to $\mathcal{G}(\mathbf{a}^\dagger)$. These results then follow directly via the inductive proof of Propositions 11 and 16; see Appendix B.1. □

It is perhaps natural to ask what is the meaning in our theory of a directed edge $X \to Y$ in the case where X is not in $\mathbf{A}$ and so counterfactual variables involving interventions on X do not arise in our model. We offer several possibilities but do not select amongst them:

One view is that there does exist some underlying FFRCISTG where directed edges continue to represent causal effects however, only the interventions on the variables in $\mathbf{A}$ are sufficiently well-specified to support

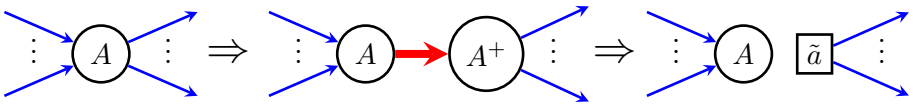
well-defined counterfactuals. Consequently we choose to restrict the interventions we consider, and indeed our domain of discourse, to just those counterfactuals for interventions in $\mathbf{A}$.

An alternative view that is more in the spirit of the section on population exclusion restrictions is to believe that there is no underlying FFRCISTG associated with a DAG $\mathcal{G}$ (even including hidden variables). However, there does exist an FFRCISTG with interventions restricted to $\mathbf{A}$, and the DAG $\mathcal{G}$ simply reflects the factorization properties of the observed distribution. This leaves open the question as to why one would expect the distribution of the observables to factorize according to anything other than a complete DAG; also why one might not also consider other factorizations such as those corresponding to chain graphs (Lauritzen, 1996).

9 An intervention-based interpretation of $\mathcal{G}(\tilde{\mathbf{a}})$

Formally, the factorization and modularity conditions associated with the graph $\mathcal{G}(\tilde{\mathbf{a}})$ can be motivated from a purely interventional perspective via an extended agnostic causal DAG[86] (Spirtes et al., 1993). In other words, we will show that someone who, possibly for philosophical reasons, did not wish to postulate the existence of counterfactual outcomes, but merely wished to reason in terms of distributions resulting from interventions, might still wish to construct a graph isomorphic to the SWIG $\mathcal{G}(\tilde{\mathbf{a}})$.

Suppose that for each variable $A \in \mathbf{A}$, on which we are intervening, it were possible to first observe the natural value of the variable, and then to intervene. We might represent this by replacing A by two 'copies', A and A^+, linked by a deterministic edge such that $A = A^+$ in the observed data with probability 1, thus:[87]



Schematic: Replacement of A with A and A^+; intervention on A^+.

[86] Also known as a 'causal Bayesian network' (Pearl, 2000); this model is also very similar to a 'causal influence diagram' Dawid and Didelez (2010).

[87] This construction generalizes that described by (Geneletti and Dawid, 2007) and (Robins et al., 2007).

The topology of the resulting manipulated graph, after intervening on the variables $\mathbf{A}^+$, would be isomorphic to $\mathcal{G}(\tilde{\mathbf{a}})$. Using the standard truncated factorization formula, the distribution over $\mathbf{V}$ (the set of all variables other than $\{A_j^+\}$) under an intervention to set A_i^+ to $\tilde{a}_i$ for all $A_i \in \mathbf{A}$, would be the extended g-formula (37). Consequently factorization and modularity would be obeyed in the obvious way.

One might wonder whether an intervention under which one observes the natural value of a treatment variable before intervening is possible even in principle for human subjects. Recent evidence indicates that if one records data in continuous time (or nearly so) it may be quite generally. Specifically by observing the activity of certain neurons one can predict a person's actions some milliseconds before the individual is consciously aware that they are taking the action.

Acknowledgments

This research was supported by the U.S. National Science Foundation (CRI 0855230) and U.S. National Institutes of Health (R01 AI032475). We thank Wen Wei Loh for helpful comments on earlier drafts.

References

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

Box, G. (1966). Use and abuse of regression. *Technometrics 8*, 625–629.

Danaei, G., A. Pan, F. B. Hu, and M. A. Hernán (2012). Hypothetical lifestyle interventions in middle-aged women and risk of type 2 diabetes: a 24-year prospective study. *Am. J. Epidemiology*, In press.

Davis, J. A. (1984). Extending Rosenberg's technique for standardizing percentage tables. *Social Forces 62*, 679–708.

Dawid, A. (2000). Causal inference without counterfactuals (with discussion). *J. Amer. Statist. Assoc. 95*, 407–448.

Dawid, A. P. and V. Didelez (2008). Identifying optimal sequential decisions. In D. McAllester and A. Nicholson (Eds.), *Proceedings of the 24th Annual*

Conference on Uncertainty in Artificial Intelligence, Corvallis, Oregon, pp. 113–120.

Dawid, A. P. and V. Didelez (2010). Identifying the consequences of dynamic treatment strategies: a decision-theoretic overview. *Stat. Surv. 4*, 184–231.

Drton, M. and T. Richardson (2008). Binary models for marginal independence. *J. Roy. Statist. Soc. Ser. B 70(2)*, 287–309.

Frangakis, C. E. and D. B. Rubin (2002). Principal stratification in causal inference. *Biometrics 58*(1), 21–29.

Geiger, D. (1990). *Graphoids: a qualitative framework for probabilistic inference*. Ph. D. thesis, UCLA.

Geiger, D. and J. Pearl (1993). Logical and algorithmic properties of conditional independence and graphical models. *The Annals of Statistics 21*, 2001–2021.

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.

Greenland, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology 14*(3), 300–306.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica 11*, 1–12.

Haneuse, S. and A. Rotnitzky (2013). Estimation of the effect of interventions that modify the received treatment. Submitted.

Huang, Y. and M. Valtorta (2006). Pearl's calculus of interventions is complete. In *Proceedings of the 22nd Conference On Uncertainty in Artificial Intelligence*.

Lajous, M., W. C. Willett, J. M. Robins, J. G. Young, E. Rimm, D. Mozaffarian, and M. A. Hernán (2012). Changes in fish consumption in midlife and the risk of coronary heart disease in men and women. *Am. J. Epidemiology*, In press.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford, UK: Clarendon Press.

Lauritzen, S. L., A. P. Dawid, B. Larsen, and H.-G. Leimer (1990). Independence properties of directed Markov fields. *Networks 20*, 491–505.

Meek, C. (1995). Strong completeness and faithfulness in Bayesian networks. In P. Besnard and S. Hanks (Eds.), *Uncertainty in Artificial Intelligence: Proceedings of the* 11th *Conference*, San Francisco, pp. 411–418. Morgan Kaufmann.

Muñoz, I. D. and M. J. van der Laan (2011). Population intervention causal effects based on stochastic interventions. Technical Report 289, Division of Biostatistics, University of California, Berkeley.

Muñoz, I. D. and M. J. van der Laan (2012). Assessing the causal effect of policies: An approach based on stochastic interventions. Technical Report 298, Division of Biostatistics, University of California, Berkeley.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych X*, 1–51. In Polish, English translation by D. Dabrowska and T. Speed in *Statistical Science* **5** 463–472, 1990.

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. (2001a). Direct and indirect effects. In J. S. Breese and D. Koller (Eds.), *UAI-01, Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence*, San Francisco, pp. 411–42. Morgan Kaufmann.

Pearl, J. (2001b). Direct and indirect effects. In J. S. Breese and D. Koller (Eds.), *Proceedings of the 17th Annual Conference on Uncertainty in Artificial Intelligence*, San Francisco, pp. 411–42. Morgan Kaufmann.

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

Pearl, J. (2010). Causal analysis in theory and practice: On mediation, counterfactuals and manipulations. Posting to UCLA Causality Blog, 3 May 2010; `http://www.mii.ucla.edu/causality/?p=133`.

Pearl, J. (2012a, July). Eight myths about causality and structural equation models. Technical Report R-393, Computer Science Department, UCLA.

Pearl, J. (2012b). Trygve Haavelmo and the emergence of causal calculus. Technical Report R-391, Computer Science Department, UCLA.

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.

Robins, J. M. (1984). A statistical method to control for the healthy worker effect in intracohort comparisons. *Am. J. Epidemiology 120*, 465.

Robins, J. M. (1985). A new theory of causality in observational survival studiesapplication to the healthy worker effect. *Biometrics 41*, 331.

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. (1987). Addendum to: *A new approach to causal inference in mortality studies with sustained exposure periods – Application to control of the healthy worker survivor effect. Computers and Mathematics with Applications 14*, 923–945.

Robins, J. M. (1989a). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In L. Sechrest, H. Freeman, and A. Mulley (Eds.), *Health Service Research Methodology: A focus on AIDS*. Washington, D.C.: U.S. Public Health Service.

Robins, J. M. (1989b). The control of confounding by intermediate variables. *Statistics in Medicine 8*, 679–701.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In M. Berkane (Ed.), *Latent variable modelling and applications to causality*, Number 120 in Lecture notes in statistics, pp. 69–117. New York: Springer-Verlag.

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In P. Green, N. Hjort, and S. Richardson (Eds.), *Highly Structured Stochastic Systems*, pp. 70–81. Oxford: Oxford University Press.

Robins, J. M., D. Blevins, G. Ritter, and M. Wulfsohn (1992). *G*-estimation of the effect of prophylaxis therapy for *Pneumocystis Carinii Pneumonia* on the survival of AIDS patients. *Epidemiology 3*, 319–336.

Robins, J. M. and S. Greenland (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology 3*, 143–155.

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 (2007). 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 (2006a). Identification of conditional interventional distributions. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, Volume 22.

Shpitser, I. and J. Pearl (2006b). Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence*.

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 (2012). Effects of treatment on the treated: Identification and generalization. *Proceedings of the 25th Conference On Uncertainty in Artificial Intelligence*.

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. J. VanderWeele, and J. M. Robins (2012). On the validity of covariate adjustment for estimating causal effects. *CoRR abs/1203.3515*.

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

Strotz, R. and H. Wold (1960). Recursive versus non-recursive systems: An attempt at synthesis. *Econometrica 28*(2), 417–427.

Taubman, S. L., M. A. Mittleman, J. M. Robins, and M. A. Hernán (2008). Alternative approaches to estimating the effects of hypothetical interventions. In *JSM Proceedings, Health Policy Statistics Section*, Alexandria, VA.

Taubman, S. L., J. M. Robins, M. A. Mittleman, and M. A. Hernán (2009). Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *Int. J. Epidemiology 38*(6), 1599611.

Tian, J. (2008). Identifying dynamic sequential plans. In *24th Conference on Uncertainty in Artificial Intelligence (UAI-08)*. AUAI Press.

Tian, J. and J. Pearl (2002). A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, pp. 567–573.

Young, J., J. M. Robins, and M. A. Hernán (2012). Identification and estimation of risk under threshold interventions using observational data. Submitted.

A Graphs

We now introduce graphical models, based on directed acyclic graphs. We first require the following definitions: A directed acyclic graph (DAG) $\mathcal{G}(\mathbf{V}, \mathbf{E})$ with vertex set $\mathbf{V}$ and edge set $\mathbf{E}$ is a graph containing directed edges ($\rightarrow$) subject to the restriction that there are no directed cycles $V \rightarrow \cdots \rightarrow V$. We define the *parents* of V to be $\text{pa}_{\mathcal{G}}(V) \equiv \{X \,|\, X \rightarrow V\}$. If $B \rightarrow a$ then

we say that B is a *parent* of A, and A is a *child* of B. A vertex A is said to be an *ancestor* of a vertex D if *either* there is a directed path $A \to \cdots \to D$ from A to D, *or* $A = D$; similarly D is said to be a *descendant* of A.

A.1 Factorization criteria

Definition 47. *A distribution $P(\mathbf{V})$ is said to* factorize *with respect to a DAG $\mathcal{G}$ if*

$$P(\mathbf{V}) = \prod_{V \in \mathbf{V}} P(V \mid \mathrm{pa}_{\mathcal{G}}(V)). \tag{104}$$

A.2 Markov properties

The Markov property associated with a DAG is given by the following definitions:

A path π is said to *d-connect vertices A and B conditional on a set* $\mathbf{C}$ if A and B are the endpoints of π and:

(i) Every non-collider on π is not in $\mathbf{C}$, and

(ii) Every collider on π is an ancestor of $\mathbf{C}$ (or is in $\mathbf{C}$).

If there is no path d-connecting A and B given $\mathbf{C}$ then A and B are said to be *d-separated given* $\mathbf{C}$ in $\mathcal{G}$. The following well-known result connects d-separation and conditional independence in distributions factoring according to $\mathcal{G}$.

Theorem 48. *In any distribution $P(\mathbf{V})$ that factorizes according to $\mathcal{G}$, if X and Y are d-separated given $\mathbf{Z}$ then $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ in P.*

B Proofs of main results

We begin with two useful Lemmas relating counterfactual distributions to actual distributions.

Lemma 49. *If $P(\mathbf{V})$ and $P(\mathbb{V}(\tilde{\mathbf{a}}))$ factorize with respect to $\mathcal{G}$ and $\mathcal{G}(\tilde{\mathbf{a}})$ respectively and obey modularity then for any $\mathbf{W} \subseteq \mathbf{V} \setminus \mathbf{A}$,*

$$P(\mathbb{W}(\tilde{\mathbf{a}}) = \mathbf{w}, \mathbb{A}(\tilde{\mathbf{a}}) = \tilde{\mathbf{a}}) = P(\mathbf{W} = \mathbf{w}, \mathbf{A} = \tilde{\mathbf{a}}).$$

It is an immediate consequence of this lemma that if $P(\mathbb{W}^*(\tilde{\mathbf{a}}) = \mathbf{w}^*, \mathbb{A}(\tilde{\mathbf{a}}) = \tilde{\mathbf{a}}) = P(\mathbf{W}^* = \tilde{\mathbf{w}}^*, \mathbf{A} = \tilde{\mathbf{a}})$ then the same equality holds replacing $\mathbf{W}^*$ by any subset $\mathbf{W} \subseteq \mathbf{W}^*$. In other words the equality is preserved if we marginalize over variables that are not in $\mathbf{A}$ (or $\mathbb{A}(\tilde{\mathbf{a}})$).

Proof: This follows by summing both sides of (37) over $v_i \in \mathbf{V} \setminus \mathbf{W}$. □.

We illustrate this Lemma with an example. Consider the template in Figure 9(iii). We have the following factorization for all $z, m, \tilde{m}, y$:

$$P(Z = z, M = m, Y(\tilde{m}) = y) = P(Z = z)P(M = m \mid Z = z)P(Y = y \mid Z = z, M = \tilde{m}).$$

Consequently setting $m = \tilde{m}$ we obtain:

$$\begin{aligned} P(Z = z, M = \tilde{m}, Y(\tilde{m}) = y) &= P(Z = z)P(M = \tilde{m} \mid Z = z)P(Y = y \mid Z = z, M = \tilde{m}) \\ &= P(Z = z, M = \tilde{m}, Y = y). \end{aligned} \tag{105}$$

We see that summing over z or y leads to equalities for the following margins:

$$P(M = \tilde{m}, Y(\tilde{m}) = y) = P(M = \tilde{m}, Y = y) \quad P(Z = z, M = \tilde{m}) = P(Z = z, M = \tilde{m}),$$

the latter being a tautology. However, we do not have equality of $P(Z, Y(\tilde{m}))$ and $P(Z, Y)$ since the equality (105) of the joint distributions only holds when $m = \tilde{m}$.

The next Lemma shows that we may preserve factorization and modularity when summing over a variable $A \in \mathbf{A}$ in the special case where A has no children (and hence no descendants) in the graph. It further follows from Proposition 17 that (17) is also preserved when marginalizing over A.

Lemma 50. *Given a DAG $\mathcal{G}$, let $T \in \mathbf{V}$ such that $ch_{\mathcal{G}}(T) = \emptyset$. Let $\mathbf{V}' = \mathbf{V} \setminus \{T\}$ and $\mathbb{V}'(\tilde{\mathbf{a}}) = \mathbb{V}(\tilde{\mathbf{a}}) \setminus \{T(\mathbf{a}_T)\}$. Further, let $\mathcal{G}' = \mathcal{G}_{\mathbf{V}'}$ and $\mathcal{G}'(\tilde{\mathbf{a}}) = (\mathcal{G}(\tilde{\mathbf{a}}))_{\mathbb{V}'(\tilde{\mathbf{a}}) \cup \tilde{\mathbf{a}}}$ be the induced subgraphs obtained by removing T and $T(\mathbf{a}_T)$ from $\mathcal{G}$ and $\mathcal{G}(\tilde{\mathbf{a}})$. If $P(\mathbf{V})$ and $P(\mathbb{V}(\tilde{\mathbf{a}}))$ factorize and obey modularity with respect to $\mathcal{G}$ and $\mathcal{G}(\tilde{\mathbf{a}})$ respectively then $P(\mathbf{V}')$ and $P(\mathbb{V}'(\tilde{\mathbf{a}}))$ factorize and obey modularity with respect to $\mathcal{G}'$ and $\mathcal{G}'(\tilde{\mathbf{a}}')$ respectively.*

Proof: This follows from the definition of factorization and modularity. □

B.1 Proof that under the FFRCISTG independence assumption the NPSEM obeys factorization and modularity

We prove Propositions 11 and 16 together:

Proof: We will first establish that $P(\mathbb{V}(\mathbf{a}^\dagger))$ factorizes according to $\mathcal{G}(\mathbf{a}^\dagger)$ by reverse induction on the size of the set $\mathbf{A}$. The claim holds immediately for the case $\mathbf{A} = \mathbf{V}$ since then the random nodes in $\mathcal{G}(\mathbf{a}^\dagger)$ are all disconnected and take the form $Y(\mathbf{pa}_Y)$ (with all parents fixed). The (trivial) factorization of $P(\mathbb{V}(\mathbf{a}^\dagger))$ then follows directly from the FFRCISTG independence (17).

Now consider a set $\mathbf{A} \subsetneq \mathbf{V}$ and suppose that the factorization claim is true for all sets larger than $\mathbf{A}$. Let C be a vertex in $\mathbf{V} \setminus \mathbf{A}$ such that in $\mathcal{G}$, all the descendants of C (other than C itself) are in $\mathbf{A}$; such a vertex is guaranteed to exist since $\mathcal{G}$ is acyclic. Let $\mathbf{B} \equiv \mathbf{A} \cup \{C\}$. Let $\mathcal{G}(\mathbf{a}^\dagger)$ and $\mathcal{G}(\mathbf{b}^\dagger)$ be the corresponding SWIGs, and $V(\mathbf{a}^\dagger_V)$ and $V(\mathbf{b}^\dagger_V)$, respectively, indicate the random node corresponding to a given vertex $V \in \mathbf{V}$.

By the inductive hypothesis $P(\mathbb{V}(\mathbf{b}^\dagger))$ factorizes according to $\mathcal{G}(\mathbf{b}^\dagger)$, so for all $\tilde{\mathbf{v}}$ and $\mathbf{b}^\dagger$,

$$P\left(\mathbb{V}(\mathbf{b}^\dagger) = \tilde{\mathbf{v}}\right) = \prod_{V(\mathbf{b}^\dagger_V) \in \mathbb{V}(\mathbf{b}^\dagger)} P\left(V(\mathbf{b}^\dagger_V) = \tilde{v} \,\middle|\, \left(\mathrm{pa}_{\mathcal{G}(\mathbf{b}^\dagger)}(V(\mathbf{b}^\dagger_V)) \setminus \mathbf{b}^\dagger\right) = \tilde{\mathbf{w}}_V\right), \tag{106}$$

where $\tilde{\mathbf{w}}_V \equiv \tilde{\mathbf{v}}_{\mathrm{pa}_{\mathcal{G}(\mathbf{b}^\dagger)}(V(\mathbf{b}^\dagger_V)) \setminus \mathbf{b}^\dagger}$ is the subvector of $\tilde{\mathbf{v}}$ corresponding to the random parents of $V(\mathbf{b}^\dagger_V)$ in $\mathcal{G}(\mathbf{b}^\dagger)$.

First observe that it follows from recursive substitution (see Definition 1 (iii)) that for all $\tilde{\mathbf{v}}$ and $\mathbf{a}^\dagger$, if we define $\mathbf{b}^\dagger \equiv (\mathbf{a}^\dagger, c^\dagger = \tilde{c})$, then

$$P(\mathbb{V}(\mathbf{a}^\dagger) = \tilde{\mathbf{v}}) = P(\mathbb{V}(\mathbf{b}^\dagger) = \tilde{\mathbf{v}}),$$

where $\tilde{c}$ is the value assigned to C by $\tilde{\mathbf{v}}$; the equality follows since the event on the LHS includes $C(\mathbf{a}^\dagger_C) = \tilde{c} = c^\dagger$. It is thus sufficient to prove that when $\mathbf{b}^\dagger \equiv (\mathbf{a}^\dagger, c^\dagger = \tilde{c})$, the RHS of (106) is equivalent to the factorization associated with $\mathcal{G}(\mathbf{b}^\dagger)$.

We now list the ways in which $\mathcal{G}(\mathbf{a}^\dagger)$ and $\mathcal{G}(\mathbf{b}^\dagger)$ differ:

(i) There is a fixed node c in $\mathcal{G}(\mathbf{b}^\dagger)$ that is not present in $\mathcal{G}(\mathbf{a}^\dagger)$.

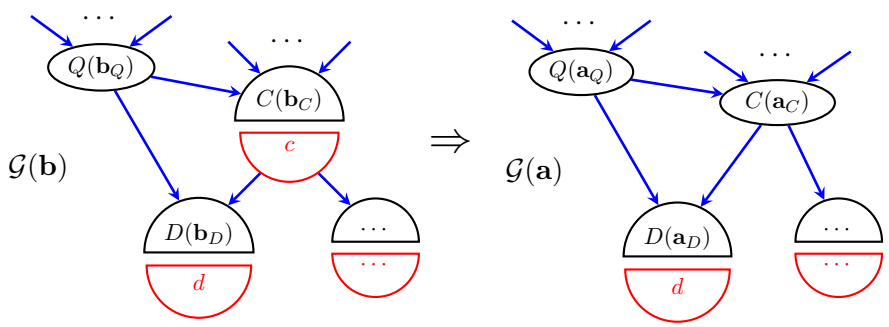
(ii) For any vertex D that is a child of C in $\mathcal{G}$, by the choice of C, $D \in \mathbf{A}$. Further, $\mathbf{a}^\dagger_D = (\mathbf{b}^\dagger_D \setminus \{c^\dagger\}) \cup \mathbf{b}^\dagger_C$; in words, the set $\mathbf{a}^\dagger_D$ labelling the random node corresponding to D in $\mathcal{G}(\mathbf{a}^\dagger)$ consists of the set labelling this node in $\mathcal{G}(\mathbf{b}^\dagger)$ after having removed c and then adding the set $\mathbf{b}^\dagger_C$ that labels C in $\mathcal{G}(\mathbf{b}^\dagger)$.

(iii) For all other vertices V such that $V \notin \mathrm{ch}_{\mathcal{G}}(C)$, the labelling is unchanged so $\mathbf{a}_V^\dagger = \mathbf{b}_V^\dagger$, in particular we have $\mathbf{a}_C^\dagger = \mathbf{b}_C^\dagger$, hence $C(\mathbf{a}_C^\dagger) = C(\mathbf{b}_C^\dagger)$.

(iv) In $\mathcal{G}(\mathbf{b}^\dagger)$, $C(\mathbf{b}_C^\dagger)$ has no children, while in $\mathcal{G}(\mathbf{b}^\dagger)$ it has edges to $D(\mathbf{a}_D^\dagger)$ for every $D \in \mathrm{ch}_{\mathcal{G}}(C)$. No other edges differ.

(v) If a vertex $Q \in \mathrm{pa}_{\mathcal{G}}(D) \setminus \mathbf{B}$, where $D \in \mathrm{ch}_{\mathcal{G}}(C)$ then by (iii), $Q(\mathbf{a}_Q^\dagger) = Q(\mathbf{b}_Q^\dagger)$. Thus the set of random parents of $D(\mathbf{a}_D^\dagger)$ in $\mathcal{G}(\mathbf{a}^\dagger)$ is equal to the set of random parents of $D(\mathbf{b}_D^\dagger)$ in $\mathcal{G}(\mathbf{b}^\dagger)$ together with $C(\mathbf{a}_C^\dagger) = \mathbf{b}_C^\dagger$. Consequently,

$$\mathrm{pa}_{\mathcal{G}(\mathbf{a}^\dagger)}(D(\mathbf{a}_D^\dagger)) \setminus \mathbf{a}^\dagger = \left(\mathrm{pa}_{\mathcal{G}(\mathbf{b}^\dagger)}(D(\mathbf{b}_D^\dagger)) \setminus \mathbf{b}^\dagger\right) \cup \{C(\mathbf{b}_C^\dagger)\}.$$



Schematic depicting the transformation $\mathcal{G}(\mathbf{b}) \mapsto \mathcal{G}(\mathbf{a})$, where $\mathbf{b} = \mathbf{a} \cup \{c\}$. By construction, C is chosen so that every descendant of C in $\mathcal{G}$ is in $\mathbf{A}$. Thus every child of c in $\mathcal{G}(\mathbf{b})$ has no descendants.
Note that $\mathbf{a}_Q = \mathbf{b}_Q$, $\mathbf{a}_C = \mathbf{b}_C$, $\mathbf{a}_D = (\mathbf{b}_D \setminus \{c\}) \cup \mathbf{b}_C$.

These facts are summarized in the schematic diagram above. It follows that in the factorizations associated with $\mathcal{G}(\mathbf{b}^\dagger)$ and $\mathcal{G}(\mathbf{a}^\dagger)$, the only terms that differ are those associated with vertices $D \in \mathrm{ch}_{\mathcal{G}}(C)$.

Consider such a vertex D, let $\tilde{d} = \tilde{v}_D$, $\tilde{c} = \tilde{v}_C$ and $\tilde{\mathbf{w}}_D$ the entries in $\tilde{\mathbf{v}}$

corresponding to D, C and $\mathrm{pa}_{\mathcal{G}(\mathbf{b}^\dagger)}(D(\mathbf{b}^\dagger_D)) \setminus \mathbf{b}^\dagger$ respectively. Now

$$
\begin{aligned}
&P\left(D(\mathbf{b}^\dagger_D)=\tilde{d} \;\middle|\; \left(\mathrm{pa}_{\mathcal{G}(\mathbf{b}^\dagger)}(D(\mathbf{b}^\dagger_D)) \setminus \mathbf{b}^\dagger\right)=\tilde{\mathbf{w}}_D\right) \\
&\quad = P\left(D(\mathbf{b}^\dagger_D)=\tilde{d} \;\middle|\; \left(\mathrm{pa}_{\mathcal{G}(\mathbf{b}^\dagger)}(D(\mathbf{b}^\dagger_D)) \setminus \mathbf{b}^\dagger\right)=\tilde{\mathbf{w}}_D, C(\mathbf{b}^\dagger_C)=c^\dagger\right) \\
&\quad = P\left(D(\mathbf{a}^\dagger_D)=\tilde{d} \;\middle|\; \left(\mathrm{pa}_{\mathcal{G}(\mathbf{b}^\dagger)}(D(\mathbf{b}^\dagger_D)) \setminus \mathbf{b}^\dagger\right)=\tilde{\mathbf{w}}_D, C(\mathbf{b}^\dagger_C)=c^\dagger\right) \\
&\quad = P\left(D(\mathbf{a}^\dagger_D)=\tilde{d} \;\middle|\; \left(\mathrm{pa}_{\mathcal{G}(\mathbf{a}^\dagger)}(D(\mathbf{a}^\dagger_D)) \setminus \mathbf{a}^\dagger\right)=(\tilde{\mathbf{w}}_D, c^\dagger)\right) \\
&\quad = P\left(D(\mathbf{a}^\dagger_D)=\tilde{d} \;\middle|\; \left(\mathrm{pa}_{\mathcal{G}(\mathbf{a}^\dagger)}(D(\mathbf{a}^\dagger_D)) \setminus \mathbf{a}^\dagger\right)=(\tilde{\mathbf{w}}_D, \tilde{c})\right), \qquad (107)
\end{aligned}
$$

but $(\tilde{\mathbf{w}}_D, \tilde{c}) = \tilde{\mathbf{v}}_{\mathrm{pa}_{\mathcal{G}(\mathbf{a}^\dagger)}(V(\mathbf{a}^\dagger_V))\setminus\mathbf{a}^\dagger}$, the sub-vector of $\tilde{\mathbf{v}}$ corresponding to the random parents of $D(\mathbf{a}^\dagger_D)$ in $\mathcal{G}(\mathbf{a}^\dagger)$. Here the first equality follows from the Markov property for $\mathcal{G}(\mathbf{b}^\dagger)$, the second follows from recursive substitution, the third follows from the fact that the random parents of $D(\mathbf{a}^\dagger_D)$ in $\mathcal{G}(\mathbf{a}^\dagger)$ are the random parents of $D(\mathbf{b}^\dagger_D)$ in $\mathcal{G}(\mathbf{b}^\dagger)$ together with $C(\mathbf{a}^\dagger_C) = C(\mathbf{b}^\dagger_C)$. The fourth follows since $c^\dagger = \tilde{c}$.

Thus $P(\mathbb{V}(\mathbf{a}^\dagger))$ factorizes according to $\mathcal{G}(\mathbf{a}^\dagger)$, as required. This establishes Proposition 11.

Notice that (107) shows that the conditional density associated with $D(\mathbf{b}^\dagger_D)$ in $\mathcal{G}(\mathbf{b}^\dagger)$ is equal to the conditional density associated with $D(\mathbf{a}^\dagger_D)$ in $\mathcal{G}(\mathbf{a}^\dagger)$, when the vertex $C(\mathbf{a}^\dagger_C)$ that is a (random) parent of $D(\mathbf{a}^\dagger_D)$ in $\mathcal{G}(\mathbf{b}^\dagger)$ takes the value $c^\dagger$. Consequently, starting with $\mathcal{G}(\mathbf{a}^\dagger)$, and repeatedly removing vertices from the set of fixed vertices, via (107), we will eventually arrive back at the original graph $\mathcal{G}$ and the factorization of the actual variables $P(\mathbf{V})$, hence

$$
\begin{aligned}
&P\left(D(\mathbf{a}^\dagger_D)=\tilde{d} \;\middle|\; \left(\mathrm{pa}_{\mathcal{G}(\mathbf{a}^\dagger)}(D(\tilde{\mathbf{a}}_D)) \setminus \mathbf{a}^\dagger\right)=\tilde{\mathbf{w}}_D\right) \\
&\qquad = P\Big(D=\tilde{d} \;\Big|\; \left(\mathrm{pa}_{\mathcal{G}}(D) \setminus \mathbf{A}\right)=\tilde{\mathbf{w}}_D, \;\; \left(\mathrm{pa}_{\mathcal{G}}(Y) \cap \mathbf{A}\right) = \mathbf{a}^\dagger_{\mathrm{pa}_{\mathcal{G}}(Y)\cap\mathbf{A}}\Big).
\end{aligned}
$$

We have thus shown that Proposition 16 holds. □

B.2 Proof of the necessary and sufficient condition for the g-formula

Recall that we have factual variables:

$$H_1, L_1, A_1, \ldots H_{K-1}, L_{K-1}, A_{K-1}, H_K, L_K, A_K, Y$$

where H_t, L_t and A_t are, respectively, hidden variables, covariates and treatments at stage t in our sequence. Further, let $W_t \equiv (H_t, L_t)$, be the observed and unobserved covariates at stage t. We define:

$$b^{\text{full}}_{\mathbf{a}^\dagger}(y \mid \overline{\mathbf{h}}_m, \bar{\mathbf{l}}_m) \equiv \sum_{w_{m+1},\ldots,w_K} p(y|\overline{\mathbf{w}}_K, \bar{\mathbf{a}}^\dagger_K) \prod_{j=m+1}^{K} p(w_j|\overline{\mathbf{w}}_{j-1}, \bar{\mathbf{a}}^\dagger_{j-1}) \quad (108)$$

$$b_{\mathbf{a}^\dagger}(y \mid \bar{\mathbf{l}}_m) \equiv \sum_{l_{m+1},\ldots,l_K} p(y|\bar{\mathbf{l}}_K, \bar{\mathbf{a}}^\dagger_K) \prod_{j=m+1}^{K} p(l_j|\bar{\mathbf{l}}_{j-1}, \bar{\mathbf{a}}^\dagger_{j-1}) \quad (109)$$

to be, respectively, the g-functional at m for the full data and the g-functional at m for the observed data, for regime $\mathbf{A} = \mathbf{a}^\dagger$.

Lemma 51. *It follows by definition that:*

$$b^{\text{full}}_{\mathbf{a}^\dagger}(y \mid \overline{\mathbf{h}}_m, \bar{\mathbf{l}}_m) = \sum_{h_{m+1},l_{m+1}} b^{\text{full}}_{\mathbf{a}^\dagger}(y \mid \overline{\mathbf{h}}_{m+1}, \bar{\mathbf{l}}_{m+1})\, p(h_{m+1}, l_{m+1} \mid \overline{\mathbf{w}}_m, \overline{\mathbf{A}}_m = \bar{\mathbf{a}}^\dagger_m), \quad (110)$$

$$b_{\mathbf{a}^\dagger}(y \mid \bar{\mathbf{l}}_m) = \sum_{l_{m+1}} b_{\mathbf{a}^\dagger}(y \mid \bar{\mathbf{l}}_{m+1})\, p(l_{m+1} \mid \bar{\mathbf{l}}_m, \overline{\mathbf{A}}_m = \bar{\mathbf{a}}^\dagger_m). \quad (111)$$

Our main result is as follows:

Theorem 22. *Let $\mathbf{a}^\dagger$ be an instantiation for $\mathbf{A}$. If $P(\mathbf{V})$ and $P(\mathbb{V}(\mathbf{a}^\dagger))$ factor according to $\mathcal{G}$ and $\mathcal{G}(\mathbf{a}^\dagger)$ respectively, and obey modularity (30) then for all $j = 0, \ldots, K$, and all $\bar{\mathbf{l}}_j$,*

$$b_{\mathbf{a}^\dagger}(y \mid \bar{\mathbf{l}}_j) = P(Y(\mathbf{a}^\dagger) = y \mid \overline{\mathbb{L}}_j(\mathbf{a}^\dagger) = \bar{\mathbf{l}}_j, \overline{\mathbb{A}}_{j-1}(\mathbf{a}^\dagger) = \bar{\mathbf{a}}^\dagger_{j-1}) \quad (44)$$

if and only if for $k = 1, \ldots, K$:

$$Y(\mathbf{a}^\dagger) \perp\!\!\!\perp I(A_k(\mathbf{a}^\dagger) = a^\dagger_k) \mid \overline{\mathbb{L}}_k(\mathbf{a}^\dagger), \overline{\mathbb{A}}_{k-1}(\mathbf{a}^\dagger) = \bar{\mathbf{a}}^\dagger_{k-1}. \quad (45)$$

We first prove the following:

Lemma 52. *If $P(\mathbb{V}(\mathbf{a}^\dagger))$ factorizes according to $\mathcal{G}(\mathbf{a}^\dagger)$ then for all $m = 1, \ldots, K$,*

$$Y(\mathbf{a}^\dagger) \perp\!\!\!\perp \overline{\mathbb{A}}_m(\mathbf{a}^\dagger) \mid \overline{\mathbb{L}}_m(\mathbf{a}^\dagger),\ \overline{\mathbb{H}}_m(\mathbf{a}^\dagger). \quad (112)$$

Proof: Suppose for a contradiction that there is a path π d-connecting some node $A'(\mathbf{a}'^{\dagger}_{A'})$ in $\overline{\mathbb{A}}_m(\mathbf{a}^\dagger)$ to $Y(\mathbf{a}^\dagger)$ given $\overline{\mathbb{L}}_m(\mathbf{a}^\dagger) \cup \overline{\mathbb{H}}_m(\mathbf{a}^\dagger) \cup \mathbf{a}^\dagger$ in $\mathcal{G}(\mathbf{a}^\dagger)$. By construction, since $\overline{\mathbf{A}}_m \subseteq \mathbf{A}$ all the nodes in $\overline{\mathbb{A}}_m(\mathbf{a}^\dagger)$ are the random half of nodes that have been split. Hence the first node after $A'(\mathbf{a}'^{\dagger}_{A'})$ on π is a parent of $A'(\mathbf{a}'^{\dagger}_{A'})$ in $\mathcal{G}(\mathbf{a}^\dagger)$, and is hence a non-collider on π. But since every parent of $A'(\mathbf{a}'^{\dagger}_{A'})$ is in $\overline{\mathbb{L}}_m(\mathbf{a}^\dagger) \cup \overline{\mathbb{H}}_m(\mathbf{a}^\dagger) \cup \mathbf{a}^\dagger$ this is a contradiction. □

Lemma 53. *If $P(\mathbf{V})$ and $P(\mathbb{V}(\mathbf{a}^\dagger))$ obey modularity (30) and factorize according to $\mathcal{G}$ and $\mathcal{G}(\mathbf{a}^\dagger)$ respectively, then:*

(i) *for $m = 1, \ldots, K$:*

$$P(Y(\mathbf{a}^\dagger)=y \mid \overline{\mathbb{H}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{h}}_m, \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{l}}_m, \overline{\mathbb{A}}_{m-1}(\mathbf{a}^\dagger)=\overline{\mathbf{a}}^\dagger_{m-1})$$
$$= P(Y(\mathbf{a}^\dagger)=y \mid \overline{\mathbb{H}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{h}}_m, \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{l}}_m, \overline{\mathbb{A}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{a}}^\dagger_m)$$
$$= b^{\text{full}}_{\mathbf{a}^\dagger}(y \mid \overline{\mathbf{l}}_m, \overline{\mathbf{h}}_m).$$

(ii) *for $m = 1, \ldots, K$:*

$$P(\overline{\mathbb{H}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{h}}_m \mid \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{l}}_m, \overline{\mathbb{A}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{a}}^\dagger_m) =$$
$$P(\overline{\mathbf{H}}_m=\overline{\mathbf{h}}_m \mid \overline{\mathbf{L}}_m=\overline{\mathbf{l}}_m, \overline{\mathbf{A}}_m=\overline{\mathbf{a}}^\dagger_m);$$

(iii) *for $m = 1, \ldots, K$:*

$$P(\overline{\mathbb{H}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{h}}_m \mid \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{l}}_m, \overline{\mathbb{A}}_{m-1}(\mathbf{a}^\dagger)=\overline{\mathbf{a}}^\dagger_{m-1}) =$$
$$P(\overline{\mathbf{H}}_m=\overline{\mathbf{h}}_m \mid \overline{\mathbf{L}}_m=\overline{\mathbf{l}}_m, \overline{\mathbf{A}}_{m-1}=\overline{\mathbf{a}}^\dagger_{m-1}).$$

Proof of (i): It follows from Corollary 18 that

$$P(Y(\mathbf{a}^\dagger)=y, \overline{\mathbb{H}}_K(\mathbf{a}^\dagger)=\overline{\mathbf{h}}_K, \overline{\mathbb{L}}_K(\mathbf{a}^\dagger)=\overline{\mathbf{l}}_K)$$
$$= p(y|\overline{\mathbf{w}}_K, \overline{\mathbf{a}}^\dagger_K) \prod_{j=1}^{K} p(w_j|\overline{\mathbf{w}}_{j-1}, \overline{\mathbf{a}}^\dagger_{j-1}).$$

Summation over $w_{m+1}, \ldots, w_K, y$ then implies that:

$$P(\overline{\mathbb{H}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{h}}_m, \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{l}}_m) = \prod_{j=1}^{m} p(w_j|\overline{\mathbf{w}}_{j-1}, \overline{\mathbf{a}}^\dagger_{j-1}).$$

Hence

$$P(Y(\mathbf{a}^\dagger)=y, \underline{\mathbb{H}}_{m+1}(\mathbf{a}^\dagger)=\underline{\mathbf{h}}_{m+1}, \underline{\mathbb{L}}_{m+1}(\mathbf{a}^\dagger)=\underline{\mathbf{l}}_K) \mid \overline{\mathbb{H}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{h}}_m, \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{l}}_m)$$
$$= p(y|\overline{\mathbf{w}}_K, \overline{\mathbf{a}}^\dagger_K) \prod_{j=m+1}^{K} p(w_j|\overline{\mathbf{w}}_{j-1}, \overline{\mathbf{a}}^\dagger_{j-1}),$$

where we use $\underline{\mathbb{H}}_{m+1}(\mathbf{a}^\dagger) = \{H_{m+1}(\mathbf{a}^\dagger), \ldots, H_{K-1}(\mathbf{a}^\dagger)\}$, likewise for $\underline{\mathbb{L}}_{m+1}(\mathbf{a}^\dagger)$. It then follows by summing over $w_{m+1}, \ldots, w_K$ that:

$$P(Y(\mathbf{a}^\dagger)=y \mid \overline{\mathbb{H}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{h}}_m, \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\bar{\mathbf{l}}_m) = b^{\text{full}}_{\mathbf{a}^\dagger}(y \mid \bar{\mathbf{l}}_m, \overline{\mathbf{h}}_m).$$

The conclusion then follows since by Lemma 52, we may add: $\overline{\mathbb{A}}_m(\mathbf{a}^\dagger)$ or $\overline{\mathbb{A}}_{m-1}(\mathbf{a}^\dagger)$ to the conditioning set. □

Proof of (ii)&(iii): We apply Lemma 50 to remove $Y, A_K, L_K, H_K, \ldots, H_{m+1}$ to obtain:

$$P(\overline{\mathbb{H}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{h}}_m, \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\bar{\mathbf{l}}_m, \overline{\mathbb{A}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{a}}^\dagger_m) = P(\overline{\mathbf{H}}_m=\overline{\mathbf{h}}_m, \overline{\mathbf{L}}_m=\bar{\mathbf{l}}_m, \overline{\mathbf{A}}_m=\overline{\mathbf{a}}^\dagger_m)$$

via Proposition 17. It then follows from Lemma 49 that

$$P(\overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\bar{\mathbf{l}}_m, \overline{\mathbb{A}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{a}}^\dagger_m) = P(\overline{\mathbf{L}}_m=\bar{\mathbf{l}}_m, \overline{\mathbf{A}}_m=\overline{\mathbf{a}}^\dagger_m),$$

from which (ii) follows. The proof of (iii) is essentially the same, except we initially also remove A_m. □

Lemma 54. *If $P(\mathbf{V})$ and $P(\mathbb{V}(\mathbf{a}^\dagger))$ obey modularity (30) and factorize according to $\mathcal{G}$ and $\mathcal{G}(\mathbf{a}^\dagger)$ respectively, then for $m = 1, \ldots, K$, and for a given $\bar{\mathbf{l}}_m$, the following are equivalent:*

(a) $Y(\mathbf{a}^\dagger) \perp\!\!\!\perp I(A_m(\mathbf{a}^\dagger) = a^\dagger_m) \mid \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\bar{\mathbf{l}}_m, \overline{\mathbb{A}}_{m-1}(\mathbf{a}^\dagger)=\overline{\mathbf{a}}^\dagger_{m-1}$,

(b) $\sum_{\overline{\mathbf{h}}_m} b^{\text{full}}_{\mathbf{a}^\dagger}(y \mid \overline{\mathbf{h}}_m, \bar{\mathbf{l}}_m) p(\overline{\mathbf{h}}_m \mid \bar{\mathbf{l}}_m, \overline{\mathbf{a}}^\dagger_{m-1}) = \sum_{\overline{\mathbf{h}}_m} b^{\text{full}}_{\mathbf{a}^\dagger}(y \mid \overline{\mathbf{h}}_m, \bar{\mathbf{l}}_m) p(\overline{\mathbf{h}}_m \mid \bar{\mathbf{l}}_m, \overline{\mathbf{a}}^\dagger_m)$. (113)

Proof: By Lemma 53(i), (ii) we have:

$$\begin{aligned}
&P(Y(\mathbf{a}^\dagger)=y \mid \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\bar{\mathbf{l}}_m, \overline{\mathbb{A}}_{m-1}(\mathbf{a}^\dagger)=\overline{\mathbf{a}}^\dagger_{m-1}) \\
&= \sum_{\overline{\mathbf{h}}_m} P(Y(\mathbf{a}^\dagger)=y \mid \overline{\mathbb{H}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{h}}_m, \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\bar{\mathbf{l}}_m, \overline{\mathbb{A}}_{m-1}(\mathbf{a}^\dagger)=\overline{\mathbf{a}}^\dagger_{m-1}) \\
&\qquad \times P(\mathbb{H}_m(\mathbf{a}^\dagger)=\overline{\mathbf{h}}_m \mid \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\bar{\mathbf{l}}_m, \overline{\mathbb{A}}_{m-1}(\mathbf{a}^\dagger)=\overline{\mathbf{a}}^\dagger_{m-1}) \\
&= \sum_{\overline{\mathbf{h}}_m} b^{\text{full}}_{\mathbf{a}^\dagger}(y \mid \overline{\mathbf{h}}_m, \bar{\mathbf{l}}_m) p(\overline{\mathbf{h}}_m \mid \bar{\mathbf{l}}_m, \overline{\mathbf{a}}^\dagger_{m-1}).
\end{aligned}$$

Similarly, we obtain from Lemma 53(i), (iii):

$$\begin{aligned}
&P(Y(\mathbf{a}^\dagger)=y \mid \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\bar{\mathbf{l}}_m, \overline{\mathbb{A}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{a}}^\dagger_m) \\
&= \sum_{\overline{\mathbf{h}}_m} P(Y(\mathbf{a}^\dagger)=y \mid \overline{\mathbb{H}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{h}}_m, \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\bar{\mathbf{l}}_m, \overline{\mathbb{A}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{a}}^\dagger_m) \\
&\qquad \times P(\mathbb{H}_m(\mathbf{a}^\dagger)=\overline{\mathbf{h}}_m \mid \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\bar{\mathbf{l}}_m, \overline{\mathbb{A}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{a}}^\dagger_m) \\
&= \sum_{\overline{\mathbf{h}}_m} b^{\text{full}}_{\mathbf{a}^\dagger}(y \mid \overline{\mathbf{h}}_m, \bar{\mathbf{l}}_m) p(\overline{\mathbf{h}}_m \mid \bar{\mathbf{l}}_m, \overline{\mathbf{a}}^\dagger_m).
\end{aligned}$$

The conclusion then follows because the independence (a) is equivalent to:

$$P(Y(\mathbf{a}^\dagger)=y \mid \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\bar{\mathbf{l}}_m, \overline{\mathbb{A}}_{m-1}(\mathbf{a}^\dagger)=\overline{\mathbf{a}}^\dagger_{m-1}) \\ = P(Y(\mathbf{a}^\dagger)=y \mid \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\bar{\mathbf{l}}_m, \overline{\mathbb{A}}_m(\mathbf{a}^\dagger)=\overline{\mathbf{a}}^\dagger_m).$$

Note that the indicator variable $I(A_m(\mathbf{a}^\dagger) = a^\dagger_m)$ arises in the independence (a) because the last equality only holds for $A_m(\mathbf{a}^\dagger) = a^\dagger_m$, and not necessarily for other values. □

Lemma 55 (Collapse of the g-formula)**.** *Equality* (113) *holds for* $m = 1, \ldots, K$ *if and only if for* $m = 0, \ldots, K$*, and all* $\bar{\mathbf{l}}_m$*,*

$$\sum_{\overline{\mathbf{h}}_m} b^{\text{full}}_{\mathbf{a}^\dagger}(y \mid \overline{\mathbf{h}}_m, \bar{\mathbf{l}}_m) p(\overline{\mathbf{h}}_m \mid \bar{\mathbf{l}}_m, \overline{\mathbf{A}}_{m-1}=\overline{\mathbf{a}}^\dagger_{m-1}) = b_{\mathbf{a}^\dagger}(y \mid \bar{\mathbf{l}}_m). \tag{114}$$

Proof: By backwards induction on m.

Base case: $m = K$

$$\begin{aligned}\sum_{\overline{\mathbf{h}}_K} b^{\text{full}}_{\mathbf{a}^\dagger}(y \mid \overline{\mathbf{h}}_K, \bar{\mathbf{l}}_K) p(\overline{\mathbf{h}}_K \mid \bar{\mathbf{l}}_K, \overline{\mathbf{a}}^\dagger_{K-1}) &= \sum_{\overline{\mathbf{h}}_K} b^{\text{full}}_{\mathbf{a}^\dagger}(y \mid \overline{\mathbf{h}}_K, \bar{\mathbf{l}}_K) p(\overline{\mathbf{h}}_K \mid \bar{\mathbf{l}}_K, \overline{\mathbf{a}}^\dagger_K) \\ &= \sum_{\overline{\mathbf{h}}_K} p(y \mid \overline{\mathbf{h}}_K, \bar{\mathbf{l}}_K, \overline{\mathbf{a}}^\dagger_K) p(\overline{\mathbf{h}}_K \mid \bar{\mathbf{l}}_K, \overline{\mathbf{a}}^\dagger_K) \\ &= p(y \mid \bar{\mathbf{l}}_K, \overline{\mathbf{a}}^\dagger_K) \;=\; b_{\mathbf{a}^\dagger}(y \mid \bar{\mathbf{l}}_K).\end{aligned}$$

Here the first equality is by (113), the second by the definition of $b^{\text{full}}_{\mathbf{a}^\dagger}$, and the third follows from the laws of probability. The conclusion follows from the definition of $b_{\mathbf{a}^\dagger}$.

Inductive case: Suppose true for $m' > m$.

$$\begin{aligned}&\sum_{\overline{\mathbf{h}}_m} b^{\text{full}}_{\mathbf{a}^\dagger}(y \mid \overline{\mathbf{h}}_m, \bar{\mathbf{l}}_m) p(\overline{\mathbf{h}}_m \mid \bar{\mathbf{l}}_m, \overline{\mathbf{a}}^\dagger_{m-1}) \\ &\quad= \sum_{\overline{\mathbf{h}}_m} b^{\text{full}}_{\mathbf{a}^\dagger}(y \mid \overline{\mathbf{h}}_m, \bar{\mathbf{l}}_m) p(\overline{\mathbf{h}}_m \mid \bar{\mathbf{l}}_m, \overline{\mathbf{a}}^\dagger_m) \\ &\quad= \sum_{\overline{\mathbf{h}}_m} \sum_{h_{m+1}, l_{m+1}} b^{\text{full}}_{\mathbf{a}^\dagger}(y \mid \bar{\mathbf{l}}_{m+1}, \overline{\mathbf{h}}_{m+1}) p(h_{m+1}, l_{m+1} \mid \overline{\mathbf{w}}_m, \mathbf{a}^\dagger_m) p(\overline{\mathbf{h}}_m \mid \bar{\mathbf{l}}_m, \mathbf{a}^\dagger_m) \\ &\quad= \sum_{l_{m+1}} \left(\sum_{\overline{\mathbf{h}}_{m+1}} b^{\text{full}}_{\mathbf{a}^\dagger}(y \mid \bar{\mathbf{l}}_{m+1}, \overline{\mathbf{h}}_{m+1}) p(\overline{\mathbf{h}}_{m+1} \mid \bar{\mathbf{l}}_m, \mathbf{a}^\dagger_m) \right) p(l_{m+1} \mid \bar{\mathbf{l}}_m, \mathbf{a}^\dagger_m) \\ &\quad= \sum_{l_{m+1}} b_{\mathbf{a}^\dagger}(y \mid \bar{\mathbf{l}}_{m+1}) p(l_{m+1} \mid \bar{\mathbf{l}}_m, \mathbf{a}^\dagger_m) \\ &\quad= b_{\mathbf{a}^\dagger}(y \mid \bar{\mathbf{l}}_m).\end{aligned}$$

Here: the first equality is by (113); the second follows from Lemma 51; the third is algebra; the fourth follows from the induction hypothesis; the fifth is again by Lemma 51. □

Proof of Theorem 22: It follows from Lemma 54 that under the conditions of the Theorem the counterfactual independence (45) is equivalent to (113) for $m = 1, \ldots, K$. It then follows from Lemma 55 that (44) holds form $m = 1, \ldots, K$. The result for $m = 0$ follows from (44) with $m = 1$ since:

$$b_{\mathbf{a}^\dagger}(y) = \sum_{l_1} b_{\mathbf{a}^\dagger}(y \mid l_1) p(l_1)$$

and

$$P(Y(\mathbf{a}^\dagger) = y) = \sum_{l_1} P(Y(\mathbf{a}^\dagger) = y \mid L_1 = l_1) p(l_1).$$

□

Proof based on consistency

In this section we give an alternative proof of Theorem 22 under the additional assumption of consistency. Our purpose in so doing is to show the utility of the assumption of consistency as a proof device.

Given $\mathbf{O} = (Y, \overline{\mathbf{L}}_K, \overline{\mathbf{A}}_K)$, define

$$f_{\overline{\mathbf{a}}^\dagger}(\mathbf{o}) \equiv p(y \mid \bar{\mathbf{l}}_K, \overline{\mathbf{a}}^\dagger_K) \prod_{j=0}^{K} p(l_j \mid \bar{\mathbf{l}}_{j-1}, \overline{\mathbf{a}}^\dagger_{j-1}) I(a_j = a_j^\dagger), \tag{115}$$

where $I(a_j = a_j^\dagger)$ is the indicator function.[88] Note that this is a degenerate distribution under which

$$f_{\overline{\mathbf{a}}^\dagger}(a_m^* \mid \bar{\mathbf{l}}_m, \overline{\mathbf{a}}_{m-1}) = I(a_m^* = a_m^\dagger). \tag{116}$$

In the statement of Theorem 22 and proof earlier in this section we went to some lengths to show that the conclusions followed directly from the properties of factorization and modularity alone. An alternative approach is to assume consistency:

Theorem 56. *Under consistency*

$$\begin{aligned} &Y(\mathbf{a}^\dagger) \perp\!\!\!\perp I(A_m = a_m^\dagger) \mid \overline{\mathbf{L}}_m, \overline{\mathbf{A}}_{m-1} = \overline{\mathbf{a}}^\dagger_{m-1} \quad \textit{for } m = 1, \ldots, K, \\ \Leftrightarrow \quad &P(Y(\mathbf{a}^\dagger) = y \mid \overline{\mathbf{L}}_m = \bar{\mathbf{l}}_m, \overline{\mathbf{A}}_{m-1} = \overline{\mathbf{a}}^\dagger_{m-1}) = f_{\overline{\mathbf{a}}^\dagger}(y \mid \bar{\mathbf{l}}_m, \overline{\mathbf{a}}^\dagger_{m-1}), \\ &\qquad \textit{for } m = 0, \ldots, K. \end{aligned} \tag{117}$$

[88] The inclusion of the indicator functions is required here in order for (115) to specify a unique joint over $\mathbf{O}$.

Note that since whenever factorization and modularity hold it is always possible to construct a (pseudo)-counterfactual model under which consistency also holds, an argument similar to that used in the proof of Theorem 56 may be used to establish Theorem 22 via a 'coupling' argument.

Proof: We first prove the equivalence for $m = K$.
($\Rightarrow$) By consistency:

$$P(Y(\mathbf{a}^\dagger_K){=}y \mid \overline{\mathbf{L}}_K{=}\bar{\mathbf{l}}_K, \overline{\mathbf{A}}_K = \bar{\mathbf{a}}^\dagger_K) = f_{\bar{\mathbf{a}}^\dagger}(y \mid \bar{\mathbf{l}}_K, \bar{\mathbf{a}}^\dagger_K).$$

Further $P(Y(\mathbf{a}^\dagger) = y \mid \bar{\mathbf{l}}_K, \bar{\mathbf{a}}^\dagger_K) = P(Y(\mathbf{a}^\dagger) = y \mid \bar{\mathbf{l}}_K, \bar{\mathbf{a}}^\dagger_{K-1})$ by assumption and $f_{\bar{\mathbf{a}}^\dagger}(y \mid \bar{\mathbf{l}}_K, \bar{\mathbf{a}}^\dagger_K) = f_{\bar{\mathbf{a}}^\dagger}(y \mid \bar{\mathbf{l}}_K, \bar{\mathbf{a}}^\dagger_{K-1})$ by (116). Thus the right hand side of (117) holds for K.

($\Leftarrow$) Conversely,

$$\begin{aligned} P(Y(\mathbf{a}^\dagger){=}y \mid \overline{\mathbf{L}}_K{=}\bar{\mathbf{l}}_K, \overline{\mathbf{A}}_{K-1}{=}\bar{\mathbf{a}}^\dagger_{K-1}) &= f_{\bar{\mathbf{a}}^\dagger}(y \mid \bar{\mathbf{l}}_K, \bar{\mathbf{a}}^\dagger_{K-1}) \\ &= f_{\bar{\mathbf{a}}^\dagger}(y \mid \bar{\mathbf{l}}_K, \bar{\mathbf{a}}^\dagger_K) \\ &= P(Y(\mathbf{a}^\dagger){=}y \mid \overline{\mathbf{L}}_K{=}\bar{\mathbf{l}}_K, \overline{\mathbf{A}}_K{=}\bar{\mathbf{a}}^\dagger_K), \end{aligned}$$

where the first equality is by assumption, the second by (116), and the last by consistency.

Suppose the equivalence is true for $m' > m \geq 1$, we prove it for m.

($\Rightarrow$) By the inductive assumption,

$$P(Y(\mathbf{a}^\dagger){=}y \mid \overline{\mathbf{L}}_{m+1}{=}\bar{\mathbf{l}}_{m+1}, \overline{\mathbf{A}}_m = \bar{\mathbf{a}}^\dagger_m) = f_{\bar{\mathbf{a}}^\dagger}(y \mid \bar{\mathbf{l}}_{m+1}, \bar{\mathbf{a}}^\dagger_m).$$

Further,

$$P(Y(\mathbf{a}^\dagger){=}y \mid \overline{\mathbf{L}}_m{=}\bar{\mathbf{l}}_m, \overline{\mathbf{A}}_m = \bar{\mathbf{a}}^\dagger_m) = f_{\bar{\mathbf{a}}^\dagger}(y \mid \bar{\mathbf{l}}_m, \bar{\mathbf{a}}^\dagger_m) \;=\; f_{\bar{\mathbf{a}}^\dagger}(y \mid \bar{\mathbf{l}}_m, \bar{\mathbf{a}}^\dagger_{m-1}),$$

by integration over a common law $p(l_{m+1}|\bar{\mathbf{l}}_m, \bar{\mathbf{a}}_m) = f_{\bar{\mathbf{a}}^\dagger}(l_{m+1}|\bar{\mathbf{l}}_m, \bar{\mathbf{a}}_m)$, and (116). But

$$P(Y(\mathbf{a}^\dagger){=}y \mid \overline{\mathbf{L}}_m{=}\bar{\mathbf{l}}_m, \overline{\mathbf{A}}_m = \bar{\mathbf{a}}^\dagger_m) \;=\; P(Y(\mathbf{a}^\dagger){=}y \mid \overline{\mathbf{L}}_m{=}\bar{\mathbf{l}}_m, \overline{\mathbf{A}}_{m-1} = \bar{\mathbf{a}}^\dagger_{m-1})$$

by the left hand side of (117).
($\Leftarrow$) We have:

$$\begin{aligned} &P(Y(\mathbf{a}^\dagger){=}y \mid \overline{\mathbf{L}}_{m+1}{=}\bar{\mathbf{l}}_{m+1}, \overline{\mathbf{A}}_{m+1} = \bar{\mathbf{a}}^\dagger_{m+1}) \\ &\quad = P(Y(\mathbf{a}^\dagger){=}y \mid \overline{\mathbf{L}}_{m+1}{=}\bar{\mathbf{l}}_{m+1}, \overline{\mathbf{A}}_m = \bar{\mathbf{a}}^\dagger_m) \;=\; f_{\bar{\mathbf{a}}^\dagger}(y \mid \bar{\mathbf{l}}_{m+1}, \bar{\mathbf{a}}^\dagger_m), \end{aligned}$$

by the inductive assumption and the right hand side of (117), for $m+1$. Thus

$$\begin{aligned} P(Y(\mathbf{a}^\dagger)=y \mid \overline{\mathbf{L}}_m=\bar{\mathbf{l}}_m, \overline{\mathbf{A}}_m=\bar{\mathbf{a}}^\dagger_m) &= f_{\bar{\mathbf{a}}^\dagger}(y \mid \bar{\mathbf{l}}_m, \bar{\mathbf{a}}^\dagger_m) \\ &= f_{\bar{\mathbf{a}}^\dagger}(y \mid \bar{\mathbf{l}}_m, \bar{\mathbf{a}}^\dagger_{m-1}) \\ &= P(Y(\mathbf{a}^\dagger)=y \mid \overline{\mathbf{L}}_m=\bar{\mathbf{l}}_m, \overline{\mathbf{A}}_{m-1}=\bar{\mathbf{a}}^\dagger_{m-1}), \end{aligned}$$

where the first is by integration over the common law $p(l_{m+1}|\bar{\mathbf{l}}_m, \bar{\mathbf{a}}_m) = f_{\bar{\mathbf{a}}^\dagger}(\bar{\mathbf{l}}_m, \bar{\mathbf{a}}_m)$, the second by (116), and the last by the right hand side of (117). Hence the independence on the left hand side of (117) holds for m.

Finally, note that the right hand side of (117) with $m=1$ implies the equality with $m=0$, but integration over $p(l_1) = f_{\bar{\mathbf{a}}^\dagger}(l_1)$. □

B.3 Proofs of identification results for dynamic regimes

As in the previous section, let $\mathbf{O} = \mathbf{A} \cup \mathbf{L} \cup \{Y\}$ be the set of observed variables, $\mathbf{H}$ the unobserved, $\mathbf{V} = \mathbf{O} \cup \mathbf{H}$ and $W_t = (H_t, L_t)$, be the observed and unobserved covariates at stage t. Given $\mathbf{Z}_m \subseteq \mathbf{O} \setminus (\overline{\mathbf{L}}_m \cup \overline{\mathbf{A}}_{m-1})$, let

$$b^{\text{full}}_{\mathbf{a}^\dagger}(\mathbf{z}_m \mid \overline{\mathbf{h}}_m, \bar{\mathbf{l}}_m) \equiv \sum_{\underline{\mathbf{h}}_{m+1}, \mathbf{s}_m} p(y|\overline{\mathbf{w}}_K, \bar{\mathbf{a}}^\dagger_K) \prod_{j=1}^{K} p(w_j, a_j|\overline{\mathbf{w}}_{j-1}, \bar{\mathbf{a}}^\dagger_{j-1}), \quad (118)$$

$$b_{\mathbf{a}^\dagger}(\mathbf{z}_m \mid \bar{\mathbf{l}}_m) \equiv \sum_{\mathbf{s}_m} p(y|\bar{\mathbf{l}}_K, \bar{\mathbf{a}}^\dagger_K) \prod_{j=1}^{K} p(l_j, a_j|\bar{\mathbf{l}}_{j-1}, \bar{\mathbf{a}}^\dagger_{j-1}), \quad (119)$$

where $\mathbf{S}_m = \mathbf{O} \setminus \mathbf{Z}_m$.

We first state without proof the following simple multivariate extensions of results in the previous section.

Lemma 57. *Suppose $P(\mathbf{V})$ and $P(\mathbb{V}(\mathbf{a}^\dagger))$ obey modularity (30) and factorize according to $\mathcal{G}$ and $\mathcal{G}(\mathbf{a}^\dagger)$ respectively. For $j = 1, \ldots, K$, let:*

$$\mathbb{Z}_j(\mathbf{a}^\dagger) \subseteq \mathbb{O}(\mathbf{a}^\dagger) \setminus \left(\overline{\mathbb{L}}_j(\mathbf{a}^\dagger) \cup \overline{\mathbb{A}}_j(\mathbf{a}^\dagger)\right). \quad (120)$$

Then for $m = 1, \ldots, K$, and for a given $\bar{\mathbf{l}}_m$, the following are equivalent:

(a) $\mathbb{Z}_m(\mathbf{a}^\dagger) \perp\!\!\!\perp I(A_m(\mathbf{a}^\dagger) = a^\dagger_m) \mid \overline{\mathbb{L}}_m(\mathbf{a}^\dagger)=\bar{\mathbf{l}}_m, \overline{\mathbb{A}}_{m-1}(\mathbf{a}^\dagger)=\bar{\mathbf{a}}^\dagger_{m-1}$,

(b) $$\sum_{\overline{\mathbf{h}}_m} b^{\text{full}}_{\mathbf{a}^\dagger}(\mathbf{z}_m \mid \overline{\mathbf{h}}_m, \bar{\mathbf{l}}_m) p(\overline{\mathbf{h}}_m \mid \bar{\mathbf{l}}_m, \bar{\mathbf{a}}^\dagger_{m-1}) = \sum_{\overline{\mathbf{h}}_m} b^{\text{full}}_{\mathbf{a}^\dagger}(\mathbf{z}_m \mid \overline{\mathbf{h}}_m, \bar{\mathbf{l}}_m) p(\overline{\mathbf{h}}_m \mid \bar{\mathbf{l}}_m, \bar{\mathbf{a}}^\dagger_m). \quad (121)$$

Lemma 58 (Multivariate g-formula collapse)**.** *Equality* (121) *holds for* $m = 1, \ldots, K$ *if and only if for* $m = 0, \ldots, K$*, and all* $\bar{\mathbf{l}}_m$*,*

$$\sum_{\overline{\mathbf{h}}_m} b^{\text{full}}_{\mathbf{a}^\dagger}(\mathbf{z}_m \mid \overline{\mathbf{h}}_m, \bar{\mathbf{l}}_m) p(\overline{\mathbf{h}}_m \mid \bar{\mathbf{l}}_m, \overline{\mathbf{a}}^\dagger_{m-1}) = b_{\mathbf{a}^\dagger}(\mathbf{z}_m \mid \bar{\mathbf{l}}_m). \tag{122}$$

We define the *full dynamic extended g-formula* to be:

$$f^{g,\text{full}}(\mathbf{z}) \equiv \sum_{\mathbf{a}^+, \mathbf{h}, \mathbf{s}} p(y|\overline{\mathbf{w}}_K, \overline{\mathbf{a}}^+_K) \prod_{j=1}^{K} p(w_j, a_j|\overline{\mathbf{w}}_{j-1}, \overline{\mathbf{a}}^+_{j-1}) \prod_{t=1}^{K} q^g_t(a^+_t \mid \mathbf{pa}^+_t). \tag{123}$$

where: $\mathbf{S} = \mathbf{O} \setminus \mathbf{Z}$; $q^g_t \in \mathfrak{Q}(g)$ is the density corresponding to g_t; $\mathbf{pa}^+_t$ is the subvector of $(\overline{\mathbf{a}}^+_{t-1}, \bar{\mathbf{l}}_t, \bar{\mathbf{a}}_t)$ of inputs to g_t. Note that by construction there are no unobserved variables in $\mathbf{PA}^+_t$.

Recall that the dynamic extended g-formula for the observed data is:

$$f^g(\mathbf{z}) \equiv \sum_{\mathbf{a}^+, \mathbf{s}} p(y|\bar{\mathbf{l}}_K, \overline{\mathbf{a}}^+_K) \prod_{j=1}^{K} p(l_j, a_j|\bar{\mathbf{l}}_{j-1}, \overline{\mathbf{a}}^+_{j-1}) \prod_{t=1}^{K} q^g_t(a^+_t \mid \mathbf{pa}^+_t). \tag{64}$$

Lemma 59 (Dynamic g-formula collapse)**.** *If* (121) *holds for* $m = 1, \ldots, K$ *then*

$$f^g(\mathbf{z}) = f^{g,\text{full}}(\mathbf{z}). \tag{124}$$

Proof:

$$\begin{aligned}
f^{g,\text{full}}(\mathbf{z}) &\equiv \sum_{\mathbf{a}^+, \mathbf{h}, \mathbf{s}} p(y|\overline{\mathbf{w}}_K, \overline{\mathbf{a}}^+_K) \prod_{j=1}^{K} p(w_j, a_j|\overline{\mathbf{w}}_{j-1}, \overline{\mathbf{a}}^+_{j-1}) \prod_{t=1}^{K} q^g_t(a^+_t \mid \mathbf{pa}^+_t) \\
&= \sum_{\mathbf{a}^+, \mathbf{s}} \prod_{t=1}^{K} q^g_t(a^+_t \mid \mathbf{pa}^+_t) \sum_{\mathbf{h}} p(y|\overline{\mathbf{w}}_K, \overline{\mathbf{a}}^+_K) \prod_{j=1}^{K} p(w_j, a_j|\overline{\mathbf{w}}_{j-1}, \overline{\mathbf{a}}^+_{j-1}) \\
&= \sum_{\mathbf{a}^+, \mathbf{s}} \prod_{t=1}^{K} q^g_t(a^+_t \mid \mathbf{pa}^+_t) p(y|\bar{\mathbf{l}}_K, \overline{\mathbf{a}}^+_K) \prod_{j=1}^{K} p(l_j, a_j|\bar{\mathbf{l}}_{j-1}, \overline{\mathbf{a}}^+_{j-1}),
\end{aligned}$$

where the first equality uses the fact that $\mathbf{PA}^+_t \subseteq \mathbf{O}$, since the dynamic regime does not depend on $\mathbf{H}$; the second follows from Lemma 58. □

Proof of Theorem 31: This follows directly from Lemmas 57 and 59. □

B.4 Proof of Lemma relating to perturbed dynamic regimes

We now prove Lemma 33, which shows that checking the d-separation condition (65) for identification of a dynamic regime in $\mathcal{G}(g)$, is equivalent to checking a d-separation in the dSWIG $\mathcal{G}(g_{-k})$ corresponding to a 'perturbed' dynamic regime g_{-k} (for each k).

Proof: We will define

$$\mathbb{T}(g_{-k}) \equiv \overline{\mathbb{A}}_{k-1}(g_{-k}) \cup \overline{\mathbb{L}}_k(g_{-k}) \cup \overline{\mathbb{A}}^+_{k-1}(g_{-k});$$
$$\mathbb{T}(\mathbf{a}^\dagger) \equiv \overline{\mathbb{A}}_{k-1}(\mathbf{a}^\dagger) \cup \overline{\mathbb{L}}_k(\mathbf{a}^\dagger) \cup \mathbf{a}^\dagger \text{ in } \mathcal{G}(\mathbf{a}^\dagger)$$

(i) $\Rightarrow$ (ii): Suppose for a contradiction that there is a path π d-connecting $A_k(g_{-k})$ and $Y(g_{-k})$ given $\mathbb{T}(g_{-k})$ in $\mathcal{G}(g_{-k})$, but that $A_k(\mathbf{a}^\dagger)$ is d-separated from $\mathbb{Z}_k(\mathbf{a}^\dagger)$ by $\mathbb{T}(\mathbf{a}^\dagger)$ in $\mathcal{G}(\mathbf{a}^\dagger)$.

First note that the ordering of the vertices in $\mathbb{A}(g) \cup \mathbb{A}^+(g) \cup \mathbb{L}(g)$ is a topological ordering of the observed vertices in $\mathcal{G}(g)$ and hence also of the observed vertices $\mathcal{G}(g_{-k})$. Further since $\mathbb{T}(g_{-k})$ is subset of the variables prior to $A_k(g_{-k})$, any vertex $V(g_{-k})$ on π that is in $\mathbb{A}(g_{-k}) \cup \mathbb{A}^+(g_{-k}) \cup \mathbb{L}(g_{-k})$ and is after $A_k(g_{-k})$ is not an ancestor of $\mathbb{T}(g_{-k})$. Hence $V(g_{-k})$ is a non-collider on π and the subpath of π between $V(g_{-k})$ and $Y(g_{-k})$ takes the form of a directed path from $V(g_{-k})$ to $Y(g_{-k})$.

If the sequence of vertices corresponding to π in $\mathcal{G}(\mathbf{a}^\dagger)$ does not form a path d-connecting $A_k(\mathbf{a}^\dagger)$ and $Y(\mathbf{a}^\dagger)$ in $\mathcal{G}(\mathbf{a}^\dagger)$ then either:

(a) There is a dashed edge on π that is not present in $\mathcal{G}(\mathbf{a}^\dagger)$, or

(b) There is a non-collider on π that is not in $\mathbb{T}(g_{-k})$, but the corresponding vertex in $\mathcal{G}(\mathbf{a}^\dagger)$ is in $\mathbb{T}(\mathbf{a}^\dagger)$, or

(c) There is a collider on π that is in $\text{an}_{\mathcal{G}(g_{-k})}(\mathbb{T}(g_{-k}))$ but is not in $\text{an}_{\mathcal{G}(\mathbf{a}^\dagger)}(\mathbb{T}(\mathbf{a}^\dagger))$.

Let $V^*(g_{-k})$ be the first vertex on π, starting from $A_k(g_{-k})$ at which one of (a), (b) or (c) holds. We consider each in turn.

If (a) holds then there is a dashed edge $V^*(g_{-k}) \dashrightarrow A^+_\ell(g_{-k})$ on π. By construction of $\mathcal{G}(g_{-k})$, $V^*(g_{-k})$ is not $A_k(g_{-k})$, since all such dashed edges are removed in $\mathcal{G}(g_{-k})$. Since every parent of $A^+_\ell(g_{-k})$ is observed, $V^*(g_{-k})$ is in $\mathbb{A}(g_{-k}) \cup \mathbb{A}^+(g_{-k}) \cup \mathbb{L}(g_{-k})$. But since π is d-connecting and $V^*(g_{-k})$ is a non-collider, $V^*(g_{-k}) \notin \mathbb{T}(g_{-k})$. This implies that both $V^*(g_{-k})$ and $A^+_\ell(g_{-k})$ are ordered after $A_k(g_{-k})$. It then follows that $V^*(g_{-k}), A^+_\ell(g_{-k}) \notin \text{an}_{\mathcal{G}(g_{-k})}(\mathbb{T}(g_{-k}))$. Hence $V^*(g_{-k}) = V(g_{-k})$ and π takes the form:

$$A_k(g_{-k}) \cdots V(g_{-k}) \dashrightarrow A^+_\ell(g_{-k}) \rightarrow \cdots \rightarrow Y(g_{-k}).$$

Since $V^*(g_{-k}) \in \text{an}_{\mathcal{G}(g_{-k})}(Y(g_{-k}))$, it follows that $V^*(g) \in \text{an}_{\mathcal{G}(g)}(Y(g))$. This then implies that $V^*(\mathbf{a}^\dagger) \in \mathbb{Z}_k(\mathbf{a}^\dagger)$. By the definition of $V^*(g_{-k})$, the subpath of pi between $A_k(\mathbf{a}^\dagger)$ and $V^*(\mathbf{a}^\dagger)$ d-connects in $\mathcal{G}(\mathbf{a}^\dagger)$ given $\mathbb{T}(\mathbf{a}^\dagger)$, which is a contradiction.

If (b) holds then since the only vertices that are conditioned on in $\mathbb{T}(\mathbf{a}^\dagger)$, but for which the corresponding vertex is not conditioned on in $\mathbb{T}(g_{-k})$ are fixed vertices a_ℓ in $\mathcal{G}(\mathbf{a}^\dagger)$, corresponding to a vertex $A_\ell^+(g_{-k})$ that occurs after $A_k(\mathbf{a}^\dagger)$ in the ordering. Consequently, $A_\ell^+(g_{-k}) \notin \text{an}_{\mathcal{G}(g_{-k})}(\mathbb{T}(g_{-k}))$. The path π thus takes one of the following form:

$$
\begin{aligned}
A_k(g_{-k}) \cdots \dashrightarrow A_\ell^+(g_{-k}) \rightarrow \cdots \rightarrow Y(g_{-k}),\\
A_k(g_{-k}) \cdots \leftarrow A_\ell^+(g_{-k}) \dashleftarrow \cdots \rightarrow Y(g_{-k}),\\
A_k(g_{-k}) \cdots \leftarrow A_\ell^+(g_{-k}) \rightarrow \cdots \rightarrow Y(g_{-k}).
\end{aligned}
$$

By the definition of $V^*(g)$, the first cannot arise since (a) would be true at the vertex prior to $A_k(g_{-k})$ on π. The other two possibilities also give rise to a contradiction because there cannot be a collider on π between $A_k(g_{-k})$ and $A_\ell^+(g_{-k})$, since $A_\ell^+(g_{-k}) \notin \text{an}_{\mathcal{G}(g_{-k})}(\mathbb{T}(g_{-k}))$, but this would imply $A_\ell^+(g_{-k}) \in \text{an}_{\mathcal{G}(g_{-k})}(A_k(g_{-k}))$, but this would violate the assumption that we have a topological ordering on $\mathcal{G}(g)$.

If (c) holds then first note that the vertex $V^*(g_{-k})$ in $\mathcal{G}(\mathbf{a}^\dagger)$ is not a fixed vertex (since by definition of $V^*(g_{-k})$ neither of the edges adjacent to $V^*(g_{-k})$ on π are dashed). Thus let $V^*(\mathbf{a}^\dagger)$ denote the corresponding vertex in $\mathcal{G}(\mathbf{a}^\dagger)$. Since $V^*(g_{-k}) \in \text{an}_{\mathcal{G}(g_{-k})}(\mathbb{T}(g_{-k}))$, but $V^*(\mathbf{a}^\dagger) \notin \text{an}_{\mathcal{G}(\mathbf{a}^\dagger)}(\mathbb{T}(\mathbf{a}^\dagger))$, it follows that on every directed path from $V^*(g_{-k})$ to some vertex in $\text{an}_{\mathcal{G}(g_{-k})}(\mathbb{T}(g_{-k}))$ there is at least one dashed edge. (Note that the set $\mathbb{T}(\mathbf{a}^\dagger)$ contains all the vertices corresponding to vertices in $\mathbb{T}(g_{-k})$.) Let $V^*(g_{-k}) \dashrightarrow A_\ell^+(g_{-k})$ be the first such edge on a directed path from $V^*(g_{-k})$ to a vertex in $\mathbb{T}(g_{-k})$. As in case (a), since every parent of $A_\ell^+(g_{-k})$ is observed, $V^*(g_{-k})$ is in $\mathbb{A}(g_{-k}) \cup \mathbb{A}^+(g_{-k}) \cup \mathbb{L}(g_{-k})$, but by hypothesis, $V^*(g_{-k}) \notin \mathbb{T}(g_{-k})$, so $V^*(g_{-k})$ is ordered after $A_k(g_{-k})$, but this contradicts $V^*(g_{-k})$ being on a directed path to some vertex in $\mathbb{T}(g_{-k})$.

(i) $\Leftarrow$ (ii): Suppose that there is a path d-connecting $A_k(\mathbf{a}^\dagger)$ to some vertex in $\mathbb{Z}_k(\mathbf{a}^\dagger)$ given $\mathbb{T}(\mathbf{a}^\dagger)$ in $\mathcal{G}(\mathbf{a}^\dagger)$. Let $\pi^\dagger$ be a shortest such d-connecting path, with endpoint $Q(\mathbf{a}^\dagger) \in \mathbb{Z}_k(\mathbf{a}^\dagger)$.

By definition of $\mathbb{Z}_k(g)$, the vertex

$$
Q(g) \in \big(\mathbb{A}(g) \cup \mathbb{A}^+(g) \cup \mathbb{L}(g)\big) \cap \text{an}_{\mathcal{G}(g)}(Y(g))
$$

but is ordered after $A_k(g)$. Consequently, if $Q(g_{-k})$ is the corresponding vertex in $\mathcal{G}(g_{-k})$ then $Q(g_{-k}) \notin \text{an}_{\mathcal{G}(g_{-k})}(\mathbb{T}(g_{-k}))$ and $Q(g_{-k}) \in \text{an}_{\mathcal{G}(g_{-k})}(Y(g_{-k}))$. Every edge in $\mathcal{G}(\mathbf{a}^\dagger)$ is also present in $\mathcal{G}(g_{-k})$. Thus let $\pi^\dagger(g)$ be the path corresponding to $\pi^\dagger$ in $\mathcal{G}(g_{-k})$. By the construction of $\mathcal{G}(\mathbf{a}^\dagger)$, fixed vertices have no parents, so there is no collider on $\pi^\dagger$ that is an ancestor of a fixed vertex in $\mathcal{G}(\mathbf{a}^\dagger)$. Thus every collider on $\pi^\dagger(g)$ is an ancestor of a vertex in $\mathbb{T}(g_{-k})$. It follows that by concatenating $\pi^\dagger(g)$ and a directed path from $Q(g_{-k})$ to $Y(g_{-k}$ we form a path that d-connects $A_k(g_{-k})$ and $Y(g_{-k})$ given $\mathbb{T}(g_{-k})$. □

C Relationship to Fully Randomized Causally Interpretable Structured Tree Graphs

In this section we consider the relationship of our results to the counterfactual models of Robins (1986, 1987) through the following development. We first review the assumptions relating to counterfactual variables that are used in Robins (1986, 1987).[89]

We are given an ordered set $\mathbf{V}$ of random variables and a set of K treatments $\mathbf{A} = \{A_1, \ldots, A_K\} \subseteq \mathbf{V}$ with A_m preceeding A_{m+1}, in the ordering. In Robins (1986) the ordering was taken to be temporal except for the discussion of the healthy worker survivor effect; see Section 11 of (Robins, 1986) and §4.2.4.

We shall denote by L_m the subset of $\mathbf{V}$ between A_{m-1} and A_m in the ordering and let $D_m = \{L_m, A_m\}$. Note that A_m and L_m are disjoint and that $\mathbf{D} \equiv \overline{\mathbf{D}}_{K+1}$ is all of $\mathbf{V}$.

We assume the counterfactual $\mathbb{V}(\mathbf{r})$ is well defined for any assignment $\mathbf{r}$ to a subset $\mathbf{R} \subset \mathbf{A}$ defined as follows:

Definition 60 (Counterfactual Existence Assumption for $(\mathbf{A}, \mathbf{V})$)**.**

(i) *All one-treatment-ahead counterfactuals $D_m(\overline{\mathbf{a}}_{m-1})$ exist for any setting of $\overline{\mathbf{a}}_{m-1}$ in the state space $\overline{\mathfrak{A}}_{m-1}$ of $\overline{\mathbf{A}}_{m-1}$, for $m = 1, \ldots, K+1$.*[90]

[89] Robins' counterfactual models were defined in terms of so-called structured tree graphs rather than in terms of a set of variables $\mathbf{V}$; the latter being the leading special case of the former. For consistency with the remainder of this paper this appendix treats only this special case.

[90] Here as elsewhere, $\overline{\mathbf{A}}_s = (A_1, \ldots, A_s)$ and $\overline{\mathbf{A}}_0$ denotes an empty vector.

(ii) *Both the factual variables* D_m *and the counterfactuals* $D_m(\mathbf{r})$ *for any* $\mathbf{R} \subset \mathbf{A}$ *exist and are obtained recursively from the one-treatment-ahead counterfactuals* $D_j(\overline{\mathbf{a}}_{j-1})$*, for* $j \leq K+1$*, as follows:*

$$D_m(\tilde{\mathbf{r}}) = D_m\left(\tilde{\mathbf{r}}_{\overline{\mathbf{A}}_{m-1} \cap \mathbf{R}},\ \mathbf{B}(\tilde{\mathbf{r}})\right), \tag{125}$$

where $\mathbf{B}(\tilde{\mathbf{r}}) \equiv \{A_j(\tilde{\mathbf{r}}) \mid A_j \in \overline{\mathbf{A}}_{m-1}, A_j \notin \mathbf{R}\}$ *are the counterfactual variables in* $\mathbf{A}$ *that are prior to* D_m*, but not in* $\mathbf{R}$*, evaluated under the intervention* $\tilde{\mathbf{r}}$*.*

Assumption (ii) combines the assumptions of consistency, recursive substitution, and the assumption that earlier variables (with respect to the ordering) are not caused by later treatments. For example, $D_3 = D_3(A_1, A_2(A_1))$ and $D_3(a_1) = D_3(a_1, A_2(a_1)), D_3(a_1, a_s) = D_3(a_1)$ for $s > 2$.

C.1 Definition of the CISTG model associated with $(\mathbf{A}, \mathbf{V})$

Suppose the counterfactual existence assumption holds for $(\mathbf{A}, \mathbf{V})$. The *causally interpreted structured tree graph model with interventions* $\mathbf{A}$ *and variables* $\mathbf{V}$, where $\mathbf{A} \subseteq \mathbf{V}$, denoted CISTG$(\mathbf{A}, \mathbf{V})$, is the set of all joint distributions over these factuals and counterfactuals.

If there is no strict superset $\mathbf{A}_{\text{sup}}$ of $\mathbf{A}$ such that the counterfactual existence assumption holds for $(\mathbf{A}_{\text{sup}}, \mathbf{V})$, we say CISTG$(\mathbf{A}, \mathbf{V})$ is a *finest* CISTG. Intuitively a CISTG is 'finest' if the set of treatments $\mathbf{A}$ cannot be enlarged without encountering counterfactuals judged to be ill-defined. If all variables in $\mathbf{V}$ have associated well defined counterfactuals then CISTG$(\mathbf{V}, \mathbf{V})$ is well defined and is the unique finest CISTG.

Given a CISTG, it follows from the existence assumption above that we may always consider a smaller set of interventions and variables, without fear that any of the counterfactuals are not well-defined:

Lemma 61. *If the counterfactual existence assumption holds for* $(\mathbf{A}, \mathbf{V})$*, it also holds for any* $(\mathbf{A}_{\text{sub}}, \mathbf{V}_{\text{sub}})$ *such that* $\mathbf{V}_{\text{sub}} \subset \mathbf{V}$ *and* $\mathbf{A}_{\text{sub}} \subset \mathbf{V}_{\text{sub}} \cap \mathbf{A}$*; and then it holds for* $(\mathbf{A}_{\text{sub}}, \mathbf{V}_{\text{sub}})$*.*

Proof: Follows immediately from recursive substitution, (Def. 60(ii)). □

In the following development, Y_m is a given subvector of responses $L_m = (Y_m, W_m)$ where Y_m and W_m are disjoint.

C.2 Randomized CISTG models

We now define three classes of CISTG models: 'fully randomized', 'randomized' and 'randomized with respect to $\mathbf{Y}$'. Here, the models differ with respect to which intervention distributions are identified. In the fully randomized model, the effect of $\mathbf{A}$ on every variable in $\mathbf{V}$ is identified both unconditionally as well as conditional on treatments and covariates that are earlier in the ordering; this includes the effect that earlier interventions A_j have on the value of later treatments A_k that a subject would receive were it not for intervention on A_k. In the randomized model, the effect of $\mathbf{A}$ on every variable in $\mathbf{V} \setminus \mathbf{A}$ is identified (again conditional on earlier variables), but the effect of earlier treatments on the values that later treatment variables would take are not identified. Finally, in models randomized with respect to $\mathbf{Y}$, it is only the effect of $\mathbf{A}$ on $\mathbf{Y}$ (conditional on earlier variables) that is identified.

Definition 62. *A submodel of a* CISTG$(\mathbf{A}, \mathbf{V})$ *model is* fully randomized (FRCISTG $(\mathbf{A}, \mathbf{V})$)*;* randomized (RCISTG $(\mathbf{A}, \mathbf{V})$)*;* randomized with respect to $\mathbf{Y}$ (RCISTG: $\mathbf{Y}(\mathbf{A}, \mathbf{V})$) *if each distribution in the submodel obeys the following independences:*

$$\{Z_{m+1}(\bar{\mathbf{a}}_m^\dagger), \ldots, Z_{K+1}(\bar{\mathbf{a}}_K^\dagger)\} \perp\!\!\!\perp A_m \mid \overline{\mathbf{L}}_m = \bar{\mathbf{l}}_m, \overline{\mathbf{A}}_{m-1} = \bar{\mathbf{a}}_{m-1}^\dagger, \quad \text{for all } m \leq K, \bar{\mathbf{a}}_K^\dagger, \bar{\mathbf{l}}_m, \tag{126}$$

for $\mathbf{Z} = \mathbf{D}$*,* $\mathbf{Z} = \mathbf{L}$*, or* $\mathbf{Z} = \mathbf{Y}$*, respectively.*

Note that by (ii), (126) is equivalent to:

$$\{Z_{m+1}(\bar{\mathbf{a}}_m^\dagger), \ldots, Z_{K+1}(\bar{\mathbf{a}}_K^\dagger)\} \perp\!\!\!\perp A_m(\bar{\mathbf{a}}_{m-1}^\dagger) \mid \overline{\mathbb{L}}_m(\bar{\mathbf{a}}_{m-1}^\dagger) = \bar{\mathbf{l}}_m, \overline{\mathbb{A}}_{m-1}(\bar{\mathbf{a}}_{m-2}^\dagger) = \bar{\mathbf{a}}_{m-1}^\dagger, \quad \text{for all } m \leq K, \bar{\mathbf{a}}_K^\dagger, \bar{\mathbf{l}}_m, \tag{127}$$

where $\overline{\mathbb{L}}_m(\bar{\mathbf{a}}_{m-1}^\dagger) = (L_1, L_2(a_1^\dagger), \ldots, L_{m-1}(\bar{\mathbf{a}}_{m-2}^\dagger), L_m(\bar{\mathbf{a}}_{m-1}^\dagger))$.

Observe that in Definition 62 the randomized CISTG models are all defined relative to a fixed ordering of the variables $\mathbf{V}$. In this section the ordering will be fixed implicitly via variable names and indices to be $\langle L_1, A_1, \ldots, L_{K+1} \rangle$. However, in section §C.3, we will consider the same set of variables under different orderings, and will thus index each model via the specific ordering thus: FRCISTG$_\prec(\mathbf{A}, \mathbf{V})$.

The following is obvious:

Proposition 63. *The FRCISTG* $(\mathbf{A}, \mathbf{V})$ *model is contained in the RCISTG* $(\mathbf{A}, \mathbf{V})$ *model; further the RCISTG* $(\mathbf{A}, \mathbf{V})$ *model is contained in the RCISTG:*$\mathbf{Y}$ $(\mathbf{A}, \mathbf{V})$ *model.*

Given an FRCISTG$(\mathbf{A}, \mathbf{V})$, we say an FRCISTG $(\mathbf{A}, \mathbf{V})$ is a *finest* FRCISTG$(\mathbf{A}, \mathbf{V})$, (i.e. an FFRCISTG$(\mathbf{A}, \mathbf{V})$) if the associated CISTG$(\mathbf{A}, \mathbf{V})$ is a finest CISTG.

Consider a DAG $\mathcal{G}$ and a CISTG $(\mathbf{V}, \mathbf{V})$ for which the CISTG ordering is a topological with respect to $\mathcal{G}$. Then the FFRCISTG model associated with $\mathcal{G}$ introduced in §3 is precisely the FFRCISTG $(\mathbf{V}, \mathbf{V})$ model if $\mathcal{G}$ is complete; otherwise the FFRCISTG is a submodel. This follows from the fact that Robins and Richardson (2011) prove that (126) with $\mathbf{Z} = \mathbf{D}$ is equivalent to the independencies (17) that defined the FFRCISTG model in §3.2.

Lemma 64. *Suppose we are given a distribution* $P(\mathbb{V})$ *in the* CISTG$(\mathbf{A}, \mathbf{V})$ *model and a subset* $\mathbf{A}_{\text{sub}}$ *of* $\mathbf{A}$*. Let* $P(\mathbb{V}^*)$ *be the marginal distribution of* $P(\mathbb{V})$ *over the counterfactual variables in the* CISTG$(\mathbf{A}_{\text{sub}}, \mathbf{V})$*. If* $P(\mathbb{V})$ *is in the* FRCISTG$(\mathbf{A}, \mathbf{V})$*, then* $P(\mathbb{V}^*)$ *is in* FRCISTG$(\mathbf{A}_{\text{sub}}, \mathbf{V})$*.*

Proof: The marginal distribution $P(\mathbb{V}^*)$ is in the FRCISTG$(\mathbf{A}_{\text{sub}}, \mathbf{V})$ model because the independencies defining the FRCISTG$(\mathbf{A}_{\text{sub}}, \mathbf{V})$ are implied by those defining the FRCISTG$(\mathbf{A}, \mathbf{V})$: the conditioning events are the same, $\mathbf{A}_{\text{sub}}$ is a subset of $\mathbf{A}$, and the counterfactuals on the lefthand side of the independence of (126) under FRCISTG $(\mathbf{A}_{\text{sub}}, \mathbf{V})$ are functions of and only of the counterfactuals on the lefthand side of (126) under FRCISTG $(\mathbf{A}, \mathbf{V})$ and the variables conditioned on. □

Lemma 64 does not hold with 'FRCISTG' replaced by either 'RCISTG' or 'RCISTG:$\mathbf{Y}$', or with $\mathbf{V}_{\text{sub}}$ substituted for $\mathbf{V}$. The latter implies that the FRCISTG model need not preserved when marginalizing over hidden variable $\mathbf{H} = \mathbf{V} \setminus \mathbf{V}_{\text{sub}}$; in other words if there are unobserved confounders then we are no longer 'fully randomized'.

Theorem 65. *For any distribution in a model FRCISTG* $(\mathbf{A}, \mathbf{V})$*:*

(a) *The joint distribution of* $\mathbb{V}(\mathbf{a}^\dagger) \equiv \overline{\mathbb{D}}_{K+1}(\overline{\mathbf{a}}^\dagger_K) = \{D_1(\overline{\mathbf{a}}^\dagger_1), \ldots, D_{K+1}(\overline{\mathbf{a}}^\dagger_K)\}$

is given in terms of the distributions of the factuals by

$$P(D_m(\overline{\mathbf{a}}^\dagger_{m-1}) = d_m \mid \overline{\mathbb{L}}_{m-1}(\overline{\mathbf{a}}^\dagger_{m-2}) = \overline{\mathbf{l}}_{m-1}) \\ = P(D_m = d_m \mid \overline{L}_{m-1} = \overline{\mathbf{l}}_{m-1}, \overline{A}_{m-1} = \overline{\mathbf{a}}^\dagger_{m-1}), \tag{128}$$

$$P(\overline{\mathbb{D}}_{K+1}(\overline{\mathbf{a}}^\dagger_K) = \overline{d}_{K+1}) = \prod_{m=1}^{K+1} P(D_m = d_m | \overline{\mathbf{L}}_{m-1} = \overline{\mathbf{l}}_{m-1}, \overline{\mathbf{A}}_{m-1} = \overline{\mathbf{a}}^\dagger_{m-1}) \tag{129}$$

where the RHS of (129) is the extended g-formula associated with $(\mathbf{A}, \mathbf{V})$. *Thus* $P(\mathbb{V}(\mathbf{a}^\dagger))$ *is identified whenever the extended g-formula is a unique function of the distribution of the factuals. Note* $D_m = (L_m, A_m)$ *and* $d_m = (l_m, a_m)$ *where* a_m *need not equal* $a^\dagger_m$.

(b) *The following independence holds:*

$$D_{m+1}(\overline{\mathbf{a}}^\dagger_m) \perp\!\!\!\perp \overline{\mathbb{A}}_m(\overline{\mathbf{a}}^\dagger_{m-1}) \mid \overline{\mathbb{L}}_m(\overline{\mathbf{a}}^\dagger_{m-1}) = \overline{l}_m, \\ \text{for all } m \leq K, \overline{\mathbf{a}}^\dagger_{K-1}, \overline{l}_m. \tag{130}$$

(c) *If (129) is a unique function of the distribution of the factuals then* $P(\mathbf{V})$ *factors relative to a DAG* $\mathcal{G}$ *(consistent with an ordering on* $\mathbf{V}$*) if and only if* $P(\mathbb{V}(\mathbf{a}^\dagger))$ *factors relative to* $\mathcal{G}(\mathbf{a}^\dagger)$. *In that case* $(\mathcal{G}, P(\mathbf{V}))$ *and* $(\mathcal{G}(\mathbf{a}^\dagger), P(\mathbb{V}(\mathbf{a}^\dagger)))$ *satisfy the modularity assumption (30).*

The condition in (a) that the g-formula is a unique function of the distribution of the factuals will hold if $P[\overline{\mathbf{L}}_i = \overline{\mathbf{l}}_i, \overline{\mathbf{A}}_{i-1} = \overline{\mathbf{a}}^\dagger_{i-1}] > 0$ implies $P\left[A_i = a^\dagger_i | \overline{\mathbf{L}}_i = \overline{l}_i, \overline{\mathbf{A}}_{i-1} = \overline{\mathbf{a}}^\dagger_{i-1}\right] > 0$.

Remark: If we were to replace A_m by $I(A_m = a^\dagger_m)$ in (126), as in the definition of the MCM considered in Robins and Richardson (2011), then the conclusions of Theorem 65 do not hold.[91]

Theorem 66. *For any distribution in the RCISTG* $(\mathbf{A}, \mathbf{V})$ *model*

[91] (Robins and Richardson, 2011, Appendix A, p.147) present a counterfactual distribution with a ternary treatment A and binary response $Y(a)$ which satisfies (126) but with A_m replaced by $I(A_m = a^\dagger_m)$. Under that model the distribution $P(A, Y(a))$ is not identified (even though the observed distribution is positive). Consequently (129) fails to hold.

(a) *the joint distribution of* $\mathbb{L}(\mathbf{a}^\dagger) \equiv \overline{\mathbb{L}}_{K+1}(\overline{\mathbf{a}}_K^\dagger) = \{L_1, L_2(\overline{\mathbf{a}}_1^\dagger), \ldots, L_{K+1}(\overline{\mathbf{a}}_K^\dagger)\}$ *is given in terms of the distributions of the factuals by*

$$\begin{aligned}
&P(L_m(\overline{\mathbf{a}}_{m-1}^\dagger) = l_m \,|\, \overline{\mathbb{L}}_{m-1}(\overline{\mathbf{a}}_{m-2}^\dagger) = \bar{\mathbf{l}}_{m-1}) \\
&\quad = P(L_m = l_m \,|\, \overline{L}_{m-1} = \bar{\mathbf{l}}_{m-1}, \overline{A}_{m-1} = \overline{\mathbf{a}}_{m-1}^\dagger) \\
&P(L_1(\mathbf{a}_1^\dagger) = l_1, \ldots, L_{K+1}(\overline{\mathbf{a}}_K^\dagger) = l_{K+1} \,|\, \overline{\mathbb{L}}_m(\overline{\mathbf{a}}_{m-1}^\dagger) = \bar{\mathbf{l}}_m) \\
&\quad = \prod_{m=1}^{K+1} P(L_m = l_m | \overline{\mathbf{L}}_{m-1} = \bar{l}_{m-1}, \overline{\mathbf{A}}_{m-1} = \overline{\mathbf{a}}_{m-1}^\dagger) \qquad (131)
\end{aligned}$$

where the RHS is the (unextended) g-formula.

Thus $P(\mathbb{L}(\mathbf{a}^\dagger) = \mathbf{l})$*, where* $\mathbf{l} = (l_1, \ldots, l_{K+1})$*, is identified whenever the extended g-formula is a unique function of the distribution of the factuals.*

(b) *If the density* (131) *is a unique function of the distribution of the factuals and the density* $P(\mathbf{V})$ *of the factuals factors according to a DAG* $\mathcal{G}$ *then the counterfactual distribution* $P(\mathbb{L}(\mathbf{a}^\dagger))$ *factors with respect to the induced subgraph of* $\mathcal{G}(\mathbf{a}^\dagger)$ *on* $\mathbb{L}(\mathbf{a}^\dagger)$*.*

Theorem 67. *For any distribution in the RCISTG:* $\mathbf{Y}(\mathbf{A}, \mathbf{V})$ *model the joint distribution of* $\mathbb{Y}(\mathbf{a}^\dagger) = \overline{Y}_{K+1}(\overline{\mathbf{a}}_K) = \{Y_1, Y_2(a_1^\dagger), \ldots, Y_{K+1}(\overline{\mathbf{a}}_K^\dagger)\}$ *satisfies for* $j = 1, \ldots, K$*:*

$$\begin{aligned}
&P\Big(Y_{j+1}(\overline{\mathbf{a}}_j^\dagger) = y_{j+1}, \ldots, Y_{K+1}(\overline{\mathbf{a}}_K^\dagger) = y_{K+1} \mid \overline{\mathbb{Y}}_j(\overline{\mathbf{a}}_{j-1}^\dagger) = \overline{\mathbf{y}}_j, \\
&\qquad\qquad \overline{\mathbb{W}}_j(\overline{\mathbf{a}}_{j-1}^\dagger) = \overline{\mathbf{w}}_j, \overline{\mathbb{A}}_j(\overline{\mathbf{a}}_{j-1}^\dagger) = \overline{\mathbf{a}}_j^\dagger\Big) \\
&\quad = \sum_{w_j, \ldots, w_K} \prod_{m=j+1}^{K} P(\{Y_m, W_m\} = (y_m, w_m) \mid \overline{\mathbf{Y}}_{m-1} = \overline{\mathbf{y}}_{m-1}, \qquad (132) \\
&\qquad\qquad \overline{\mathbf{W}}_{m-1} = \overline{\mathbf{w}}_{m-1}, \overline{\mathbf{A}}_{m-1} = \overline{\mathbf{a}}_{m-1}^\dagger)
\end{aligned}$$

where the RHS is the conditional g-formula for $\mathbf{Y}$*. Thus* $P(\mathbb{Y}(\mathbf{a}^\dagger) = \mathbf{y})$ *is identified whenever the RHS with* $j = 0$ *is a unique function of the distribution of the factuals.*

Note that for univariate Y, (132) corresponds to $b_{\mathbf{a}^\dagger}(y \mid \bar{\mathbf{l}}_j)$ given in Definition 21. We now illustrate the definitions and results given above with examples.

Example of an FRCISTG

Consider an FRCISTG $(\mathbf{A}, \mathbf{V})$ whose factual density $P(\mathbf{V})$ factors according to the DAG in Figure 12(i) with $\mathbf{V} = (H, A_0, L, A_1, Y)$. Let $\mathbf{V}_{obs} = (A_0, L, A_1, Y)$ be the observed variables. Then CISTG $(\mathbf{A}, \mathbf{V}_{obs})$ is a FRCIST $(\mathbf{A}, \mathbf{V}_{\text{obs}})$.

Proof: By Theorem 65 the counterfactual densities $P(\mathbb{V}(\mathbf{a}^\dagger))$ given by the FRCISTG $(\mathbf{A}, \mathbf{V})$ factors according to $\mathcal{G}(\mathbf{a}^\dagger)$ in Figure 12(ii). By d-separation on Figure 12(ii), we have

$$Y(a_0, a_1) \perp\!\!\!\perp A_1(a_0) \mid A_0, L(a_0) \quad \text{and} \quad \{Y(a_0, a_1), L(a_0), A_1(a_0)\} \perp\!\!\!\perp A_0$$

which imply the defining independencies (126) for the FRCISTG $(\mathbf{A}, \mathbf{V}_{\text{obs}})$.

Note in this example, as will be the case in the following examples, the factual densities $P(\mathbf{V}_{\text{obs}})$ associated with FRCISTG $(\mathbf{A}, \mathbf{V}_{\text{obs}})$ do not factor according to any incomplete DAG.

Example of an RCISTG

Consider an FRCISTG $(\mathbf{A}, \mathbf{V})$ with $\mathbf{V} = (H_1, H_2, A_1, L, A_2, Y)$ whose factual density $P(\mathbf{V})$ factors according to the DAG in Figure 19(i). Let $\mathbf{V}_{\text{obs}} = (A_1, A_2, L, Y)$ be the observed variables. Then the CISTG $(\mathbf{A}, \mathbf{V}_{\text{obs}})$ is an RCISTG $(\mathbf{A}, \mathbf{V}_{\text{obs}})$ but not an FRCISTG $(\mathbf{A}, \mathbf{V}_{\text{obs}})$.

Proof: By Theorem 65, the counterfactual densities $P(\mathbb{V}(\mathbf{a}^\dagger))$ implied by the FRCISTG $(\mathbf{A}, \mathbf{V})$ factor according to a graph $\mathcal{G}(\mathbf{a}^\dagger)$ shown in Figure 19(ii). By d-separation, we have

$$Y(a_1, a_2) \perp\!\!\!\perp A_2(a_1) \mid A_1, L(a_1) \quad \text{and} \quad \{Y(a_1, a_2), L(a_1)\} \perp\!\!\!\perp A_1,$$

which imply the defining independencies (126) for RCISTG $(\mathbf{A}, \mathbf{V}_{\text{obs}})$. However, $A_2(a_1)$ and A_1 are d-connected (given the empty set). Thus, by completeness (Theorem 19), there exists a law in FRCISTG $(\mathbf{A}, \mathbf{V})$ with $A_2(a_1)$ and A_1 dependent. Since the marginal of that law over $\mathbf{V}_{\text{obs}}$ is in CISTG $(\mathbf{A}, \mathbf{V}_{\text{obs}})$, we conclude that CISTG $(\mathbf{A}, \mathbf{V}_{\text{obs}})$ is not an FRCISTG $(\mathbf{A}, \mathbf{V}_{\text{obs}})$.

Example of an RCISTG: $\mathbf{Y}$ $(\mathbf{A}, \mathbf{V})$

Consider an FRCISTG $(\mathbf{A}, \mathbf{V})$ with $\mathbf{V} = (H_1, H_2, A_0, L, A_1, Y)$ whose factual densities $P(\mathbf{V})$ factor according to the DAG in Figure 13(i). Let

$\mathbf{V}_{\text{obs}} = (A_0, A_1, L, Y)$ be the observed variables. Then CISTG $(\mathbf{A}, \mathbf{V}_{\text{obs}})$ is an RCISTG: $\mathbf{Y}$ $(\mathbf{A}, \mathbf{V}_{\text{obs}})$ but not an RCISTG $(\mathbf{A}, \mathbf{V}_{\text{obs}})$.

Proof: By Theorem 65, the counterfactual densities $P(\mathbb{V}(\mathbf{a}^\dagger))$ implied by the FRCISTG $(\mathbf{A}, \mathbf{V})$ factors according to $\mathcal{G}(\mathbf{a}^\dagger)$ in Figure 13(ii). By d-separation, we have $Y(a_0, a_1) \perp\!\!\!\perp A_1(a_0) | A_0, L(a_0)$ and $Y(a_0, a_1) \perp\!\!\!\perp A_0$, which imply the defining independencies (126) for RCISTG: $\mathbf{Y}$ $(\mathbf{A}, \mathbf{V}_{\text{obs}})$. However, $A_1(a_0)$ and $L_1(a_0)$ are both d-connected to A_0. Thus, by completeness (Theorem 19), there exists a law in FRCISTG $(\mathbf{A}, \mathbf{V})$ with neither $A_1(a_0)$ nor $L_1(a_0)$ independent of A_0. Since the marginal of that law over $\mathbf{V}_{\text{obs}}$ is in CISTG $(\mathbf{A}, \mathbf{V}_{\text{obs}})$, we conclude that CISTG $(\mathbf{A}, \mathbf{V}_{\text{obs}})$ is not an RCISTG $(\mathbf{A}, \mathbf{V}_{\text{obs}})$.

C.3 Individual Level Exclusion Restrictions and Graphs

In the development in Section 3 the graph was taken as a primitive and the NPSEM was then built on top of this. Our development in the previous section, which follows Robins (1986, 1987) differed from this in that the definitions of CISTG$(\mathbf{A}, \mathbf{V})$ and FRCISTG $(\mathbf{A}, \mathbf{V})$ do not make any mention of graphs. In more detail:

- The definition of CISTG$(\mathbf{A}, \mathbf{V})$ in §C.2 makes no mention of a graph $\mathcal{G}$, nor of individual-level exclusion restrictions;
- The definition of FRCISTG$(\mathbf{A}, \mathbf{V})$ in §C.2 is relative to an ordering of the variable set $\mathbf{V}$;
- The underlying counterfactual existence assumption (Def. 60) does not refer to a graph, in contrast to the assumption in §3 (Def. 1);
- In §3 every variable could be intervened upon, so $\mathbf{A} = \mathbf{V}$.

Here we explain how starting from the definitions in the previous section, by additional additional background information, one may arrive at the definition in Section 3. This thus shows that in logical terms, the graphical theory presented in §3 may be reduced to a purely counterfactual theory.

C.4 CISTG with exclusion restrictions

Clause (i) of the counterfactual existence assumption given by Definition 60 takes the set of one-treatment-ahead counterfactuals $D_m(\bar{\mathbf{a}}_{m-1})$ as primitive. All other counterfactuals are then derived as simple recursive functions of these via clause (ii). Note that formally, the set of counterfactual variables

that are 'one-treatment-ahead' is determined by the ordering of the set $\mathbf{V}$. Since we will wish to consider the effect of different choices for the ordering we will write $\text{CISTG}_{\prec}(\mathbf{A}, \mathbf{V})$ to indicate the dependence on the choice of an ordering $\prec$ of $\mathbf{V}$.

Robins (1986, §8) considered $\text{CISTG}(\mathbf{A}, \mathbf{V})$ models that also satisfied individual level exclusion restrictions:

Definition 68. *Given a pair* $(\mathbf{A}, \mathbf{V})$*, an ordering* $\prec'$ *on* $\mathbf{V}$ *satisfying the counterfactual existence assumption, and any* $V \in \mathbf{V}$*, let* $\mathbf{A}_V$ *be the smallest subset of* $\mathbf{A}$ *satisfying the individual exclusion restriction that, for all* $\mathbf{a}$*,* $V(\mathbf{a}) = V(\mathbf{a}_V)$ *for each unit. The variables in* $\mathbf{A}_V^{\prec}$ *are the* individual level direct causes of V (relative to $\mathbf{A}$) *under ordering* $\prec$*.*

We will show below that so long as each variable is ordered after its direct causes, the choice of ordering here is not important.

Definition 69. *The* $\text{CISTG}_{\prec}(\mathbf{A}, \mathbf{V})$ with individual level direct causes $\left\{\mathbf{A}_V^{\prec};\ V \in \mathbf{V}\right\}$ model *is the submodel of the* $\mathit{CISTG}_{\prec}(\mathbf{A}, \mathbf{V})$ *model in which for each* V*,* $\mathbf{A}_V$ *are the direct causes of* V

Consider a $\text{CISTG}_{\prec}(\mathbf{A}, \mathbf{V})$ with individual level direct causes $\left\{\mathbf{A}_V^{\prec}; V \in \mathbf{V}\right\}$. Let $\mathfrak{O}_{\prec}$ be the set of orderings of $\mathbf{V}$ such that for every $V \in \mathbf{V}$, each element of $\mathbf{A}_V^{\prec}$ precedes V in the ordering; note that by definition $\prec \in \mathfrak{O}_{\prec}$.

Proposition 70. *Let* $\mathbb{V}_{\prec}$ *be the set of counterfactual random variables in* $\mathcal{C}_{\prec}$*, a* $\text{CISTG}_{\prec}(\mathbf{A}, \mathbf{V})$*. Let* $\mathbb{B}_{\prec^*} \subseteq \mathbb{V}_{\prec}$ *be the subset of random variables in* $\mathbb{V}_{\prec}$ *that are 'one-treatment-ahead' counterfactuals under a different order* $\prec^* \in \mathfrak{O}_{\prec}$*. Finally, let* $\mathbb{V}_{\prec^*}$ *be the set of counterfactual random variables derived from* $\mathbb{B}_{\prec^*}$ *via recursive substitution* (125)*. Then* $\mathbb{V}_{\prec^*} = \mathbb{V}_{\prec}$*.*[92]

Proof: Given a variable $D \in \mathbf{V}$ with indices m and m^* under the orderings $\prec$ and $\prec^*$, let $D_{\prec,m}(\tilde{\mathbf{r}})$ and $D_{\prec^*,m^*}(\tilde{\mathbf{r}})$ be the corresponding variables in $\mathbb{V}_{\prec}$ and $\mathbb{V}_{\prec^*}$ respectively. By construction, any one-treatment ahead counterfactual $D_{\prec^*,m^*}(\overline{\mathbf{a}}^{\dagger}_{\prec^*,m-1}) \in \mathbb{B}_{\prec^*} \subseteq \mathbb{V}_{\prec}$, so

$$D_{\prec^*,m^*}(\overline{\mathbf{a}}^{\dagger}_{\prec^*,m^*-1}) = D_{\prec,m}(\overline{\mathbf{a}}^{\dagger}_{\prec^*,m^*-1}), \tag{133}$$

here $\overline{\mathbf{a}}^{\dagger}_{\prec^*,m^*-1}$ is an assignment to $\overline{\mathbf{A}}_{\prec^*,m^*-1}$ the set of treatment variables prior to D under the orderings $\prec^*$. It is thus sufficient to show by induction

[92] Here we are asserting equality of the individual random variables within each set. In other words, for every variable $T \in \mathbf{V}$ and intervention $\tilde{\mathbf{c}}$, where $\mathbf{C} \subseteq \mathbf{A}$, $T_{\prec}(\tilde{\mathbf{c}}) = T_{\prec^*}(\tilde{\mathbf{c}})$, for every unit of the population, where here $T_{\prec}(\tilde{\mathbf{c}})$ and $T_{\prec^*}(\tilde{\mathbf{c}})$ denote the associated counterfactuals associated variables within the sets $\mathbb{V}_{\prec}$ and $\mathbb{V}_{\prec^*}$.

that each application of recursive substitution results in the same random variable. For any assignment $\tilde{\mathbf{r}}$ to $\mathbf{R} \subseteq \mathbf{A}$, we have the following:

$$\begin{aligned} D_{\prec^*,m^*}(\tilde{\mathbf{r}}) &= D_{\prec^*,m^*}\left(\tilde{\mathbf{r}}_{\overline{\mathbf{A}}_{\prec^*,m^*-1}\cap\mathbf{R}},\ \mathbf{W}^*_{\prec^*}(\tilde{\mathbf{r}})\right) \\ &= D_{\prec,m}\left(\tilde{\mathbf{r}}_{\overline{\mathbf{A}}_{\prec^*,m^*-1}\cap\mathbf{R}},\ \mathbf{W}^*_{\prec^*}(\tilde{\mathbf{r}})\right) \\ &= D_{\prec,m}\left(\tilde{\mathbf{r}}_{\overline{\mathbf{A}}_{\prec^*,m^*-1}\cap\mathbf{R}},\ \mathbf{W}_{\prec^*}(\tilde{\mathbf{r}})\right) \\ &= D_{\prec,m}\left(\tilde{\mathbf{r}}_{\mathbf{A}^{\prec}_D\cap\mathbf{R}},\ \mathbf{W}_D(\tilde{\mathbf{r}})\right) \\ &= D_{\prec,m}\left(\tilde{\mathbf{r}}_{\overline{\mathbf{A}}_{\prec,m-1}\cap\mathbf{R}},\ \mathbf{W}_{\prec}(\tilde{\mathbf{r}})\right) = D_{\prec,m}(\tilde{\mathbf{r}}), \end{aligned} \tag{134}$$

where

$$\begin{aligned} \mathbf{W}^*_{\prec^*}(\tilde{\mathbf{r}}) &\equiv \{A_{\prec^*}(\tilde{\mathbf{r}}) \mid A \in \overline{\mathbf{A}}_{\prec^*,m^*-1}, A \notin \mathbf{R}\} && \subseteq \mathbb{V}_{\prec^*}, \\ \mathbf{W}_{\prec^*}(\tilde{\mathbf{r}}) &\equiv \{A_{\prec}(\tilde{\mathbf{r}}) \mid A \in \overline{\mathbf{A}}_{\prec^*,m^*-1}, A \notin \mathbf{R}\} && \subseteq \mathbb{V}_{\prec}, \\ \mathbf{W}_D(\tilde{\mathbf{r}}) &\equiv \{A_{\prec}(\tilde{\mathbf{r}}) \mid A \in \mathbf{A}^{\prec}_D, A \notin \mathbf{R}\} && \subseteq \mathbb{V}_{\prec}, \\ \mathbf{W}_{\prec}(\tilde{\mathbf{r}}) &\equiv \{A_{\prec}(\tilde{\mathbf{r}}) \mid A \in \overline{\mathbf{A}}_{\prec,m-1}, A \notin \mathbf{R}\} && \subseteq \mathbb{V}_{\prec}, \end{aligned}$$

and $\overline{\mathbf{A}}_{\prec,m-1}$ are the treatment variables prior to D under the ordering $\prec$. In (134) the first equality follows by (125); the second by (133); the third is by the inductive hypothesis, which implies $\mathbf{W}^*_{\prec^*}(\tilde{\mathbf{r}}) = \mathbf{W}_{\prec^*}(\tilde{\mathbf{r}})$; the fourth and fifth follow by the definition of $\mathbf{A}^{\prec}_D$ since both $\prec^*$ and $\prec$ are in $\mathfrak{O}_{\prec}$; the sixth is again by (125). □

It follows directly from Proposition 70 that if a set of counterfactuals forms a CISTG$_{\prec}(\mathbf{A},\mathbf{V})$ then it also forms a CISTG$_{\prec^*}(\mathbf{A},\mathbf{V})$ for any other ordering $\prec^* \in \mathfrak{O}_{\prec}$. Consequently, $\prec \in \mathfrak{O}_{\prec^*} \Leftrightarrow \prec^* \in \mathfrak{O}_{\prec}$, hence $\mathfrak{O}_{\prec^*} = \mathfrak{O}_{\prec}$. Thus a CISTG and the set of direct causes are invariant to the choice of ordering in this class, consequently we suppress the ordering and write CISTG$(\mathbf{A},\mathbf{V})$ and $\mathbf{A}_V$. Such a result was first given in Section 11.F of Robins (1986) .

The invariance to choice of order extends to the FFRCISTG$_{\prec}(\mathbf{V},\mathbf{V})$ submodel of the CISTG$(\mathbf{V},\mathbf{V})$ with direct causes:

Lemma 71. *Consider a counterfactual distribution $P(\mathbb{V})$ that is in the* FRCISTG$_{\prec}(\mathbf{V},\mathbf{V})$ *with individual direct causes $\{\mathbf{A}_V \mid V \in \mathbf{V}, \mathbf{A}_V \subset \mathbf{V} = \mathbf{A}\}$. For any $\prec^*$ in $\mathfrak{O}_{\prec}$, $P(\mathbb{V})$ is also in the* FRCISTG$_{\prec^*}(\mathbf{V},\mathbf{V})$ *with the same individual direct causes.*

Thus given a $\text{CISTG}_{\prec}(\mathbf{V}, \mathbf{V})$ the $\text{FRCISTG}_{\prec^*}(\mathbf{V}, \mathbf{V})$ submodels are equivalent for all $\prec^* \in \mathfrak{O}_{\prec}$. We may naturally associate a graph $\mathcal{G}$ with an $\text{FRCISTG}_{\prec}(\mathbf{V}, \mathbf{V})$ model with direct causes by setting the parents of $\text{pa}_{\mathcal{G}}(V) \equiv \mathbf{A}_V$ for all $V \in \mathbf{V}$. Under this correspondence the $\text{FRCISTG}_{\prec}(\mathbf{V}, \mathbf{V})$ is precisely the NPSEM model associated with $\mathcal{G}$ under the FFRCISTG[93] independence assumption (17), introduced in Section 3.

Proof: This follows from the fact that Robins and Richardson (2011) prove that the defining independences (126) with $\mathbf{Z} = \mathbf{D}$ for an $\text{FRCISTG}_{\prec}(\mathbf{V}, \mathbf{V})$ with direct causes $\{\mathbf{A}_V\}$ are equivalent to the independences (17) for the graph $\mathcal{G}$ with $\text{pa}_{\mathcal{G}}(V) \equiv \mathbf{A}_V$ for $V \in \mathbf{V}$; the latter set of independence relations are invariant to the choice of topological ordering $\prec$ for $\mathcal{G}$. □

Corollary 72. *Given a distribution $P(\mathbb{V})$ in the $\text{FRCISTG}_{\prec}(\mathbf{V}, \mathbf{V})$ with direct causes $\{\mathbf{A}_V\}$. Let $\mathbf{A}^* \subseteq \mathbf{V}$, and let $P(\mathbb{V}^*)$ be the marginal distribution over the counterfactual variables in $\text{CISTG}(\mathbf{A}^*, \mathbf{V})$. For any ordering $\prec^*$ in $\mathfrak{O}_{\prec}$, $P(\mathbb{V}^*)$ is in $\text{FRCISTG}_{\prec^*}(\mathbf{A}^*, \mathbf{V})$.*

Proof: This follows directly from Lemmas 71 and 64. □

The above Lemma and Corollary would not hold in general had we taken as our premise that $P(\mathbb{V})$ is in the $\text{FRCISTG}_{\prec}(\mathbf{A}, \mathbf{V})$, for $\mathbf{A}$ a strict subset of $\mathbf{V}$. Likewise it would not hold for either the $\text{RCISTG}(\mathbf{A}, \mathbf{V})$ or RCISTG $\mathbf{Y} : (\mathbf{A}, \mathbf{V})$ submodels of a $\text{CISTG}_{\prec}(\mathbf{A}, \mathbf{V})$ with direct causes $\{\mathbf{A}_V\}$. As the examples in the next section show these models are not invariant to the choice of ordering within $\mathfrak{O}_{\prec}$. Consequently, the g-formula will represent a causal quantity only under some topological orderings and not others. However, we note that, given a DAG $\mathcal{G}$, the ID algorithm of Tian and Pearl (2002) and the IDC algorithm of Shpitser and Pearl (2006a) will identify all the intervention distributions that are identified by a g-formulae, in addition to identifying other intervention distributions, such as $P(Y(a_1, a_2))$ in Figure 14 that are not identified via g-formula.

C.4.1 Examples

Consider the graph $\mathcal{G}$ shown in Figure 17(a) with $\mathbf{V} = \langle H_1, A, S, Y \rangle$, denoting the implied ordering $\prec$. Consider the $\text{FRCISTG}_{\prec}(\mathbf{V}, \mathbf{V})$, with direct effects relative to $\mathbf{A} = \mathbf{V}$ given by the parents in $\mathcal{G}$:

$$\mathbf{A}_{H_1} = \emptyset;\ \mathbf{A}_A = \{H_1\};\ \mathbf{A}_S = \{H_1\};\ \mathbf{A}_Y = \{A, S\}.$$

[93] Note that if $\mathbf{A} = \mathbf{V}$ then clearly the set of counterfactuals cannot be enlarged. Thus the FRCISTG is 'finest' and thus an *F*FRCISTG.

In this example the set of orderings

$$\mathfrak{O}_{\prec} = \{\langle H_1, A, S, Y\rangle, \langle H_1, S, A, Y\rangle\}$$

corresponding to the topological orderings of $\mathcal{G}$. As noted above, this model corresponds to the FFRCISTG associated with $\mathcal{G}$.

By Corollary 72 given any distribution $P(\mathbb{V})$ in the FRCISTG$_{\prec}(\mathbf{V}, \mathbf{V})$, the margin $P(\mathbb{V}^*)$ over the counterfactual variables in CISTG$(\mathbf{A}, \mathbf{V})$ also obeys the FRCISTG$_{\prec^*}(\mathbf{A}, \mathbf{V})$ assumptions for both orderings $\prec^*$ in $\mathfrak{O}_{\prec}$.

However, if we now consider the subset $\mathbf{V}_{\text{obs}} = \{A, S, Y\} \subseteq \mathbf{V}$, a margin $P(\mathbb{V}')$ over the counterfactual variables in CISTG$(\mathbf{A}, \mathbf{V}_{\text{obs}})$ is an FRCISTG under the ordering $\langle S, A, Y\rangle$ but not $\langle A, S, Y\rangle$. This may be seen by inspecting the template $\mathcal{G}(a)$ shown in Figure 17(b), since we have: $Y(a) \perp\!\!\!\perp A \mid S$ as required by $\langle S, A, Y\rangle$, but not $Y(a) \perp\!\!\!\perp A$, which is the ordering $\langle A, S, Y\rangle$ requires.

For a second example, consider the FRCISTG$_{\prec}(\mathbf{V}, \mathbf{V})$, with

$$\mathbf{V} = \langle H_0, A_0, D_1, H_1, A_1, D_2\rangle$$

with direct causes given by the graph $\mathcal{G}$ shown in Figure 18(a), again using $\prec$ to denote the ordering.

Given a distribution $P(\mathbb{V})$ in FRCISTG$_{\prec}(\mathbf{V}, \mathbf{V})$, let $P(\mathbb{V}^*)$ be the marginal distribution over the counterfactuals in the CISTG$(\mathbf{A}, \mathbf{V}_{\text{obs}})$, where $\mathbf{A} = \{A_0, A_1\}$ and $\mathbf{V}_{\text{obs}} = \{A_0, D_1, A_1, D_2\}$. Under the (natural) ordering induced by $\prec$ on $\mathbf{V}_{\text{obs}}$, $P(\mathbb{V}^*)$ is not in RCISTG$_{\prec}(\mathbf{A}, \mathbf{V}_{\text{obs}})$ nor even in RCISTG$_{\prec}$ $\mathbf{Y} : (\mathbf{A}, \mathbf{V}_{\text{obs}})$, with $\mathbf{Y} = \{D_2\}$. However, under the ordering

$$\prec^* = \langle D_1, A_0, D_2, A_1\rangle,$$

which is in $\mathfrak{O}_{\prec}$ for the CISTG$_{\prec}(\mathbf{A}, \mathbf{V}_{\text{obs}})$, $P(\mathbb{V}^*)$ is in the RCISTG$_{\prec^*}(\mathbf{A}, \mathbf{V}_{\text{obs}})$. Consequently the distribution $P(D_2(a_0, a_1)) = P(D_2(a_0))$ is identified. See Robins (1986, §11) and §4.2.4 for further discussion.

C.5 Proofs of results relating to RCISTGs

We now prove Theorems 65, 66 and 67.

Lemma 73. *The defining independence* (126) *for an FRCISTG* $(\mathbf{A}, \mathbf{V})$ *is equivalent to the following holding for all* $1 \leq m \leq K + 1$*:*

$$D_{m+1}(\overline{\mathbf{a}}_m^{\dagger}) \perp\!\!\!\perp \overline{\mathbb{A}}_m(\overline{\mathbf{a}}_{m-1}^{\dagger}) \mid \overline{\mathbb{L}}_m(\overline{\mathbf{a}}_{m-1}^{\dagger}).$$

Proof: The defining independence (126) for an FRCISTG $(\mathbf{A}, \mathbf{V})$ is that:

$$\left\{D_{k+1}(\overline{\mathbf{a}}_k^\dagger), \ldots, D_{K+1}(\overline{\mathbf{a}}_K^\dagger)\right\} \perp\!\!\!\perp A_k \mid \overline{\mathbf{L}}_k = \bar{\mathbf{l}}_k, \overline{\mathbf{A}}_{k-1} = \overline{\mathbf{a}}_{k-1}^\dagger,$$

for all k, all $\mathbf{a}^\dagger$ and all $\bar{\mathbf{l}}_k$. Under consistency, this holds if and only if

$$\left\{D_{k+1}(\overline{\mathbf{a}}_k^\dagger), \ldots, D_{K+1}(\overline{\mathbf{a}}_K^\dagger)\right\} \perp\!\!\!\perp A_k(\overline{\mathbf{a}}_{k-1}^\dagger) \mid L_1 = l_1, A_1 = a_1^\dagger, L_2(a_1^\dagger) = l_2, \ldots, L_{k-1}(\overline{\mathbf{a}}_{k-2}^\dagger) = l_{k-1}, A_{k-1}(\overline{\mathbf{a}}_{k-2}^\dagger) = \overline{\mathbf{a}}_{k-1}^\dagger, L_k(\overline{\mathbf{a}}_{k-1}^\dagger) = l_k,$$

which may be re-written more compactly as:

$$\left\{D_{k+1}(\overline{\mathbf{a}}_k^\dagger), \ldots, D_{K+1}(\overline{\mathbf{a}}_K^\dagger)\right\} \perp\!\!\!\perp A_k(\overline{\mathbf{a}}_{k-1}^\dagger) \mid \overline{\mathbb{L}}_k(\overline{\mathbf{a}}_{k-1}^\dagger) = \bar{\mathbf{l}}_k, \overline{\mathbb{A}}_{k-1}(\overline{\mathbf{a}}_{k-2}^\dagger) = \overline{\mathbf{a}}_{k-1}^\dagger. \tag{135}$$

Since $D_k \equiv (L_k, A_k)$ it follows from the weak union and decomposition properties that (135) is equivalent to the following conditional independences:

$$D_{m+1}(\overline{\mathbf{a}}_m^\dagger) \perp\!\!\!\perp A_k(\overline{\mathbf{a}}_{k-1}^\dagger) \mid \overline{\mathbb{L}}_m(\overline{\mathbf{a}}_{m-1}^\dagger) = \bar{\mathbf{l}}_m, A_{k+1}(\overline{\mathbf{a}}_k^\dagger), \ldots, A_m(\overline{\mathbf{a}}_{m-1}^\dagger), \overline{\mathbb{A}}_{k-1}(\overline{\mathbf{a}}_{k-2}^\dagger) = \overline{\mathbf{a}}_{k-1}^\dagger, \quad \text{for } 1 \leq k \leq m \leq K.$$

This is then equivalent to it holding for $1 \leq k \leq m \leq K$ that:

$$D_{m+1}(\overline{\mathbf{a}}_m^\dagger) \perp\!\!\!\perp A_k(\overline{\mathbf{a}}_{k-1}^\dagger), \ldots, A_m(\overline{\mathbf{a}}_{m-1}^\dagger) \mid \overline{\mathbb{L}}_m(\overline{\mathbf{a}}_{m-1}^\dagger) = \bar{\mathbf{l}}_m, \overline{\mathbb{A}}_{k-1}(\overline{\mathbf{a}}_{k-2}^\dagger) = \overline{\mathbf{a}}_{k-1}^\dagger.$$

The conclusion follows, taking $k = 1$. □

Proof of Theorem 65: Lemma 73 establishes the independence (130) given in (b). This is the (ordered) local Markov property for the SWIG $\overline{\mathcal{G}}(\mathbf{a}^\dagger)$ obtained from a complete DAG $\overline{\mathcal{G}}$ for $P(\mathbf{V})$, with edges oriented according to the ordering $(L_1, A_1, \ldots, L_K, A_K, L_{K+1})$. By results in (Lauritzen et al., 1990) the ordered local Markov property is equivalent to the global Markov property (24). It then follows from (e.g. Lauritzen, 1996, Thm. 3.27) that:

$$P(\overline{\mathbb{D}}_{K+1}(\overline{\mathbf{a}}_K^\dagger) = \mathbf{d}) = \prod_{m=1}^{K+1} P(D_m(\overline{\mathbf{a}}_{m-1}^\dagger) = d_m \mid \overline{\mathbb{L}}_{m-1}(\overline{\mathbf{a}}_{m-2}^\dagger) = \bar{\mathbf{l}}_{m-1}). \tag{136}$$

It further follows from (130) and consistency that:

$$\begin{aligned} &P(D_m(\overline{\mathbf{a}}_{m-1}^\dagger) = d_m \mid \overline{\mathbb{L}}_{m-1}(\overline{\mathbf{a}}_{m-2}^\dagger) = \bar{\mathbf{l}}_{m-1}) \\ &\quad = P(D_m(\overline{\mathbf{a}}_{m-1}^\dagger) = d_m \mid \overline{\mathbb{L}}_{m-1}(\overline{\mathbf{a}}_{m-2}^\dagger) = \bar{\mathbf{l}}_{m-1}, \overline{\mathbb{A}}_{m-1}(\overline{\mathbf{a}}_{m-2}^\dagger) = \overline{\mathbf{a}}_{m-1}^\dagger) \\ &\quad = P(A_m = a_m^\dagger \mid \overline{\mathbf{L}}_m = \bar{\mathbf{l}}_m, \overline{\mathbf{A}}_{m-1} = \overline{\mathbf{a}}_{m-1}^\dagger) \qquad (137) \\ &\qquad \times P(L_m = l_m \mid \overline{\mathbf{L}}_{m-1} = \bar{\mathbf{l}}_{m-1}, \overline{\mathbf{A}}_{m-1} = \overline{\mathbf{a}}_{m-1}^\dagger). \qquad (138) \end{aligned}$$

which establishes (a). The 'only if' in (c) then follows because given an incomplete DAG $\mathcal{G}$, factorization of $P(\mathbf{V})$ with respect to $\mathcal{G}$ implies that the conditioning set on the left hand side of (137) and (138) may be reduced to

$$\text{pa}_{\mathcal{G}(\mathbf{a}^\dagger)}(A_m(\bar{\mathbf{a}}^\dagger_{m-1})) \quad \text{and } \text{pa}_{\mathcal{G}(\mathbf{a}^\dagger)}(L_m(\bar{\mathbf{a}}^\dagger_{m-1}))$$

respectively. Since this holds for every $\mathbf{a}^\dagger$ and $\mathbf{l}$ then (136) implies that $P(\mathbf{V})$ factorizes with respect to $\mathcal{G}$. The proof of the converse is essentially the same. □

Proof of Theorem 66: Property (a) follows from Theorem 4.1 of Robins (1986). The second claim follows from (131) by an argument similar to that establishing (c). □

Note that whereas (c) in Theorem 65 asserts that $P(\mathbf{V})$ factorizes with respect to $\mathcal{G}$ if and only if $P(\mathbb{V}(\mathbf{a}^\dagger))$ factorizes with respect to $\mathcal{G}(\mathbf{a}^\dagger))$ for all $\mathbf{a}^\dagger$, in Theorem 66 (b) the implication is only one way. This is because where $\mathbf{V}$ contains the variables in $\mathbf{A}$, under the RCISTG $(\mathbf{A}, \mathbf{V})$ model only the distribution over the counterfactual variables in $\mathbf{V} \setminus \mathbf{A} = \mathbf{L}$ is identified.

Proof of Theorem 67: This follows from the Corollary to Theorem AD.1 in Robins (1987) and Lemma 58 in Appendix §B.3. □.

D A brief history of Pearl's approaches to linking graphs and counterfactuals

Pearl (2000, §3.6.3, §7.1) provides a 'translation' between graphs and counterfactuals. This translation is built on two pillars: interpreting causal DAGs as a non-parametric structural equation model with independent errors (NPSEM-IE), and constructing 'Twin-networks' for the purpose of making inferences regarding counterfactuals. We examine each in turn.

D.1 Counterfactual models based on NPSEM-IEs

The first is to associate a DAG with a non-parametric structural equation model with independent errors (NPSEM-IE). It has long been understood that NPSEMs may be viewed as a counterfactual model via the classical idea (see e.g. Box, 1966; Haavelmo, 1943; Pearl, 2000; Spirtes et al., 1993; Strotz and Wold, 1960) of constructing counterfactual distributions by removing equations from the model and replacing them with equations that fix a variable to a given value. Though this approach is intuitive and provides

a bridge to the literature in the social sciences, the requirement that the error terms be independent is strong. As shown in Table 6, this assumption implies super exponentially many additional 'cross-world' counterfactual independence relations compared to the FFRCISTG model. As commented on at length elsewhere (Dawid, 2000; Robins, 2003; Robins and Richardson, 2011) and in §6 of this paper, this has the consequence that in order to posit an NPSEM-IE model researchers must be in a position to make hypotheses about the underlying causal process that they cannot even in principle verify via any randomized experiment on the variables in the graph. Furthermore these additional independence assumptions lead to additional identification results, such as for the Pure Direct Effect (Robins and Greenland, 1992) also known as the Natural Direct Effect (Pearl, 2001a); see §6.1.1.

D.2 The twin network procedure

The second pillar of Pearl's synthesis between graphs and counterfactuals under the NPSEM-IE, is the 'twin network' method, later generalized to multi-networks. The method is claimed to provide a graphical method for evaluating the truth of counterfactual independence queries. The basic idea is to create a a set of linked graphs – consisting of the the actual world graph plus separate graphs for each counterfactual world referenced by the query.

The actual variable (e.g. V) and counterfactual variables (e.g. $V(a)$, $V(b)$) are linked through their common error term U_V.[94] Examples are shown in Figures 29(a) and 31(ii). In the twin- or multi-network associated with a DAG, the error terms associated with different variables U_V, U_W are assumed independent, reflecting the NPSEM-IE independence assumption (18). The explicit inclusion of error terms in these networks means that actual and counterfactual variables are *deterministic* functions of their parents in the graph.

Pearl has advocated his unification of DAGs and potential outcome models given in Chapter 7 of Pearl (2000, 2009) via twin- and multi-networks in many places (e.g. Pearl, 2012b, p.5). However, in contrast to our SWIG approach, these network methods are difficult to correctly apply, owing to the presence of deterministic relations between variables in the networks. The most powerful argument for this claim is that, as we shall see, Pearl himself made numerous errors in applying these methods particularly in the first edition of his book Causality, but also in the second edition. In fact we take his continuing difficulty in properly applying the network methods even in his

[94] Elsewhere we have used ε_V to indicate error terms. In order to facilitate the exposition we here follows Pearl's usage.

second edition as a major motivation for the simpler SWIG based approach presented herein. We turn now to the first edition of his book Causality . Although Pearl did not warn the readers of the second edition of the errors made in the first, he did attempt to correct them (though he continued to make analogous errors in the second edition).

To substantiate our claim that network methods are difficult to apply we demonstrate below both how deterministic relations contributed to Pearl's errors and how an approach based on SWIGs would have prevented certain of his mistakes.

D.2.1 Mistaken treatment of error terms as single counterfactual variables

In the first edition, Pearl's treatment of the error terms (U_Z) in twin networks was mistaken. Specifically Pearl (2000, p. 214), in the last paragraph states:

> The twin network reveals an interesting interpretation of counterfactuals of the form $Z(\text{pa}_Z)$, where Z is any variable and PA_Z stands for the set of Z's parents. Consider the question of whether $Z(x)$ is independent of some given set of variables in the model of Figure 7.3. The answer to this question depends on whether Z^* is d-separated from that set of variables. However, any variable that is *d*-separated from Z^* would also be *d*-separated from U_Z, so the node representing U_Z can serve as a proxy for representing the counterfactual variable $Z(x)$. [counterfactual notation changed[95]]

For reference, the graph in Figure 7.3 is shown in Figure 29(ii).

Pearl (2000) p.215, first paragraph goes on to say:

> We thus obtain a simple graphical representation for any counterfactual variable of the form $Z(pa_Z)$. Using this representation we can *easily* verify from Figure 7.3 that $(Y^* \perp\!\!\!\perp X \mid \{Z, U_Z, Y\})_G$ and $(Y^* \perp\!\!\!\perp X \mid \{U_Y, U_Z, Y\})_G$ both hold in the twin-network and, therefore,
>
> $$Y(x) \perp\!\!\!\perp X \mid \{Z, Z(x), Y\} \quad \textit{and} \quad Y(x) \perp\!\!\!\perp X \mid \{Y(z), Z(x), Y\}$$
>
> must hold in the model. [emphasis added; notation changed]

[95] Here and elsewhere we replace Pearl's subscript notation for counterfactuals (e.g. Z_x) with our parenthetical notation $Z(x)$.

Pearl's last claim is not easily verified as it is false. The two d-separation relations $(Y^* \perp\!\!\!\perp X \mid \{Z, U_Z, Y\})_G$ and $(Y^* \perp\!\!\!\perp X \mid \{U_Y, U_Z, Y\})_G$ do indeed hold in the twin-network graph. Since the error term U_V for a given variable V is equivalent to the *set* of counterfactual variables $\{V(\text{pa}_V)\}$ for all instantiations pa_V of the parents of V, it follows that under the NPSEM-IE association with the graph G,

$$Y(x) \perp\!\!\!\perp X \mid \{Z, Y, Z(x') \text{ for all } x' \in \mathfrak{X}\}$$
$$Y(x) \perp\!\!\!\perp X \mid \{Y(z'), Z(x'), Y, \text{ for all } x' \in \mathfrak{X}, z' \in \mathfrak{Z}\}$$

where $\mathfrak{X}$ and $\mathfrak{Z}$ are the state-spaces for X and Z respectively. However, the latter condition does not imply that $Y(x)$ and X will be independent when conditioning on a subset of these variables; the inference that $Y(x) \perp\!\!\!\perp X \mid \{Y(z), Z(x), Y\}$ holds is erroneous.

D.3 Partial Correction in the Second Edition of Errors in the First Edition

Pearl (2009) corrects the above mistakes in the second edition at least partially. Specifically Pearl no longer claims that U_z is a proxy for $Z(x)$. Rather the sentence has been changed to read:

> ...so the node representing U_Z can serve as a *one-way* proxy for representing the counterfactual variable $Z(x)$. [emphasis added; counterfactual notation changed]

No definition of 'one-way proxy' is provided, nor is any explanation given for the change. However Pearl's thinking is clarified by comparing the first paragraph on p. 215 in the first edition (quoted above) with the same paragraph in the second edition which has been changed to read:

> We thus obtain a simple graphical representation for any counterfactual variable of the form $Z(pa_Z)$, in terms of the so called "error-term" U_Z. Using this representation we can *easily* verify from Figure 7.3 that $(U_Y \perp\!\!\!\perp X \mid \{Y^*, Z^*\})_G$ and $(U_Y \perp\!\!\!\perp U_Z \mid \{Y, Z\})_G$ both hold in the twin-network and, therefore,
>
> $$Y(z) \perp\!\!\!\perp X \mid \{Y(x), Z(x)\} \quad \textit{and} \quad Y(z) \perp\!\!\!\perp Z(x) \mid \{Y, Z\}$$
>
> must hold in the model. [emphasis added; notation changed]

Thus in the second edition all four conditional independence statements in this paragraph have been changed. In particular he now avoids conditioning on the error U_Z and/or U_Y, in order to avoid repeating the inferential error made in the first edition. Nonetheless Pearl fails to change the following statement in the second edition:

> Therefore, if U_Z obeys a certain independence relationship then $Z(x)$ (more generally $Z(pa_Z)$) must obey that relationship as well. [notation changed] Pearl (2000, 2009, p.214)

If "U_Z obeys a certain independence relationship" is interpreted as including independencies in which U_Z appears in the conditioning set then this statement is incorrect for the reason we gave above. That is, since $Z(pa_Z)$ is a deterministic function of U_Z the implication:

$$A \perp\!\!\!\perp U_Z \mid B \quad \Rightarrow \quad A \perp\!\!\!\perp Z(pa_Z) \mid B$$

is valid. However, for the same reason the implication:

$$A \perp\!\!\!\perp B \mid U_Z \quad \Rightarrow \quad A \perp\!\!\!\perp B \mid Z(pa_Z)$$

is not generally valid even though Pearl's statement might suggest otherwise. Note all these difficulties result from the deterministic relations between U_Z and $Z(x)$.

Finally we note that the erroneous independence relation $Y(x) \perp\!\!\!\perp X \mid \{Y(z), Z(x), Y\}$ in (Pearl, 2000, p.215) quoted earlier is corrected in the second edition.

However we now show that the correction itself highlights additional problems that arise in evaluating counterfactual independence relations through multi-networks. Specifically, the above independence reappears as equation (11.40) on (Pearl, 2009, p.394), in connection with the 'triple-network' graph shown in Figure 29(iii). A reader asks whether or not it follows from this triple network that in fact $Y(x) \not\perp\!\!\!\perp X \mid \{Y(z), Z(x), Y\}$ as there is an open d-connecting path:

$$Y(x) \leftarrow U_Y \rightarrow Y \leftarrow Z \leftarrow X.$$

Pearl accurately informs the reader that they are "correct" in their conclusion, since $Y(x) \perp\!\!\!\perp X \mid \{Y(z), Z(x), Y\}$ is not implied under the NPSEM-IE. [96] However, Pearl does not give a justification for drawing this conclusion.

[96] However, Pearl does not offer an explanation as to why he reached the opposite mistaken conclusion in the first edition, even though the argument that led to this false conclusion is unchanged in the second edition (Pearl, 2009), save for the addition of the qualifier 'one-way' to 'proxy' on p. 214.

Presumably the reader is applying to the query $Y(x) \perp\!\!\!\perp X \mid \{Y(z), Z(x), Y\}$, the logic applied by Pearl to evaluate the query $Y(x) \perp\!\!\!\perp X \mid Z$ in this same example. To wit Pearl states in the context of Figure 29(ii):

> To test whether $Y(x) \perp\!\!\!\perp X \mid Z$ holds in the original [NPSEM-IE] model, we *test* whether Z d-separates X from $Y(x)$ in the twin network. [(Pearl, 2000, 2009, p.214); counterfactual notation changed; emphasis added]

Given this quotation, and the reader's question as to whether it is safe to apply d-connection to a triple-network, it might easily be inferred that Pearl was endorsing not only the reader's conclusion, but also the method (d-connection) by which the reader arrived at it.

However, a close reading of Pearl's response to another question from the same reader shows that in fact this is not the case. The reader correctly points out that if it were true in general that the presence of a d-connecting path in a twin-network implied that the corresponding counterfactual independence did not hold, then an inference in (Pearl, 2001b) would be invalid; (see Pearl, 2009, p.394). This is because there may be deterministic equalities between variables that are present in the network, even though these are not explicitly represented.[97] Consequently d-separation is not a complete criterion for establishing conditional independence in multi-networks.

Pearl agrees that in a multi-network d-connection does not imply dependence, showing that indeed there is a hidden equality in the twin-network associated with the example from (Pearl, 2001b) raised by the reader. He then refers the reader to the paper (Shpitser and Pearl, 2007), that introduces 'counterfactual graphs' as a means of systematically handling these equalities. However, 'counterfactual graphs' are intended to represent conjunctions of counterfactual *events*, e.g. $Z(x = 1) = 0$, not counterfactual *variables*, e.g. $Z(x = 1)$. (Shpitser and Pearl, 2007) contains a complete algorithm for testing whether a counterfactual quantity is identified (under the NPSEM-IE). Though the paper does not include an explicit algorithm for testing independence, the **make-cg** algorithm suggests an obvious procedure for checking counterfactual independence among events (Shpitser,

[97] To give a simple example of this phenomenon, consider the graph shown in in Figure 30(i) together with the twin-network (ii) resulting from intervening to set Z to z. We see that in the twin-network T and $Y(z)$ are d-connected given X by the path $T \leftarrow U_T \rightarrow T(z) \rightarrow X(z) \rightarrow Y(z)$, but in spite of this $T \perp\!\!\!\perp Y(z) \mid X$ under the associated NPSEM-IE. This is because, since Z is not an ancestor X, $X(z) = X$, and $T(z)$ and $Y(z)$ are d-separated given $X(z)$ in the twin-network. This independence also holds under the FFRCISTG model associated with this graph, as may be seen directly from the template $\mathcal{G}(z)$ shown in Figure 30(iii).

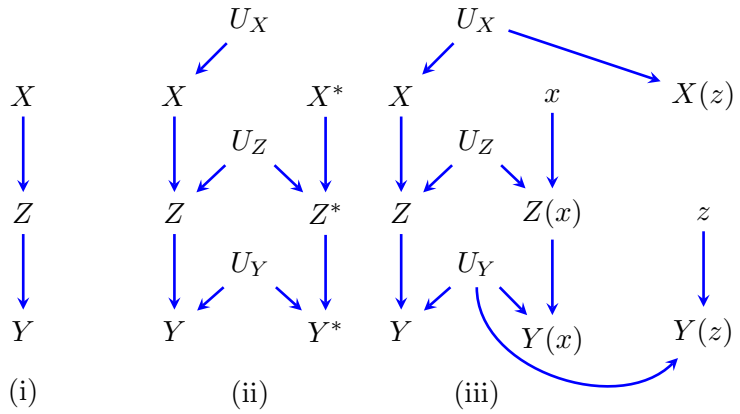


Figure 29: Twin networks. (i) A DAG; (ii) Twin network intervening on X; see Pearl (2000, 2009), p. 214, Fig. 7.3; note that in counterfactual notation $X^* \equiv x$, $Z^* \equiv Z(x)$ and $Y^* \equiv Y(x)$; (iii) 'Triple' network intervening on X and Z; see Pearl (2000, 2009), p. 394, Fig. 11.18. Compare to the corresponding SWIGs in Figure 10.

2013). Specifically, to test:

$$\{\mathbb{X}(\mathbf{a}^\dagger) = \mathbf{x}'\} \perp\!\!\!\perp \{\mathbb{Y}(\mathbf{b}^\dagger) = \mathbf{y}'\} \mid \mathbb{Z}(\mathbf{c}^\dagger) = \mathbf{z}'\}$$

one may run **make-cg** on the set of events $\{\mathbb{X}(\mathbf{a}^\dagger) = \mathbf{x}', \mathbb{Y}(\mathbf{b}^\dagger) = \mathbf{y}', \mathbb{Z}(\mathbf{c}^\dagger) = \mathbf{z}'\}$ and then test d-separation in the resulting counterfactual graph. It is conjectured that this procedure is complete (Shpitser, 2013). However, since a counterfactual independence amongst variables corresponds to an exponential number of independences among events, testing an independence amongst variables using this method would necessitate the construction of exponentially many separate counterfactual graphs.[98]

D.4 Context Specific Independence and a Further Error

In spite of acknowledging that it is unsafe to draw inferences from d-connection in twin-networks,[99] in Pearl (2009, p.353, Ex. 11.3.3) he falls afoul of exactly this pitfall in his second edition. Specifically, in the twin network shown in

[98] There is currently no known polynomial-time algorithm for testing counterfactual independence in NPSEM-IEs.

[99] Without the additional steps required to construct a counterfactual graph.

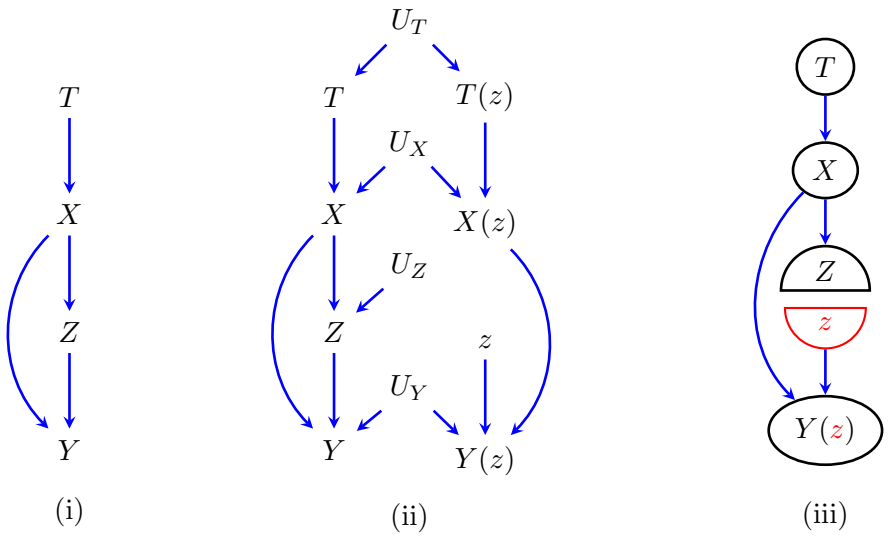


Figure 30: (i) A DAG $\mathcal{G}$. (ii) The twin-network arising from intervening to set Z to z. This example shows the difficulty arising from deterministic relations: $T(z)$ and $Y(z)$ are d-connected given X in the twin-network, but in spite of this $T(z) \perp\!\!\!\perp Y(z) \mid X$ under the associated NPSEM-IE because $X(z) = X$, and $T(z)$ and $Y(z)$ are d-separated given $X(z)$ in the twin-network. (iii) The template $\mathcal{G}(z)$ obtained by node-splitting, which makes manifest that $T = T(z)$ is d-separated from $Y(z)$ given X.

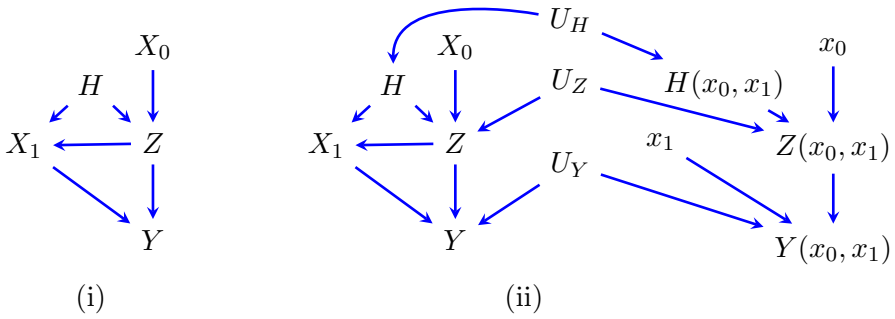


Figure 31: Failure of the simple twin network method of Pearl (2000, 2009) in Pearl's Example 11.3.3. (i) The original DAG; (ii) The twin network after intervention on X_0 and X_1. The twin network fails to reveal that $Y(x_0, x_1) \perp\!\!\!\perp X_1 \mid Z, X_0 = x_0$. This 'extra' independence holds 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 $Y(x_0, x_1) \not\perp\!\!\!\perp X_1 \mid Z, X_0 \neq x_0$. (Note that we have replaced the bi-directed edge $X_1 \leftrightarrow Z$ present in Pearl's figure, by a hidden variable $X_1 \leftarrow H \rightarrow Z$, this does not change the conclusions, but does make the construction of the twin network more transparent.)

Figure 31 he incorrectly concludes from the existence of the path:

$$X_1 \leftarrow H \rightarrow Z \leftarrow U_Z \rightarrow Z(x_0, x_1) \rightarrow Y(x_0, x_1)$$

that $Y(x_0, x_1) \not\perp\!\!\!\perp X_1 \,|\, Z, X_0 = x_0$. This fails to take into account the fact that conditional on $X_0 = x_0$ we have the additional equality: $Z = Z(x_0) = Z(x_0, x_1)$; hence the above path fails to result in dependence because it is blocked by $Z(x_0, x_1)$. We note that careful application of the **make-cg** algorithm described in (Shpitser and Pearl, 2007, 2008) would have performed a pre-processing step in which Z and $Z(x_0, x_1)$ would have been identified (conditional on $X_0 = x_0$), and thus would have avoided the problem. (The criterion of 'D-separation' of Geiger (1990), which allows for determinism would also have addressed this problem. Note here that 'D-separation' and 'd-separation' are not the same.)

As noted earlier, we see some irony in Pearl's error in this example: in (Pearl, 2009, Ex. 11.3.3) the example was supposed to illustrate an unnoticed shortcoming of the condition given by Robins for the application of the g-formula. As stated earlier, our approach based on SWIGs immunizes the user from making this mistake.

Summary

Pearl (2000, 2009) presents the NPSEM-IE as a natural way to unify graphs and counterfactuals with the twin-network method playing the role of inferential engine. However, as we have seen under close examination twin-networks do not provide a general method for testing counterfactual independence relations implied by NPSEM-IEs, since the presence of deterministic relations means that d-separation is not complete (since d-connection does not ensure that independence does not always hold). This fact led Pearl to make errors in both editions of his book. The counterfactual graphs introduced in (Shpitser and Pearl, 2007, 2008) systematically address many of the shortcomings present in twin-networks. However, there is currently no known polynomial time algorithm for testing counterfactual conditional independence amongst *variables* in NPSEM-IEs.

Though we think we have identified some weaknesses in the twin-network method as it appears in (Pearl, 2000, 2009), we do not wish to suggest that this somehow taints the other graphical tools and methods that Pearl has advocated, extended or developed.

We are in strong agreement with Pearl's basic contention that directed graphs are a very valuable tool for reasoning about causality, and by extension, potential outcomes. If anything our criticism of the approach to synthesizing graphs and counterfactuals in (Pearl, 2000, 2009) is that it is not 'graphical enough' in the sense that, as we have seen, since twin networks involve variables that are deterministically related, d-separation is no longer a complete criterion for deriving conditional independence relations.