From inference to influence: applying causal game theory to complex security environments

Vonk, M.C.

Chapter 3

Causality and Assumptions

To effectively apply causal research within complex security environments, it is essential to understand both the available causal concepts and the specific types of causal claims they support. However, causal inference research remains fragmented across multiple scientific disciplines, often leading to conceptual silos and inconsistent terminology. Synthesizing these disparate components is therefore necessary to build a coherent foundation for causal reasoning.

This chapter addresses that need by systematically organizing causal inference concepts within Pearl's causal hierarchy [27], which provides a formal structure for distinguishing three levels of causal reasoning: association, intervention, and counterfactual. Beyond mapping methods to this hierarchy, it also clarifies the assumptions required for rigorously applying causal models in practice and offers guidance for practitioners seeking to implement them effectively.

By adopting this structured approach, the chapter directly engages with the first research question, RQ1.1: *What fundamental causal concepts are necessary for structuring and differentiating causal relationships, particularly in the context of Pearl's causal hierarchy?* In addition, it explores RQ1.2: *What key assumptions underpin causal inference applications across Pearl's hierarchy?*

This dual focus equips practitioners with the foundational causal concepts needed to formulate meaningful and well-structured causal claims, while simultaneously cultivating a deeper understanding of the methodological commitments and assumptions required to substantiate them. The chapter's discussion closely follows a previously published article [251].

3.1 Introduction

Historically, the fundamental problem of causal inference, which arises from the impossibility of observing both the treated and control outcomes for the same unit, made it difficult for researchers to establish causal claims [102]. To overcome this challenge, randomized controlled trials (RCTs) became the gold standard for identifying causal effects, as the random assignment of units to treatment or control effectively eliminated confounding between assignment and outcome. However, in many contexts (e.g., studying the effects of smoking), it is either unethical or impractical to randomly assign individuals to treatment conditions. As a result, researchers must rely on *observational data* and alternative approaches to draw causal claims.

In this chapter, causal concepts that emerge from reasoning with causality using observational data within the context of *probabilistic graphical models* (PGMs) are explored. The latter are graphical representations that can be learned from observational data through causal discovery, an algorithmic approach to inferring the causal structure among variables. The next step, causal identification, determines whether a causal effect can be estimated from the available observational data, and if so, employs specific calculi to express causal queries in terms of known quantities, ensuring they have a unique solution [177]. With additional assumptions, one can perform causal inference, which involves estimating an outcome variable under hypothetical interventions. The focus is on the assumptions required at each stage, including causal discovery, identification, and inference, and on the different causal concepts that emerge from these processes.

Table 3.1: Pearl's Causal Hierarchy Queries

Level	Action	Query	Example
1. Associational $P(Y \mid x)$	Seeing	How does observing $X = x$ influence Y?	Do smokers generally tend to have more lung cancer than non-smokers?
2. Interventional $P(Y \mid do(x), z)$	Doing	How does intervening on $X = x$ affect Y given $Z = z$?	Is there a causal effect of smoking on lung cancer?
3. Counterfactual $P(Y(x) \mid x', y')$	Imagining	What would have been Y under $X = x$ given that $Y = y'$ is observed under $X = x'$?	Would a patient have lung cancer if he/she had smoked given that the patient does not have lung cancer and has never smoked?

Causal identification and causal inference can be further categorized into three levels of increasing complexity, known as Pearl's causal hierarchy [27]. These levels correspond to different types of queries: *associational* (seeing), *interventional* (doing), and *counterfactual* (imagining) [179]. Table 3.1 provides an overview of these queries, while Figure 3.1, adapted from Bareinboim et al. [27], presents a structured depiction of the key concepts at each level of Pearl's causal hierarchy, with higher levels shown towards the top of the figure to highlight increasing complexity. It has been proven by the *causal hierarchy theorem* [27] that queries at higher levels of the hierarchy can generally not be addressed with information of lower levels only.

To navigate this hierarchy, Section 3.2 introduces the *potential outcome framework* (POF), providing a useful perspective for analyzing key causal concepts and their foundational assumptions [202]. This is followed by a more formal, yet logically equivalent, approach that adopts a distinct notational and conceptual perspective: the *structural causal model* (SCM) [178]. Both serve as foundational frameworks for addressing causal queries across all three levels. Building on the SCM, its natural extension is introduced, the *spatially equivalent structural equation models* (SESEM), which allows for modeling in environments where interference may be present [137]. Then, Bayesian networks, d-separation, and some equivalent Markov assumptions at the associational level of the hierarchy are introduced in Section 3.3. Concepts and assumptions at the interventional level of the hierarchy will be introduced in Section 3.4. In Section 3.4.1, different sets of assumptions that allow non-parametric as well as parametric causal discovery are introduced. Subsequent Section 3.4.2 delineates different assumptions and concepts for non-parametric as well as parametric identi-

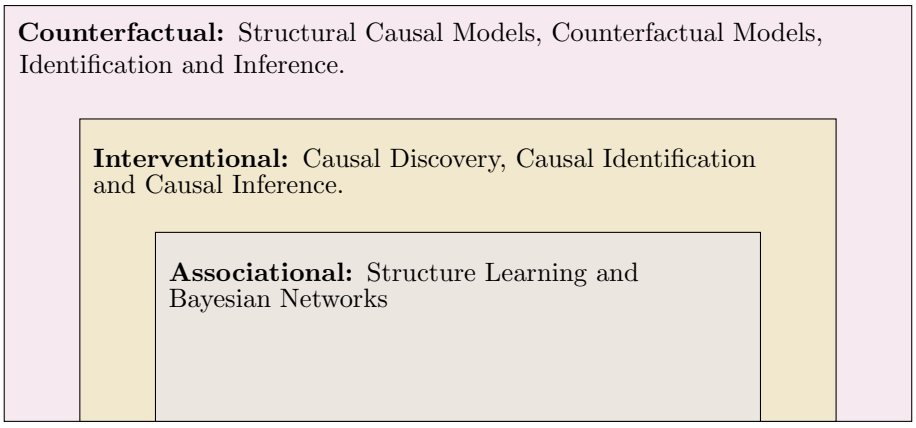

Figure 3.1: Pearl's causal hierarchy of causal concepts.

fication and inference approaches while enunciating the meeting point between the two approaches. The chapter proceeds with the introduction of various counterfactual models and inference techniques to enable reasoning at the counterfactual level of the hierarchy in Section 3.5. Finally, concluding remarks are presented in Section 3.6.

3.2 Potential Outcome Framework

In this section, the potential outcome framework (or Neyman-Rubin causal model) as developed by Rubin [202] is introduced. The potential outcomes ground the most granular sort of queries of the causal hierarchy, the counterfactual, and the framework incorporates the core assumptions of causal inference. That means that claims about potential outcomes are equivalent to counterfactual claims. The necessary methods and targets of interest will be defined along with accessory assumptions. For a full picture of these methods and assumptions, the reader is referred to Figure 3.2. This section naturally revolves around the concept of *potential outcomes*.

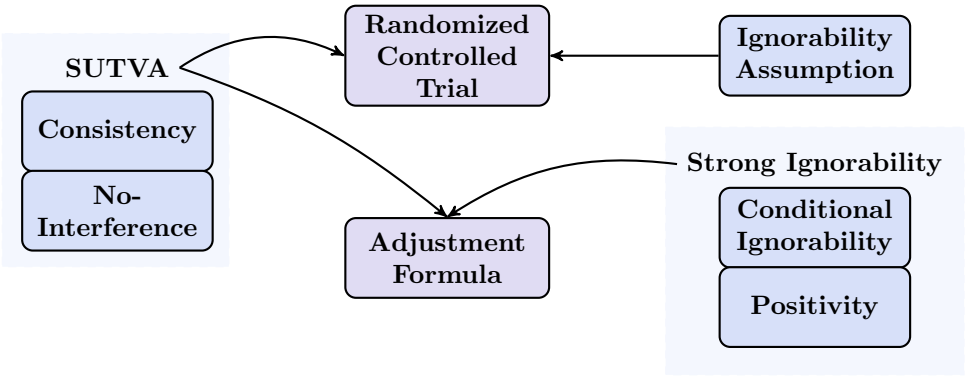

Figure 3.2: Methods for inferring causal claims under different assumptions. Boxes represent assumptions (individual or grouped); arrows indicate that satisfying the source enables the use of the target method. The ignorability assumption and stable unit-treatment value assumption (SUTVA) are implicit in randomized controlled trials for which causal claims can be made. When strong ignorability holds together with SUTVA, the adjustment formula should be invoked to calculate causal estimates. These concepts are further introduced in the following sections.

3.2.1 Potential Outcomes

Before the potential outcomes are introduced, first the treatment will be defined:

Definition 3.1 (Treatment Variable)**.** The treatment variable T is a random variable that takes on different values for treatment t.[1]

Definition 3.2 (Potential Outcome)**.** The potential outcome random variables are denoted by $Y(T = t)$ (or $Y(t)$ in short) for different treatment values $T = t$. For a unit of observation i (or unit in short) and treatment value t, the potential outcome realizations are denoted by y_i^t, the outcome that would have been observed if unit i had been exposed to treatment t.

Classically, t has been considered to take on binary values corresponding to treatment (1) and control (0) [202]. The first target of interest emerges naturally from this definition and is called the *unit-level causal effect*.

Definition 3.3 (Unit-Level Causal Effect)**.** Considering binary treatment t, the unit-level causal effect for unit i is defined as $\tau_i = y_i^1 - y_i^0$.

The potential outcome of unit i cannot be observed for treatment $t = 1$ and control $t = 0$ in a single observation, leading to the fundamental problem of causal inference [102]. This means that the unit-level causal effect cannot be calculated exactly but only estimated. y_i^t is called *counterfactual* when unit i has not been exposed to treatment t but to another treatment value $t' \neq t$. The unit-level causal effect also has its statistical population counterpart, the *average treatment effect*.

Definition 3.4 (Average Treatment Effect (ATE))**.** For binary treatment $t \in \{0, 1\}$, the average treatment effect is defined as

$$\tau = \mathbb{E}[Y(T = 1) - Y(T = 0)]. \tag{3.1}$$

For a sample of n units, the sample average treatment effect is $\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} (y_i^1 - y_i^0)$.

3.2.2 Randomized Controlled Trials

Randomized controlled trials are widely considered to be the gold standard for estimating average treatment effects, as they inherently satisfy three key assumptions. To state these assumptions, the observed outcome for unit i is introduced, denoted by y_i, which is the outcome actually measured after treatment assignment. Each unit reveals only one potential outcome; the one corresponding to the treatment actually received.

The first assumption is called *consistency*:

[1] Although the treatment variable can be multi-dimensional, the fundamental causal concepts are most clearly introduced using a single treatment variable. This concept is generalized in Section 3.5.

Assumption 1 (Consistency)**.** For each unit i that receives treatment t_i, the observed outcome equals the potential outcome under that treatment:

$$y_i = y_i^{t_i}.$$

Equivalently, at the random variable level: if $T = t$, then $Y = Y(t)$.

Informally, the assumption forces one to unambiguously define treatment and tie the potential outcomes to the observed variables. Although this assumption is sometimes known as the *no-multiple-treatment* assumption, some researchers draw a firm distinction between the two [245]. Consistency can be a strong assumption in the observational setting, but it is implicit in randomized controlled trials, because exposure to treatment is a result of experimental design [51].

The second assumption is known as the *no-interference* assumption [54]. It explicitly states that a potential outcome of a unit is not dependent on treatment received by other units. More formally,

Assumption 2 (No-Interference)**.** Let t_i be the treatment assignment of unit i for $i = 1, \ldots, n$. Then no-interference is satisfied if

$$Y_i(t_1, \ldots, t_n) = Y_i(t_i).$$

Interference is also known as *spillover*. In a randomized controlled trial, the investigator can prevent causal spillover by designing the experiment such that different units do not interact.

A combination of both consistency and no-interference leads to the *stable unit-treatment value assumption* (SUTVA) [203]. As interference is hard to restrain in the observational setting, a formal framework capable of addressing violations of interference assumptions will be introduced when presenting structural causal models in the next section. Although a randomized controlled trial poses limitations on SUTVA violations, the strength of the randomized controlled trial lies in its implication of the *ignorability* assumption:

Assumption 3 (Ignorability/Exchangeability)**.** Consider binary treatment assignment random variable T and potential outcome under treatment $Y(1)$ and control $Y(0)$. Then, ignorability is satisfied if

$$Y(0), Y(1) \perp\!\!\!\perp_P T,$$

where $\perp\!\!\!\perp_P$ means independence in probability.

In words, the potential outcomes under treatment are independent of treatment assignment. In this case, it can be ignored how units ended up in the treatment or control group. Equivalently, the group that received treatment could have been exchanged with the group receiving control, resulting in the same potential outcome.

The three assumptions together constitute the randomized controlled trial (as illustrated in Figure 3.2) and make calculation of the average treatment effect possible by means of reasoning with potential outcomes. Besides the use of potential outcomes, the potential outcome framework contains one additional element that enables one to bypass the fundamental problem of causal inference beyond randomized controlled trials, which is the assignment mechanism [118].

3.2.3 Beyond Randomized Controlled Trials

Unlike for randomized controlled trials, the ignorability assumption is easily violated when dealing with observational data, because the treatment and control groups are rarely truly exchangeable. A *confounder* can causally influence the treatment variable as well as the outcome variable as illustrated in Figure 3.3. Therefore, more lenient assumptions can be adopted to render the calculation of causal effects under the potential outcome framework still possible in the presence of confounders.

Assumption 4 (Conditional Ignorability)**.** Let $\boldsymbol{Z}$ denote confounding variables. Consider binary treatment assignment random variable T. Then conditional ignorability is satisfied if

$$Y(0), Y(1) \perp\!\!\!\perp_P T \mid \boldsymbol{Z}.$$

That means that the treatment and control group are generally not exchangeable, but they become exchangeable when conditioning on the confounding set. For that reason, *conditional ignorability* is also known as the *unconfoundedness* assumption. It

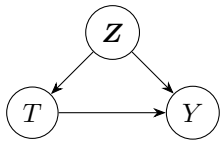

Figure 3.3: Because $\boldsymbol{Z}$ causally influences both T and Y, $\boldsymbol{Z}$ is said to *confound* the relation between T and Y.

is useful to adjust for confounding to reach conditional ignorability as long as each treatment level has non-zero probability within every subgroup defined by the confounders. The *positivity* assumption guarantees this condition holds.

Assumption 5 (Positivity)**.** Let $\boldsymbol{Z}$ denote confounding variables. Then positivity is satisfied if

$$P(T = t \mid \boldsymbol{Z}) \in (0, 1) \quad \forall\, T, \boldsymbol{Z}.$$

There is a tradeoff between conditional ignorability and positivity by virtue of adjusting for covariates [68], which is the process of conditioning on subgroups of the data that share similar covariate values. Intuitively, the more covariates are adjusted for, the smaller the subgroups become. This can lead to subgroups being entirely assigned to either treatment or control, which is a violation of the positivity assumption. Contrary, not sufficiently adjusting for high-dimensional covariates may lead to violations of conditional ignorability assumptions. Section 3.4.2 explains how this problem motivates the use of parametric approaches over non-parametric ones. Both conditional ignorability and positivity together are called *strong ignorability* [200, 119].

Vested with all of the above assumptions, one is able to calculate the average treatment effect. Assume binary treatment assignment variable T and confounding set $\boldsymbol{Z}$:

$$\begin{aligned}
\mathbb{E}[Y(1) - Y(0)] &\overset{(1)}{=} \mathbb{E}_{\boldsymbol{Z}}[\mathbb{E}[Y(1) - Y(0) \mid \boldsymbol{Z}]] \\
&\overset{(2)}{=} \mathbb{E}_{\boldsymbol{Z}}[\mathbb{E}[Y(1) \mid \boldsymbol{Z}] - \mathbb{E}[Y(0) \mid \boldsymbol{Z}]] \\
&\overset{(3)}{=} \mathbb{E}_{\boldsymbol{Z}}[\mathbb{E}[Y(1) \mid T = 1, \boldsymbol{Z}] - \mathbb{E}[Y(0) \mid T = 0, \boldsymbol{Z}]] \\
&\overset{(4)}{=} \mathbb{E}_{\boldsymbol{Z}}[\mathbb{E}[Y \mid T = 1, \boldsymbol{Z}] - \mathbb{E}[Y \mid T = 0, \boldsymbol{Z}]].
\end{aligned}$$

While the first two equalities follow from the laws of probability and expectation, the third equality is a result of conditional ignorability and positivity and the fourth equality a result of consistency. This result is also called the *adjustment formula* and the underlying assumptions are summarized in Figure 3.2. The formula requires one to have insight into the *assignment mechanism*: the conditional probabilities of treatment given covariates and potential outcomes. This is the second element that constitutes the potential outcome framework.

When conditional ignorability does not apply, causal inference becomes significantly harder. In some cases instrumental variables, those that causally influence the

treatment but not the outcome variable, can be utilized [93], the joint distribution of latent and observed confounders can be extracted from variational auto-encoders [147] and network data as a proxy for latent confounders can still be used to substantiate causal effects [89].

Consistency follows from the definitions of the structural causal models and hence the literature rejecting this assumption is not rich [178]. SUTVA can easily be violated by departures from the no-interference assumption. A framework that can account for such violations will be described; first, however, the structural causal model is introduced.

3.2.4 Structural Causal Models

Structural causal models (also known as structural equation models (SEMs)) are logically equivalent to the potential outcome framework [178] but offer a more formally structured notation:

Definition 3.5 (Structural Causal Models (SCM))**.** A structural causal model is a tuple $\mathcal{S} = (\boldsymbol{V}, \boldsymbol{W}, G, \boldsymbol{F})$ where $\boldsymbol{V} = \{V_1, \ldots, V_n\}$ is an ordered set of endogenous variables, $\boldsymbol{W}$ is a set of exogenous variables, G is a directed acyclic graph with vertex set $\boldsymbol{V}$, and $\boldsymbol{F} = \{f_1, \ldots, f_n\}$ is a set of structural functions satisfying:

1. For all $V_i \in \boldsymbol{V}$, there exist a corresponding subset of exogenous variables $\boldsymbol{W}_i \subseteq \boldsymbol{W}$ and a mapping $f_i : \Omega_{\mathbf{pa}(V_i) \cup \boldsymbol{W}_i} \to \Omega_{V_i}$ that maps the state space of endogenous parents of V_i together with $\boldsymbol{W}_i$ to the state space of V_i:

$$v_i = f_i(\boldsymbol{pa}_i, \boldsymbol{w}_i).$$

2. The exogenous variables $\boldsymbol{w}$ are drawn from a probability distribution $P(\boldsymbol{W})$ over the state space $\Omega_{\boldsymbol{W}}$.

SCMs can be either parametric or non-parametric. Non-parametric structural equation models are sometimes invoked because assumptions about functional forms between respective exogenous and endogenous variables are costly. It is important to note that the SCM does not assume the independence of exogenous variables.[2] However, when this additional property is satisfied, the models are known as non-parametric structural equation models with independent errors (NPSEM-ie) as will be

[2]In Definition 3.5 this can be observed from the fact that for $V_i, V_j \in \boldsymbol{V}$, the corresponding $\boldsymbol{W}_i$ and $\boldsymbol{W}_j$ can overlap.

elaborated on in Section 3.5. The SCMs are assumed to be acyclic, also called *recursive*. Recursiveness allows a topological sort to exist over the endogenous variables.

A natural extension of SEMs arises when the no-interference assumption is violated in the context of spatial spillover, where causal spillover occurs when changes in one unit influence outcomes in neighboring units through spatial interdependencies. This extension is known as *spatially explicit structural equation models* (SESEMs) [137]. This method merges the flexibility of structural equation models in describing complex relationships with the capability to explicitly model spatial confoundedness through variance/covariance matrices computed across various lag distances. An application of such a model will be discussed in Chapter 6.

Frequently, the true SEM is unattainable due to a limited ability to observe a system [201], and one has to settle for surrogate models that do not have equal expressive power, but can be sufficient to answer queries of the first two levels of the hierarchy. These surrogate models will be introduced in the following sections.

3.3 Associational Level

This section introduces concepts and associated assumptions necessary to address questions at the first level of Pearl's causal hierarchy, the associational level (see Figure 3.1). The section starts with some preliminaries on the relation between probability distributions and graphical models. It then explains the features of Bayesian networks and introduces Markov random fields. Because structure learning of Bayesian networks closely resembles causal discovery, Section 3.4.1 provides further information on structure learning. An overview of the items covered in this section is presented in Figure 3.4.

3.3.1 Bayesian Networks

In order to address queries at the first level, random variables need to be tied to the graphical components introduced. This is only possible when additional assumptions are invoked. Let $X_1, \ldots, X_n$ be random variables with joint probability distribution $P(x_1, \ldots, x_n)$. In *Bayesian networks* (BNs), the random variables are represented by the nodes of a directed acyclic graph and the probabilistic dependencies are represented by the edges via the *local Markov* assumption:

Assumption 6 (Local Markov)**.** Let $P(x_1, \ldots, x_n)$ be the joint probability distribution of random variables X_i corresponding to nodes $V_i \in \boldsymbol{V}$ in the directed acyclic

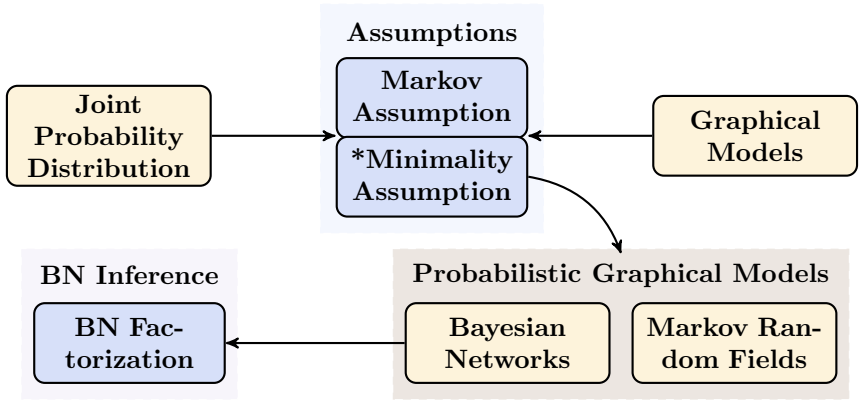

Figure 3.4: Assumptions (blue) and concepts (yellow) discussed at the associational level of the hierarchy. Probability distributions and graphical models can be tied together by means of the Markov assumptions. The minimality assumption can be adopted optionally for a parsimonious encoding of the joint distribution. The resulting object can either be a Bayesian network or a Markov random field. While inference is more extensively discussed in Chapter 5, the Bayesian network factorization assumption plays a central role.

graph $G = (\boldsymbol{V}, \boldsymbol{E})$. Then, the local Markov assumption holds if for every X_i the following holds in the graph:

$$X_i \perp\!\!\!\perp_P \mathbf{nonde}(X_i) \mid \mathbf{pa}(X_i).$$

Since the local Markov assumption ties the random variables together with the graphical structure, $\boldsymbol{V}$ is assumed to inherit all the probabilistic properties from $\boldsymbol{X}$. Henceforth, $P(v_1, \ldots, v_n)$ will be used instead of $P(x_1, \ldots, x_n)$ to denote the probability distribution of the random variables. The use of the underscore P to imply independence in probability is not superfluous as there also exists independence in the graph defined by *d-separation* and denoted by symbol $\perp\!\!\!\perp_G$.

Definition 3.6 (*d*-separation)**.** A path ρ between V_i and V_j (ignoring edge directions) is d-connected in the directed acyclic graph $G = (\boldsymbol{V}, \boldsymbol{E})$ by a set of nodes $\boldsymbol{C} \subseteq \boldsymbol{V} \backslash \{V_i, V_j\}$ if

1. ρ does not contain a chain $\cdots \to Z \to \cdots$ or fork $\cdots \leftarrow Z \to \cdots$, where Z is contained in $\boldsymbol{C}$.

2. all colliders (nodes where two arrows converge, e.g., $\cdots \to Z \leftarrow \cdots$) of the path ρ are in $\boldsymbol{C}$ or have a descendant in $\boldsymbol{C}$.

If there are no d-connecting paths between V_i and V_j given $\boldsymbol{C}$, then V_i and V_j are *d-separated* by $\boldsymbol{C}$ which is denoted by $V_i \perp\!\!\!\perp_G V_j \mid \boldsymbol{C}$.

The concept of graph independencies gives rise to a reformulation of the local Markov assumption to the global Markov assumption:

Assumption 7 (Global Markov)**.** Let $P(v_1, \ldots, v_n)$ be the joint probability distribution of random variables corresponding to the nodes $V_i \in \boldsymbol{V}$. Let $\perp\!\!\!\perp_G$ denote d-separation in the directed acyclic graph $G = (\boldsymbol{V}, \boldsymbol{E})$ and $\perp\!\!\!\perp_P$ independence in distribution. Then the global Markov assumption holds if for all $\boldsymbol{V}_I, \boldsymbol{V}_J, \boldsymbol{V}_K \subseteq \boldsymbol{V}$

$$\boldsymbol{V}_I \perp\!\!\!\perp_G \boldsymbol{V}_J \mid \boldsymbol{V}_K \implies \boldsymbol{V}_I \perp\!\!\!\perp_P \boldsymbol{V}_J \mid \boldsymbol{V}_K.$$

By relating the independencies of the graph to the independencies of the distribution, one can leverage the graphical structure for a parsimonious factorization of the joint probability distribution. This can also be directly assumed.

Assumption 8 (Bayesian Network Factorization)**.** Let $P(v_1, \ldots, v_n)$ be the joint probability distribution of random variables corresponding to the nodes $V_i \in \boldsymbol{V}$ in the directed acyclic graph $G = (\boldsymbol{V}, \boldsymbol{E})$. The Bayesian network factorization assumption holds if the distribution can be factorized according to the corresponding graphical structure:

$$P(v_1, \ldots v_n) = \prod_{i=1}^{n} P(v_i \mid \boldsymbol{pa}_i).$$

Example 1 (Graph Factorization)**.** Consider the Bayesian network displayed by Figure 3.3. According to the Bayesian network factorization assumption, the joint probability distribution $P(\boldsymbol{Z}, Y, T)$ can be factorized to $P(\boldsymbol{Z})P(T \mid \boldsymbol{Z})P(Y \mid \boldsymbol{Z}, T)$.

It has been shown that the local Markov assumption, the global Markov assumption and the Bayesian network factorization are equivalent when positivity is assumed [127]. When any of these equivalent conditions holds, the probability distribution P is said to be *Markov relative* (or Markov[3] in short) to $G = (\boldsymbol{V}, \boldsymbol{E})$.

While the Markov assumption imposes restrictions on the probability distribution via the graphical structure, an additional assumption is necessary to obtain a minimal representation of the probability distribution's conditional independence structure.

[3]The Markov assumption is in specific cases known as the causal Markov assumption. Technically, the assumption is only causal when the concomitant graphical component has causal meaning (which will be introduced in Section 3.4.2).

This assumption comes in various forms of increasing strength: SGS-minimality, P-minimality and faithfulness [263]. P-minimality will be discussed here [178], but before introducing this assumption, the concept of a *preferred* graph needs to be introduced:

Definition 3.7 (Preferred Graph)**.** Let P be the set of distributions that is Markov relative to $G = (\boldsymbol{V}, \boldsymbol{E})$ and $G' = (\boldsymbol{V}, \boldsymbol{E}')$. Then G' is (strictly) preferred to G if the conditional independence relations of G are a (proper) subset of the conditional independence relations of G'.

Assumption 9 (Minimality)**.** Let P be the set of distributions that is Markov relative to $G = (\boldsymbol{V}, \boldsymbol{E})$. Minimality is satisfied with respect to G if P is not Markov relative to a strictly preferred graph $G' = (\boldsymbol{V}, \boldsymbol{E}')$ to G.

Although minimality is a desirable assumption because it allows one to encode the joint distribution in the most parsimonious graphical structure possible, it is not required to answer queries at the first level of the hierarchy.

In concluding this section, not all independence relations are representable by a Bayesian network, as the following counterexample illustrates:

Example 2 (Limits of Bayesian Networks in Encoding Independencies)**.** Let X_1, X_2, X_3, X_4 be random variables. Then, there does not exist a Bayesian network satisfying conditional independence relations $X_1 \perp\!\!\!\perp_P X_2 \mid \{X_3, X_4\}$ and $X_3 \perp\!\!\!\perp_P X_4 \mid \{X_1, X_2\}$.

Therefore, there is another graphical structure that can represent conditional independencies: the *Markov random field*. Unlike Bayesian networks, which use directed edges to encode asymmetric relationships, Markov random fields use undirected edges to represent symmetric associations. This allows them to account for cyclic probability relations and work with potential functions, but prevents them from representing directionality. For more information about Markov random fields, the reader is referred to the work by Koller and Friedman [127]. Both Bayesian networks and Markov random fields are *probabilistic graphical models* as they unify joint probability distributions with graphical structures.

3.4 Interventional Level

This section discusses the causal assumptions and components necessary to address queries at the second level of the hierarchy. It begins with the various sets of assumptions required to conduct parametric as well as non-parametric causal discovery in Section 3.4.1, as specified in Figure 3.5. Section 3.4.2 then demonstrates how the

output of causal discovery, a causal diagram, forms the basis of both a non-parametric and a parametric approach, where the approaches differ based on a different appreciation of the fundamental problem of causal inference. The non-parametric approach adopts assumptions inherent to causal Bayesian networks that enable inference, while the parametric approach emerges by observing that the fundamental problem of causal inference requires estimation by definition. Figure 3.6 shows the specifications of the different concepts and assumptions necessary for causal inference for each of the two approaches. Finally, causal concepts that emerge when deviating from putative assumptions are discussed in Section 3.4.3.

3.4.1 Causal Discovery

This section will discuss causal discovery from the point of view of necessary assumption, expanding on previous assumptive approaches [69]. Technical details will be discussed when they are contingent on the introduced assumptions, but for a broader account of why causal discovery methods fail in the absence of assumptions, the reader is referred to a survey paper by Runge [207]. Although this section can serve as a blueprint for which method to use when certain assumptions are adopted, a more practical guide about the application of causal discovery methods can be found in the work by Malinsky and Danks [155]. While interventional data can significantly improve causal structure learning by resolving directional ambiguities that observational data cannot [96, 224], this section is restricted to recovering the structure with observational data alone. Because observational data alone is available at both the associational and interventional levels of the hierarchy (in the absence of actual interventions), structure learning at these two levels coincides [149]. Additionally, this section is limited to static causal discovery methods, which are causal discovery methods that do not account for the passage of time. There is a body of survey papers on causal discovery methods for longitudinal data and the additional assumptions necessary [17, 207].

An assumption most causal discovery methods revolve around is the *i.i.d.* assumption.

Assumption 10 (Independent and Identically Distributed (i.i.d.))**.** The observational data are independent and identically distributed.

Structure learning is first discussed under the assumptions of causal sufficiency, Markov, faithfulness, acyclicity, and independent and identically distributed data. Subsequently, causal discovery is considered in the presence of violations of the causal

sufficiency assumption, followed by a discussion of relaxations of the faithfulness assumption. Some of these approaches are summarized in Figure 3.5. However, there are assumption sets that allow conducting causal discovery beyond the assumption sets in Figure 3.5. Concepts that emerge when the Markov or the i.i.d. assumptions are violated are discussed in Section 3.4.3.

Because the goal of causal discovery is to recover as much of the graphical structure as possible from observational data, the core assumption within causal discovery should imply features of this underlying structure from the probability distributions (that are learned from the data). The strongest form of that assumption was already touched upon in Section 3.3.1 and is called *faithfulness*:

Assumption 11 (Faithfulness)**.** Let $P(v_1, \ldots, v_n)$ be the joint probability distribution of random variables $V_i \in \boldsymbol{V}$ corresponding to the nodes in the graph $G = (\boldsymbol{V}, \boldsymbol{E})$. Let $\perp\!\!\!\perp_G$ denote d-separation in a graph $G = (\boldsymbol{V}, \boldsymbol{E})$ and $\perp\!\!\!\perp_P$ be the independencies in distribution. Then the probability distribution P is faithful to G if for all $\boldsymbol{V}_I, \boldsymbol{V}_J, \boldsymbol{V}_K \subseteq \boldsymbol{V}$:

$$\boldsymbol{V}_I \perp\!\!\!\perp_P \boldsymbol{V}_J \mid \boldsymbol{V}_K \implies \boldsymbol{V}_I \perp\!\!\!\perp_G \boldsymbol{V}_J \mid \boldsymbol{V}_K.$$

A probability distribution can be faithful to a graph that is acyclic. If this is the case, then the *acyclicity* assumption holds in addition to faithfulness. Practitioners who adopt faithfulness are not necessarily expected to have access to the full probability distributions, but are equipped with appropriate independence tests to find (conditional) independencies in the data. In order to complete the first collection of assumptions necessary to conduct causal discovery, the *causal sufficiency* assumption is highlighted:

Assumption 12 (Causal Sufficiency)**.** A set of variables $\boldsymbol{V}$ is causally sufficient if there are no unobserved confounders, meaning $\boldsymbol{V}$ contains all common causes of any two or more variables in $\boldsymbol{V}$.

When causal sufficiency is assumed, the object to be construed is the directed acyclic graph that best fits the data generating process of the observational data. Because observational data can only identify conditional independence structures, multiple distinct graphical structures may be consistent with the same observational distribution. To represent this uncertainty, we introduce the following definition:

Definition 3.8 (Completed Partially Directed Acyclic Graph)**.** Directed acyclic graphs that entail the same conditional independencies are said to be in the same *Markov*

equivalence class for DAGs. The Markov equivalence class for DAGs is represented by a *completed partially directed acyclic graph* (CPDAG) for which an edge is directed if all directed acyclic graphs in the Markov equivalence class agree on the direction of the edge and undirected otherwise.

The causal sufficiency, Markov, faithfulness, acyclicity and i.i.d. assumptions make up the first assumption set that allows causal discovery.

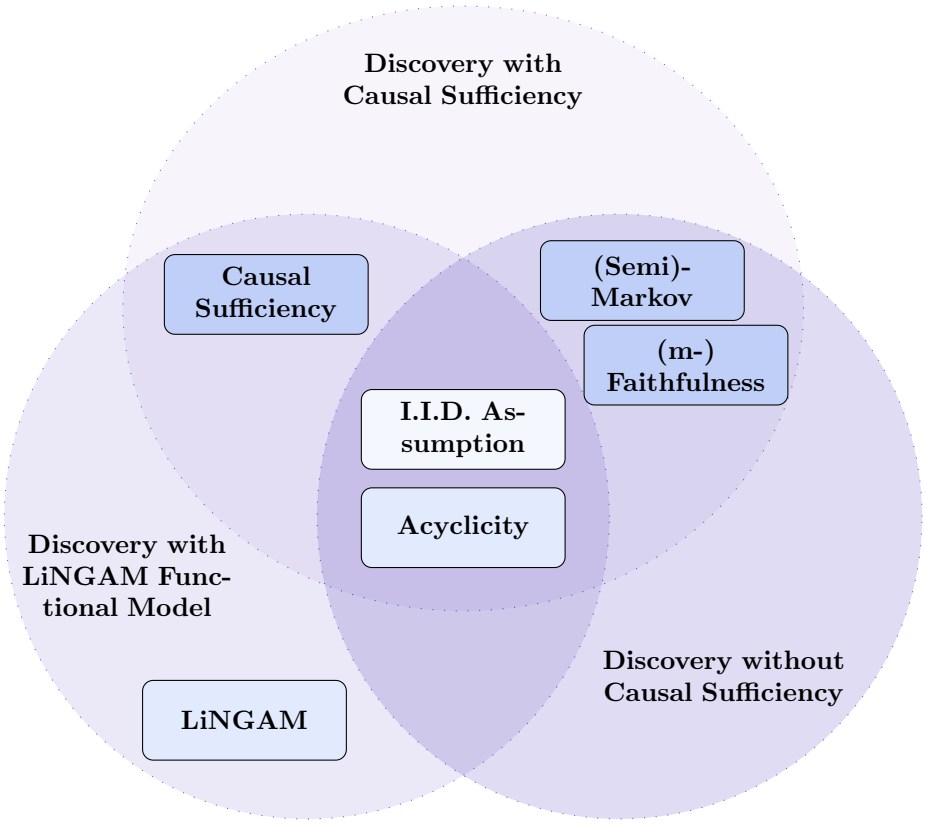

Figure 3.5: Causal discovery assumption sets: the different purple circles represent possible sets of assumptions described in Section 3.4.1 under which causal discovery can be conducted.[4] The boxes represent the assumptions necessary for causal discovery, which may have overlap with multiple assumption sets. The color of the boxes indicates the nature of the assumptions: while light blue represents sampling assumptions, ivory blue indicates assumptions on the data generating process, and darker blue is used for causal assumptions.

Causal discovery with causal sufficiency

Vested with this collection of assumptions as illustrated in the top circle of Figure 3.5, the structure of the underlying data generating process could be investigated with observational data alone. The first algorithm was the *Spirtes, Glymour and Scheines* algorithm [230], closely followed by the *Peter-Clarke* algorithm [229]. Both are *constraint-based* methods, meaning they aim to exploit the conditional independencies to inform the structure of the graph. This means that they require the use of reliable conditional independence testing methods. The algorithms output the CPDAG based on observational data.

Besides constraint-based methods, there are also *score-based* methods. Score-based methods employ the same assumptions, take in the same input, and generate the same output as constraint-based methods, but work fundamentally differently. The methods start with a specific CPDAG and fit it to the data. The fit is scored based on a scoring system and compared to the score of a slightly different CPDAG. The best fit is kept, and the algorithm continues in the same way. In order to restrain the enormous search space, they often have a forward and a backward phase. The forward phase keeps adding edges, which improves the score the most. When no edges can be added that can improve the score, the backward phase starts removing edges that improve the score the most. If there is no edge that can be removed to improve the score, the algorithm ends. Score-based methods require the use of the appropriate loss function based on the nature of the data. Common choices include the Bayesian information criterion, which balances model fit with complexity. An example of such a score-based method that has been proven to work well in simulation studies of small sample sizes is *greedy equivalent search* (GES) [48, 155].

Causal discovery without causal sufficiency

The causal sufficiency assumption can be relaxed. In this case, the possibility of missing common causes in the observational data is acknowledged, and the target of interest is expected to account for unobserved confounders. The smallest superclass of DAGs that accounts for the presence of unobserved confounders and is closed under marginalization is a *maximal ancestral graph* [193]. Similar to how multiple DAGs can encode the same independence constraints, multiple maximal ancestral graphs can also

[4]The list of assumption sets is not exhaustive as more possible assumption sets will be described that allow conducting causal discovery. Although constraint-based and score-based causal discovery algorithms require the use of appropriate conditional independence tests and scoring methods respectively, these are not mentioned as assumptions because they are algorithm-specific.

represent the same conditional independencies. This gives rise to the *partial ancestral graph* that represents the Markov equivalence class of maximal ancestral graphs with the same independence constraints.

It is important to note that the existence of unobserved confounding also leads to a slightly modified version of d-separation that represents conditional independencies with respect to the maximal ancestral graph, called *m-separation.* This leads to natural extensions of the Markov assumption and the faithfulness assumption that go by the *semi-Markov* assumption and *m-faithfulness.*

Algorithms that can extract the partial ancestral graph from observational data such as *fast causal inference* [231], *greedy fast causal inference* [171] and *really fast causal inference* [52] rely on the i.i.d. assumption, the semi-Markov assumption and the m-faithfulness assumption to an acyclic system as illustrated in the bottom right circle of Figure 3.5.

There are two main drawbacks with the algorithms introduced so far. First, either traditional faithfulness or its extension to unobserved confounder models (m-faithfulness) is assumed. Faithfulness is a strong assumption, and it can be easy to find examples where faithfulness is violated [7]. Second, the output of all introduced algorithms entails a representation of a Markov equivalence class. In order to exploit the obtained graphical structure for inference purposes, additional assumptions on the data generating process should be adopted to direct the edges in the graphical structure, which the algorithm could not provide. Both drawbacks can be skirted by assuming restrictions on the data generating process beforehand. This will be discussed in the next section.

Parametric causal discovery and relaxations of faithfulness

In Pearl's causal hierarchy, the true object of investigation is the structural causal model. Because the true SCM is almost always unattainable, one is forced to settle for a surrogate model for which at least questions of lower levels of the hierarchy can be addressed. However, by taking parametric assumptions on the distribution of the underlying SCM, other assumptions can be bypassed.

These methods are based on *functional causal models*, which are equivalent to earlier introduced SCMs [86], where one writes the dependent variable as a function of its parents and a noise term. A special case of a functional causal model is *linear non-Gaussian acyclic model* (LiNGAM) and is defined as follows:

Assumption 13 (LiNGAM)**.** A SCM $\mathcal{S}$ with an ordered set of endogenous variables

$\boldsymbol{V} = \{V_1, \ldots, V_n\}$, exogenous variables $\boldsymbol{W} = \{W_1, \ldots, W_n\}$ and a set of functions $\boldsymbol{F} = \{f_1, \ldots, f_n\}$ is assumed to be a linear non-Gaussian acyclic model if:

1. Every endogenous variable v_i is a linear combination of its parents in the topological sort and exogenous variable term w_i:

$$v_i = f_i(\boldsymbol{pa}_i, w_i) = \sum_{j:V_j \in \mathbf{pa}(V_i)} b_{ij} v_j + w_i.$$

2. The error terms w_i are drawn from exogenous variables $W_i \in \boldsymbol{W}$, which are continuous, mutually independent, and follow a non-Gaussian distribution.

When LiNGAM is assumed, methods exist to fully recover the DAG based on independent component analysis (ICA-LiNGAM) [218]. Faithfulness can be dropped, but causal sufficiency, acyclity and the i.i.d. assumptions should be adopted. The assumption set has been summarized in Figure 3.5. Complementary LiNGAM discovery methods were further developed to account for the violation of causal sufficiency [110]. In addition, there are also variants that allow for a violation of the acyclicity assumption [134].

There are also alternative assumptions (to LiNGAM) on the data generating process that can be used to sideline the faithfulness assumptions and retrieve the full DAG. Some of those assume an additive noise data generating process [109, 185]. More general methods assume a post-linear form [266], where it has been proven that in all but 5 model specification cases the causal direction is identifiable. Even though faithfulness does not have to be assumed in some cases, less restrictive assumptions do have to be adopted [185].

If one is not willing to commit to additional assumptions about the data generating process, but still considers faithfulness too strong of an assumption, one can adopt one of the many weaker versions of faithfulness [265], such as adjacency faithfulness [231, 189], 2-adjacency faithfulness [157] and frugality [74] for which causal discovery algorithms exist or could be developed.

Finally, research has shown that causal discovery algorithms can be unstable [126] or that only limited parts of the graphical structure can be discovered from pure observations [184]. For this purpose, domain knowledge can be used to refine the performance of causal discovery algorithms and has been incorporated via tiered background knowledge [10], user interactions [152], or the penalization of the search process [94]. It is recommended that practitioners assess the possibility of incorporating domain knowledge or experts to enhance the quality of the obtained graphical structure.

3.4.2 Identification and Inference

This section discusses how the concepts from causal discovery can be used for parametric as well as non-parametric inference. While the debate regarding the extent to which the result of causal discovery can be termed 'causal' is acknowledged [62], this section assumes that the ADMGs and DAGs convey causal meaning, making them *causal diagrams*. It first addresses how non-parametric causal inference has contributed to causal inference and emphasizes its assumptions. Subsequently, the assumptions adopted by the parametric approach to causal inference are described, along with points of convergence between the two approaches. Both approaches are summarized in Figure 3.6.

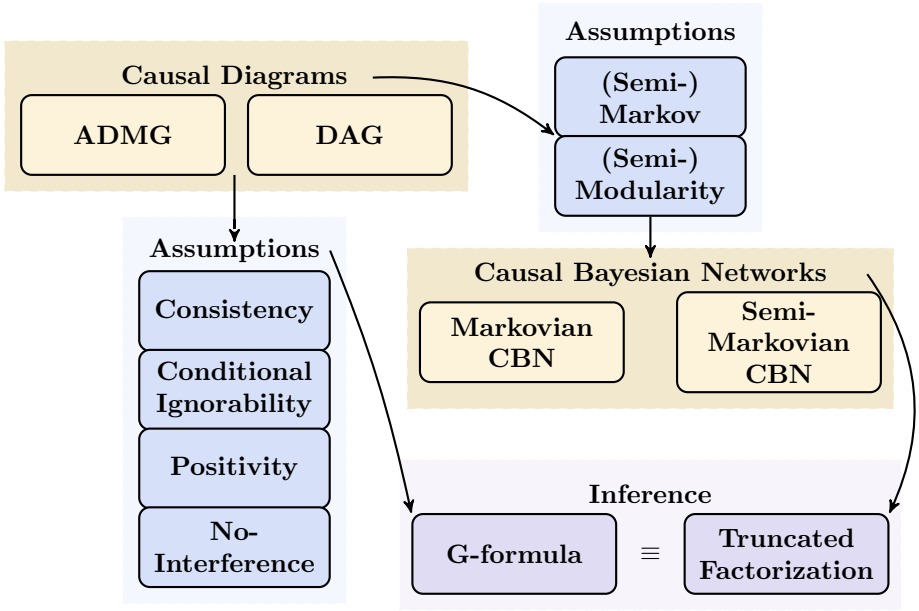

Figure 3.6: Causal diagrams are the basis for causal inference. They can be endowed with assumptions from Section 3.2 to allow inferring causal statements under the g-formula. Alternatively, the diagrams can be subjected to non-parametric assumptions to obtain causal Bayesian networks, which can be leveraged for inference with the truncated factorization formula.

Non-parametric causal inference

In order to infer causal statements, the causal meaning should be specified on top of the Bayesian networks that were introduced earlier. This leads to the definition of

causal Bayesian networks. The 'missing link' definition, as described by Bareinboim et al. [26], is adopted among multiple equivalent definitions of causal Bayesian networks, as it intuitively implicates the (SGS-)minimality assumption. The assumptions inherent to the definitions are examined. Central to this notation are (atomic) interventions; therefore, the *do*-operator and the associated interventional distribution are introduced.

Definition 3.9 (Interventional Distribution)**.** Let Y be a random variable and $\boldsymbol{S} \subset \boldsymbol{V}$ be a set of random variables. The interventional distribution $P(y \mid do(\boldsymbol{S} = \boldsymbol{s}))$ encodes the probability that $Y = y$ given that $\boldsymbol{S}$ is forced to take value $\boldsymbol{s}$ (denoted by the *do-operator* $do(\boldsymbol{S} = \boldsymbol{s})$, or $do(\boldsymbol{s})$ in short) with probability 1.

Bayesian networks that do not contain latent variables are first considered; these are referred to as *Markovian*.

Markovian causal Bayesian networks: The behavior of the do-operator within a Bayesian network can be assumed by the modularity assumption.

Assumption 14 (Modularity)**.** Let P be a probability distribution Markov relative to Bayesian network $G = (\boldsymbol{V}, \boldsymbol{E})$ and let $\boldsymbol{S} \subseteq \boldsymbol{V}$. Then an intervention $do(\boldsymbol{S} = \boldsymbol{s})$ is said to be modular if:

1. For every $V_i \in \boldsymbol{V} \backslash \boldsymbol{S}$, where $\boldsymbol{S}$ and $\mathbf{pa}(V_i)$ are disjoint in G, the interventional distribution by intervening on the parents of V_i is invariant to other interventions in the graph:

$$P(v_i \mid do(\boldsymbol{S} = \boldsymbol{s}), do(\mathbf{pa}(V_i) = \boldsymbol{pa}_i)) = P(v_i \mid do(\mathbf{pa}(V_i) = \boldsymbol{pa}_i)).$$

2. For every $V_i \in \boldsymbol{V}$, the interventional distribution by intervening on the parents of V_i yields the same distribution as observing the parents of V_i:

$$P(v_i \mid do(\boldsymbol{S} = \boldsymbol{s}), do(\mathbf{pa}(V_i) = \boldsymbol{pa}_i)) = P(v_i \mid do(\boldsymbol{S} = \boldsymbol{s}), \mathbf{pa}(V_i) = \boldsymbol{pa}_i).$$

Modularity specifies how the interventional distributions operate within the context of a Bayesian network. A *causal Bayesian network* can now be defined:

Definition 3.10 ((Markovian) Causal Bayesian Network (CBN))**.** Let P be a probability distribution Markov relative to Bayesian network $G = (\boldsymbol{V}, \boldsymbol{E})$. Then $G = (\boldsymbol{V}, \boldsymbol{E})$ is said to be a causal Bayesian network if for all $\boldsymbol{S} \subseteq \boldsymbol{V}$ and $V_i \in \boldsymbol{V} \backslash \boldsymbol{S}$:

1. $P(v_i \mid do(\boldsymbol{S} = \boldsymbol{s}))$ is Markov relative to G.

2. The intervention $do(\boldsymbol{S} = \boldsymbol{s})$ is modular.

The assumptions of the interventional distributions implicit in the definition of causal Bayesian networks immediately imply (SGS-)minimality in case the conditional probability distributions are strictly positive. In case they are deterministic, there still is good reason to assume (SGS-)minimality [264].

As the Markov assumption implies a factorization of a Bayesian network, in a similar way the modularity assumption implicit in the causal Bayesian networks enforces the *truncated factorization* for interventional distributions [26]:

Assumption 15 (Truncated Factorization)**.** Let P be a probability distribution Markov relative to Bayesian network $G = (\boldsymbol{V}, \boldsymbol{E})$. Let $\boldsymbol{S} \subseteq \boldsymbol{V}$ be the set of random variables where is intervened upon. The truncated factorization is assumed to hold if:

$$P(\boldsymbol{v} \mid do(\boldsymbol{S} = \boldsymbol{s})) = \prod_{i|V_i \notin \boldsymbol{S}} P(v_i \mid \boldsymbol{pa}_i) \quad \text{if } \boldsymbol{v} \text{ consistent with intervention } \boldsymbol{s}$$

and 0 otherwise.

The truncated factorization property implicit in Markovian causal Bayesian networks reduces marginal inference in Markovian causal Bayesian networks to marginal inference in the *mutilated Bayesian networks*. These are the networks that are obtained when removing all the arrows to these nodes where is intervened upon. Although the truncated factorization property is sometimes known as the *g-formula* [182], it will be emphasized in Section 3.4.2 that the g-formula is derived from a different appreciation of the fundamental problem of causal inference as shown in Figure 3.6.

More efficiently, the interventional distribution can be computed by means of the adjustment set and associated adjustment formula for causal Bayesian networks [244]:

Definition 3.11 (Adjustment Set)**.** Let P be a probability distribution Markov relative to Bayesian network $G = (\boldsymbol{V}, \boldsymbol{E})$. An adjustment set is a set $\boldsymbol{V}_J \subset \boldsymbol{V}$ for which:

$$P(\boldsymbol{v}_S \mid do(\boldsymbol{V}_K = \boldsymbol{v}_K)) = \begin{cases} P(\boldsymbol{v}_S \mid \boldsymbol{v}_K) & \text{if } \boldsymbol{V}_J = \emptyset, \\ \sum_{\boldsymbol{v}_J} P(\boldsymbol{v}_S \mid \boldsymbol{v}_K, \boldsymbol{v}_J) P(\boldsymbol{v}_J) & \text{otherwise.} \end{cases} \tag{3.2}$$

Semi-Markovian causal Bayesian networks: The concepts and assumptions introduced in this section do naturally extend to the case when the models allow for unobserved confounding variables, as is the case in *semi-Markovian* models. Naturally, the Markov assumption cannot be adopted but is replaced by a semi-Markov assumption. Although the full specifications of the semi-Markovian causal Bayesian network have been detailed by Bareinboim et al. [27], it is important to emphasize that inherent to that definition is an adjusted version of the Markov assumption and modularity assumption, tailor-made to account for the complexities when latent variables are involved.

As described in Section 3.4.1, the object that emerges when unobserved confounding random variables are at play is an ADMG. Naturally, the Markov assumption as defined above does not hold when unobserved confounders are involved, because the latent confounders cannot be conditioned on. By generalizing d-separation to m-separation, the Markov assumption can be extended to ADMGs [194], resulting in the semi-Markov assumption. Similarly, as in the original Markov assumption, the semi-Markov assumption can also be expressed in terms of m-separation or in terms of the truncated factorization of the distribution. It has been shown that both definitions are equivalent [194], but for specifications of the semi-Markov assumption or the associated semi-modularity assumption, the reader is referred to the article by Bareinboim et al. [27]. These assumptions together give rise to the *semi-Markovian causal Bayesian network*

Definition 3.12 (Semi-Markovian Causal Bayesian Network)**.** Let P be a probability distribution Markov defined on the ADMG $G = (\boldsymbol{V}, \boldsymbol{E})$. Then $G = (\boldsymbol{V}, \boldsymbol{E})$ is said to be a semi-Markovian causal Bayesian network if for all $\boldsymbol{S} \subseteq \boldsymbol{V}$ and $V_i \in \boldsymbol{V} \backslash \boldsymbol{S}$:

1. $P(v_i \mid do(\boldsymbol{S} = \boldsymbol{s}))$ is semi-Markov relative to $G_{\overline{\boldsymbol{S}}}$.

2. The intervention $do(\boldsymbol{S} = \boldsymbol{s})$ is semi-modular.

Obviously, the factorization implied by the semi-Markov assumption also leads to a form of truncated factorization of interventional distributions. For a full overview of this factorization and subsequent ways to marginalize out variables, the reader is referred to (the appendix of) Bareinboim et al. [27]. It has been proven that the do-calculus provides a complete toolkit necessary to rewrite interventional distributions to observational distributions, and the rules of do-calculus are implied by the assumptions implicit in the definition of the semi-Markovian Bayesian network [220]. Completeness of the do-calculus means that the do-calculus will provide an observational distribution

for each interventional distribution if it exists. When the interventional distributions cannot be written in observational terms, the distribution is called *unidentifiable*. Identification is a necessary condition for both non-parametric and parametric causal inference approaches

Parametric causal inference

Apart from some causal discovery methods, most of the concepts discussed so far are non-parametric concepts. Since potential outcomes by nature imply missing values, the fundamental problem of causal inference is essentially an *estimation problem*. That is why substantial contributions to causal inference also involve estimation. The motivation for parametric causal inference is briefly discussed, followed by an examination of the parametric counterpart of the truncated factorization (parametric g-formula), based on assumptions introduced in Section 3.2. At the third level of the hierarchy, these concepts will be extended (see Section 3.5).

The following example motivates the use of parametric methods as a result of estimation problems: according to the adjustment formula, the interventional probability $P(y \mid do(T = t))$ corresponding to the DAG of Figure 3.3 can be converted to observation probabilities:

$$P(y \mid do(T = t)) = \sum_{\boldsymbol{z}} P(y \mid T = t, \boldsymbol{z}) P(\boldsymbol{z}).$$

This is also known as the back-door adjustment [178]. Although using parametric methods would require additional assumptions on the functional form, there are two main benefits to using parametric approaches. First, when considering continuous treatment variables, the query of interest $P(y \mid do(T = t))$ might not be available from data for the intervention $do(T = t)$ of interest. Second, taking into account high-dimensional covariates $\boldsymbol{Z}$, summing over all the strata $\boldsymbol{z}$ could be intractable. Both estimation problems can be circumvented by assuming the functional form [100].

When returning to the fundamental problem of causal inference and the adjustment formula as a result of various assumptions in Section 3.2, calculating the conditional expectation $\mathbb{E}[Y \mid do(T = t)]$ of Figure 3.3 can be reduced to evaluating $\mathbb{E}_{\boldsymbol{Z}}\mathbb{E}[Y \mid T, \boldsymbol{Z}]]$. This would require the evaluation of $\mathbb{E}[Y \mid T, \boldsymbol{Z}]$ adjusted for the probability $P(\boldsymbol{z})$. However, a non-parametric evaluation of $\mathbb{E}[Y \mid T, \boldsymbol{Z}]$ is impossible when $\boldsymbol{Z}$ is high-dimensional. Therefore, one can fit a regression model to the data to receive the estimates for $\mathbb{E}[Y \mid T, \boldsymbol{Z}]$ for each combination of $(t, \boldsymbol{z})$ and only estimate the $P(\boldsymbol{z})$ for the $\boldsymbol{z}$ that are present in the data. This is called *standardization based on parametric*

models, or in a more general form, *the parametric g-formula*.

Alternatively, $\mathbb{E}[Y \mid do(T = t)]$ can be further reduced to

$$\mathbb{E}[Y \mid do(t)] = \sum_{y} \sum_{\boldsymbol{z}} \frac{yP(y, t, \boldsymbol{z})}{P(t \mid \boldsymbol{z})},$$

meaning $\boldsymbol{z}$ can be marginalized out from the joint probability if the conditional probability $P(t \mid \boldsymbol{z})$ for ending up in the treatment group $T = t$ is taken into account. When $\boldsymbol{Z}$ is high-dimensional, this cannot be completed with non-parametric methods, but parametric model specifications need to be assumed. Logistic regression would be a straightforward choice in case of binary treatment. This is an example of *inverse probability weighting*.

Together with g-estimation methods, inverse probability weighting and the parametric g-formula belong to the family of *g-methods*, a class of methods that allows the computation of the average causal effects under time-varying treatments [166]. All these methods rely on the availability of a causal diagram and on assumptions that have been described in Section 3.2. These assumptions include consistency, positivity, (conditional) ignorability, and no-interference as illustrated by Figure 3.6. The connection between the g-formula and the truncated factorization formula looms large because the latter stems from the non-parametric causality research, while the former originates in its parametric counterpart, both being derived from different assumptions.

In a similar way, expressions with the do-operator, such as $\mathbb{E}[Y \mid do(T = t)]$, can be formulated as expressions containing potential outcomes, $\mathbb{E}[Y(t)]$. Nonetheless, identifiable potential outcomes queries cannot always be reduced to observational queries via the do-calculus, as nested counterfactuals require more refined tooling for reduction. Section 3.5 explains that some properties of the do-calculus can be extended to account for the reduction of nested counterfactuals to observational queries as well [156].

3.4.3 Discovery, Identification and Inference with More Relaxations

There are also many more departures from traditional assumptions in causal discovery and inference that have been omitted so far and will be discussed here. Deviations that are henceforth considered are departures from the no-interference assumption, departures that allow for context-specific independence and departures that consider a different kind of intervention.

All of the discussed causal discovery methods in Section 3.4.1 are based on the i.i.d. assumption as illustrated by Figure 3.5. There is also an entire body of work in terms of causal discovery and inference when this assumption is violated [13, 151, 150, 139, 113, 239, 172, 180, 33, 217, 15]. As has been tenaciously demonstrated, the graphical structures that emerge as a result of causal discovery under interference, depend on the different kinds of causal interference present [172]. Causal research under interference has been bifurcating.

On the one hand, graphs with violations of the i.i.d. assumption allow directed edges for causal relationships as well as undirected edges for stable symmetric relationships. These can consequently be accounted for by either *Laurritzen-Wermuth-Frydenberg chain graphs* [138, 139, 33] or *Andersson-Madigan-Perlman chain graphs* [8] depending on the Markov property interpreted. Generalization of the former by relaxing causal sufficiency leads to segregated graphs [219, 217]. Complete identification and inference methods for segregated graphs with stable symmetric relationships are established [217]. Alternatively, an absorption of the Andersson-Madigan-Perlman chain graphs in combination with ADMGs [193] leads to a new family of graphical structures for which causal discovery methods exist for observational and interventional data [180].

On the other hand, extending the rules of d-separation to *relational d-separation*, a criterion for conditional independence in case of relational data, has given rise to an alternative representation, that enables the existence of independencies of relational data, called *abstract ground graphs* [150]. With an extension of the Peter-Clark algorithm, the *relational causal discovery* algorithm [151] makes it possible to extract the true relational causal structure in case of violations of the no-interference assumption. For every perspective, the relational causal model corresponds to an abstract ground graph. Inference is also possible under abstract ground graphs [13].

Because the Markov assumption has occasionally been defended [97] and criticized [43], there have also been attempts to relax the Markov assumption. Claiming that any variable is independent of its non-descendants given its parents excludes the possibility of conditional independence relations that only hold for a subset of realizations of conditioning variables [67]. Relaxing the Markov property to a kind of Markov property that allows for *context-specific independence* relations calls for different causal concepts that can account for this such as Bayesian multinets [80], conditional probability tables with regularity structure [37], staged trees and chain event graphs [226] and labeled directed acyclic graphs [181]. Various algorithms for causal discovery exist for staged trees [42, 142] as well as for labeled directed acyclic graphs [115] (with a

slightly adapted version of faithfulness). There are also inference methods available when context-specific independence is involved [240].

Besides the atomic or hard interventions discussed in Section 3.4.2, there are also stochastic or soft interventions. These interventions do not force the intervened variable to take on a fixed value, but merely replace the underlying causal mechanism by a known function [53, 70]. The do-calculus falls short in converting causal queries with soft interventions or conditional interventions. For that, a more general calculus is required, called *σ-calculus* [53] that can account for stochastic interventions and comes with a concomitant inference algorithm.

3.4.4 Practical Guide to Causal Inference

In this section, three practical considerations for conducting causal inference are discussed. Figure 3.7 summarizes key considerations for converting observational data into causal information, outlining the flow from data to causal graphical structures, estimands, and estimates, while highlighting the role of assumptions and expert knowledge.

First, identifying the necessary components to address inference queries of interest

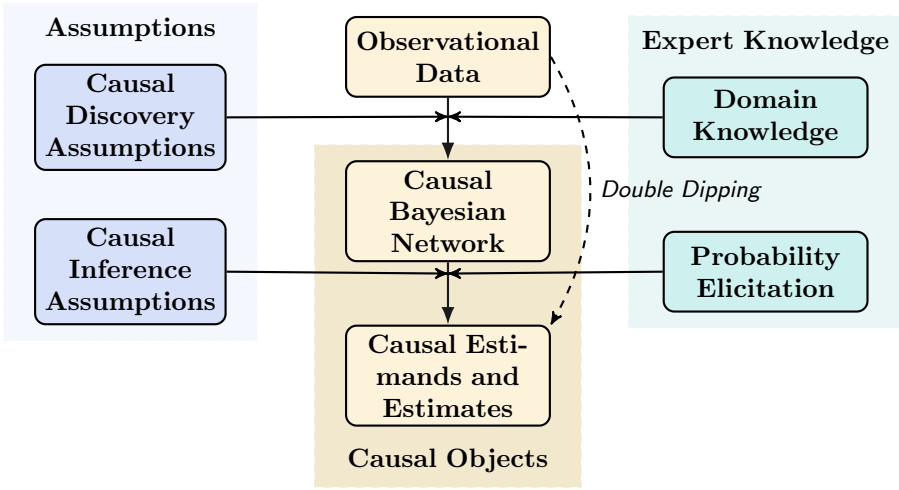

Figure 3.7: Flow of causal reasoning from data to graphical components (causal discovery) and subsequently to causal estimands and estimates (causal inference). While the light blue part indicates at what stage assumptions have to be taken into account, the light green part indicates possible supplementary information from domain experts. Double dipping occurs when data in the causal discovery stage is being reused at the causal inference stage.

requires a sound graphical structure. This graphical structure can be obtained via domain experts or causal discovery methods. When causal inference methods are applied to data that has already undergone causal discovery algorithms, this can result in *double dipping*, compromising the validity of the confidence intervals provided by the statistical methods [47]. Practitioners should be mindful of this issue and, if needed, apply available methods that can correct for this bias [87].

Second, it can be the case that there is no data available to estimate the conditional probability distributions (or one does not want to engage in double dipping). In this case, practitioners can still engage in causal inference when the conditional probability distributions are elicited from domain experts. There already exist efficient methods to infer discrete conditional probability distributions from experts [29, 5, 95], but eliciting continuous conditional probability distributions remains underdeveloped.

Finally, as discussed, causal inference relies on statistical methods, where larger sample sizes improve reliability. However, when $\boldsymbol{Z}$ is high-dimensional, non-parametric estimation of $\mathbb{E}[Y \mid T, \boldsymbol{Z}]$ becomes infeasible [100]. To address this, parametric methods such as the *parametric g-formula* and commonly parametric implementations of inverse probability weighting can be used, though they require specifying functional forms. When these assumptions are restrictive, semi-parametric alternatives offer a balance between flexibility and efficiency [34, 148].

3.5 Counterfactual Level

The components introduced in the previous two sections are not sufficient to address queries at the third level of the hierarchy. While the second level represents interventions on conditioning variables, the third level corresponds to interventions on conditioned variables. As mentioned in Section 3.2.4, the object necessary to reason at all levels of the hierarchy, including the counterfactual level, is the SCM. Next is an example of how an SCM can be utilized to reason at the counterfactual level of the hierarchy when the causal Bayesian network falls short:

Example 3 (Counterfactual Queries Require SCMs)**.** Assume the linear Gaussian (Markovian) causal Bayesian network corresponding to the graph $X \to Y$ with

$$\begin{aligned} X &\sim \mathcal{N}(1, 4) \\ Y &\sim \mathcal{N}(-0.5X + 3, 1). \end{aligned}$$

The intervention distribution $P(Y \mid do(X = 1))$ can be computed via the truncated

factorization formula and results in $\mathcal{N}(2.5, 1)$. However, the counterfactual distribution $P(Y(X = 0) \mid X = 1, Y = 4)$, meaning the probability of Y had X been set to 0 given that $X = 1$ and $Y = 4$, cannot be computed with a causal Bayesian network alone. In order to compute this counterfactual query, access to the SCM is required.

Therefore, assume the following structural equations in the SCM:

$$\begin{aligned} f_1(w_1) &= w_1 && \text{where } w_1 \sim \mathcal{N}(1, 4) \\ f_2(X, w_2) &= -0.5X + w_2 && \text{where } w_2 \sim \mathcal{N}(3, 1). \end{aligned}$$

The evidence of the counterfactual query, $X = 1$ and $Y = 4$, can be used to update the distribution of the exogenous variables in the SCM to $w_1 \sim \delta(1)$ and $w_2 \sim \delta(4)$, with $\delta()$ being the Dirac delta measure. Ingesting the intervention $X = 0$ into the updated structural equations leads to a complete evaluation of the counterfactual query: $P(Y(X = 0) \mid X = 1, Y = 4) = f_2(X = 0, w_2) = \delta(4)$.

One of the reasons much research has been dedicated to the first two levels of the hierarchy is that access to the fully specified SCM is considered to be implausible. While the above linear Gaussian (Markovian) Bayesian network gives rise to a natural separation between the endogenous and exogenous variables, the interaction between the observed and latent variables is often unknown, rendering access to the fully specified SCM 'hopeless' [27]. Despite the inaccessibility of the fully specified SCM, scholars have painstakingly reasoned with counterfactual models, because it plays an essential role in mediation analysis [197, 196]. Some counterfactual models have antagonized scholars that have argued that the introduced assumptions are not scientific because they lack the possibility of empirical validation [61].

This section generalizes the potential outcome framework introduced in Section 3.2.1, which is equivalent to the structural causal model framework presented in Section 3.2.4, thereby shedding new light on the assumptions involved at the third level of the hierarchy (see Figure 3.8). The different counterfactual models emerging from assumptions are emphasized, and the inference tools available for each model are highlighted. Throughout this section, the existence of a topological sort on the random variables is assumed.

[5]The SWIG does not immediately apply the edge g-formula, but the graphical structure of the SWIG can be generalized to allow edge interventions [222].

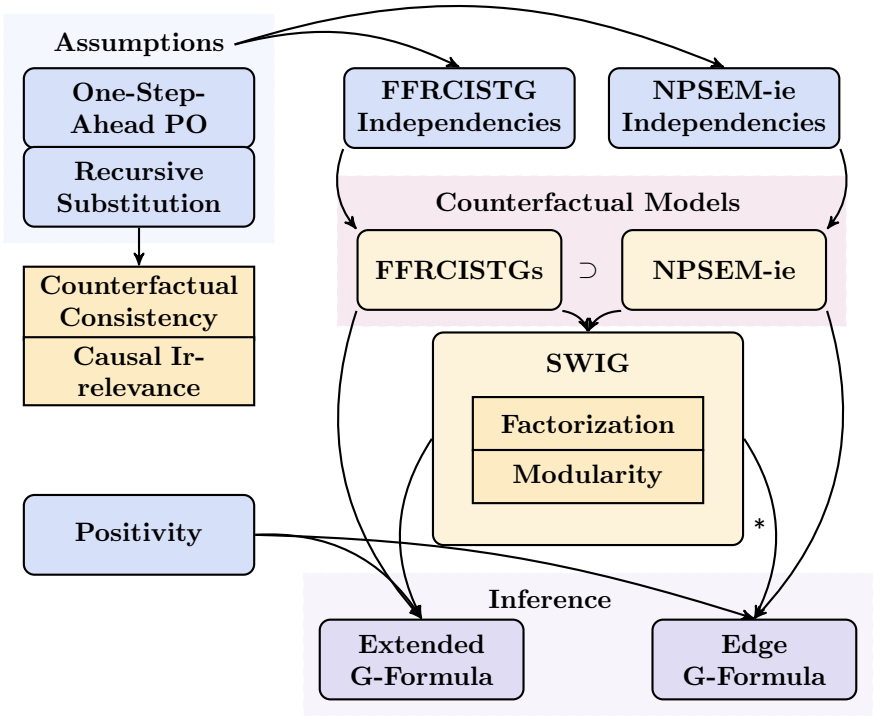

Figure 3.8: The definition of the one-step-ahead potential outcomes and recursive substitution imply the desirable counterfactual consistency and causal irrelevance property. Additional independence assumptions need to be adopted to yield a counterfactual model, which can either be a FFRCISTG or a NPSEM-ie. The SWIG unifies these models with graphical approaches and features a factorization and modularity property. Together with the positivity assumption, inference can be conducted via the extended g-formula or the edge g-formula.[5]

3.5.1 One-Step-Ahead Potential Outcomes

The very definition of counterfactuals entails the existence of a hypothetical world that may not be empirically verifiable. Therefore, the existence of *one-step-ahead potential outcomes* is assumed.

Assumption 16 (One-Step-Ahead Potential Outcomes)**.** Let $X_1, \dots, X_n$ be random variables corresponding to nodes $V_1, \dots, V_n$. Then for all $V_i \in \boldsymbol{V}$ and possible assignments of parents $\boldsymbol{pa}_i \in \Omega_{\mathbf{pa}(V_i)}$, the existence of one-step-ahead potential outcomes $V_i(\mathbf{pa}(V_i) = \boldsymbol{pa}_i)$ is assumed.

Note that $V_i(\mathbf{pa}(V_i) = \boldsymbol{pa}_i)$ corresponds to the notation introduced in the potential outcome framework of Section 3.2.1. Intuitively, the one-step-ahead potential

outcome corresponds to the response V_i had the parents of V_i been set to $\boldsymbol{pa}_i$. This is emphasized as an assumption because the assumed potential outcomes could possibly be counterfactual and therefore presuming the existence of a hypothetical world. Since not all potential outcomes naturally depend on possible assignments of parent nodes in the topological sort, it is necessary to extend the definition of potential outcomes via recursive substitution.

Assumption 17 (Recursive Substitution)**.** Let $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ be random variables corresponding to nodes $\boldsymbol{V} = \{V_1, \ldots, V_n\}$. Assume the existence of one-step-ahead potential outcomes $V_i(\mathbf{pa}(V_i) = \boldsymbol{pa}_i)$ for all $V_i \in \boldsymbol{V}$ and possible assignments of parents $\boldsymbol{pa}_i \in \Omega_{\mathbf{pa}(V_i)}$. Then for all $\boldsymbol{S} \subset \boldsymbol{V}$ and $\boldsymbol{s} \in \Omega_{\boldsymbol{S}}$ it is assumed that $V_i(\boldsymbol{s})$ can be expressed recursively:

$$V_i(\boldsymbol{s}) = V_i(\boldsymbol{s} \cap \boldsymbol{pa}_i, \{V_j(\boldsymbol{s}) \mid V_j \in \mathbf{pa}(V_i), V_j \notin \boldsymbol{S}\}).$$

$V_i(\boldsymbol{s})$ is thus the potential outcome where the parents of V_i that are in $\boldsymbol{S}$ had been set to $\boldsymbol{s}$ and variables for which $V_j \in \mathbf{pa}(V_i) \setminus \boldsymbol{S}$ are set to the values these potential outcomes would have had had $\boldsymbol{S}$ been set to $\boldsymbol{s}$, denoted by $V_j(\boldsymbol{s})$.

Example 4 (Recursive Substitution in Graphical Strucures)**.** Assume the topological sort over the random variables $\boldsymbol{Z}, T, Y$ as implied by Figure 3.3. Then, it is assumed that the one-step-ahead potential outcome $Y(\boldsymbol{z})$ is defined recursively:

$$\begin{aligned} Y(\boldsymbol{z}) &= Y(\boldsymbol{z} \cap \boldsymbol{pa}_Y, \{V_j(\boldsymbol{z}) \mid V_j \in \mathbf{pa}(Y), V_j \notin \boldsymbol{Z}\}) \\ &= Y(\boldsymbol{z}, T(\boldsymbol{z})). \end{aligned}$$

Expressing potential outcomes recursively brings along desirable properties as illustrated by Figure 3.8. First of all, it directly implies the consistency assumption introduced in Section 3.2 [156]. Second, it proves the so-called *causal irrelevance*: every potential outcome derived from recursive substitution $V_i(s)$ can be expressed as a unique minimally causal relevant subset of $W \subseteq S$: $V_i(s) = V_i(w)$. The reader can find the specifications of a minimally causal relevant subset and the proof in the work of Malinsky et al. [156]. Equivalence between the structural causal model and the potential outcome framework follows from the equivalent representation of the one-step-ahead counterfactual $V_i(\boldsymbol{pa}_i)$ as the output of the structural equation $f_i(\boldsymbol{pa}_i, \boldsymbol{w}_i)$ (by letting $\boldsymbol{w}_i = \{V_i(\boldsymbol{pa}_i) \mid \boldsymbol{pa}_i \in \Omega_{\mathbf{pa}(V_i)}\}$ and setting $f_i(\boldsymbol{pa}_i, w_i) = (\boldsymbol{w}_i)_{\boldsymbol{pa}_i} = V_i(\boldsymbol{pa}_i)$).

3.5.2 Counterfactual Models

In addition to consistency and causal irrelevance, independence relations are assumed in order to reason about counterfactuals. The literature splits along the lines of which independence assumptions to adopt. There is the more conservative *finest fully randomized causally interpretable structured tree graph* (FFRCISTG) and the more restrictive *non-parametric structural equation model with independent errors* (NPSEM-ie). First, the FFRCISTG independencies are introduced.

Assumption 18 (FFRCISTGS Independencies)**.** Assume one-step-ahead counterfactuals by recursive substitution. Let $\boldsymbol{v}$ be an assignment for random variables $\boldsymbol{V}$ and let $\boldsymbol{pa}_i$ be the restriction of that assignment to parent variables of V_i. Then for each assignment $\boldsymbol{v}$, the corresponding one-step-ahead counterfactuals consistent with $\boldsymbol{v}$ are mutually independent:

$$V_1 \perp\!\!\!\perp_P V_2(\boldsymbol{pa}_2) \perp\!\!\!\perp_P \ldots \perp\!\!\!\perp_P V_n(\boldsymbol{pa}_n),$$

where $V_i < V_{i+1}$ in the topological sort.

It is important to note that all counterfactual random variables are consistent with each other in the sense that there is no contrary assignment among them. Extra independencies across contradicting assignments are imposed by assuming independencies of the error terms in the non-parametric structural equation models. Formally, the counterfactual random variables that are independent in the NPSEM-ie model are defined as follows.

Assumption 19 (NPSEM-ie Independencies)**.** Assume one-step-ahead counterfactuals by recursive substitution. Then the set of one-step-ahead counterfactuals across possibly contradictory interventions are mutually independent:

$$\{V_1\} \perp\!\!\!\perp_P \{V_2(\boldsymbol{pa}_2) \mid \boldsymbol{pa}_2 \in \Omega_{\mathbf{pa}(V_2)}\} \perp\!\!\!\perp_P \ldots \perp\!\!\!\perp_P \{V_n(\boldsymbol{pa}_n) \mid \boldsymbol{pa}_n \in \Omega_{\boldsymbol{pa}(V_n)}\},$$

where $V_i < V_{i+1}$ by the topological sort.

Because the NPSEM-ie independencies also contain the FFRCISTGS independencies, the NPSEM-ie model is *strictly stronger* than the FFRCISTGS model. Consistency and causal irrelevance are implicit in the NPSEM-ie as well as the FFRCISTGS model.

Example 5 (Difference in Counterfactual Model Independencies)**.** Assume one-step-ahead potential outcome random variables corresponding to the nodes $\boldsymbol{Z}, T, Y$ respect-

ing the topological sort of Figure 3.3. Then, following the FFRCISTGS model, for assignment $\boldsymbol{z}_1, t$ the following independence exist:

$$\boldsymbol{Z} \perp\!\!\!\perp_P T(\boldsymbol{z}_1) \perp\!\!\!\perp_P Y(t, \boldsymbol{z}_1).$$

In addition to the previous independencies, according to the NPSEM-ie model, other independencies across contradictory assignments $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are implied, such as:

$$\boldsymbol{Z} \perp\!\!\!\perp_P T(\boldsymbol{z}_2) \perp\!\!\!\perp_P Y(t, \boldsymbol{z}_1).$$

While DAGs and ADMGs are not expressive enough to account for reasoning with one-step-ahead potential outcomes with either NPSEM-ie independencies or FFRCISTGS independencies, a more refined graphical construction called a *single world intervention graph* (SWIG) was introduced via a node-splitting operation based on causal irrelevance. The SWIG can encode the independence relations of either the NPSEM-ie or the FFRCISTGS. Similarly to how the causal Bayesian networks assume a factorization property of the interventional distributions and modularity property about the nature of interventions, the SWIGs obey properties that specify the behavior of counterfactual distributions. Both the NPSEM-ie model and the FFRCISTGS model, together with consistency, imply these factorization and modularity properties for SWIGs [195] as illustrated by Figure 3.8.

3.5.3 Inference

Inference on the counterfactual level is concerned with the identification of the relevant components necessary to address counterfactual queries. In order to calculate the distribution of counterfactuals under different interventions, the *g-formula* can be used, which has been in Section 3.4.2. This formula can be extended to account for unit-specific interventions and the distribution of that intervention [262] resulting in the *extended g-formula* [198, 195]:

Proposition 20 (Extended G-Formula)**.** Let $\boldsymbol{S} \subset \boldsymbol{V}$ and $V(\boldsymbol{s})$ be the one-step-ahead counterfactual defined by recursive substitution. Then given positivity, the joint distribution can be written as:

$$P(V_1(\boldsymbol{s}), \ldots, V_n(\boldsymbol{s})) = \prod_{i|V_i \in \boldsymbol{V}} P(V_i \mid \boldsymbol{s} \cap \boldsymbol{pa}_i, \mathbf{pa}(V_i) \notin \boldsymbol{S}).$$

The formula is equivalent to the factorization and modularity property implicit in SWIGs and has been proven to hold [195, 221]. The strength of the formula is that it rewrites counterfactual distributions in terms of observational distributions, but unlike the g-formula, the extended g-formula also accounts for nested counterfactuals by having a term for every $V_i \in \boldsymbol{V}$. Analogously, the do-calculus can be extended to rewrite nested counterfactuals such as dynamic treatment regimes or path-specific interventions. For that reason, *po-calculus* has been introduced as a generalization of the do-calculus as a result of consistency, causal irrelevance, and factorization on SWIGs [156]. Although the po-calculus implies the do-calculus for interventional queries [156], it has been shown that additional identification results consisting of nested counterfactuals follow exclusively from the po-calculus [221].

As there exists a hierarchy of causal queries, there is also a *hierarchy of interventions*. The most granular form of interventions are node interventions according to the hierarchy of interventions of [222]. Node interventions are a specific form of edge interventions which in turn are a specific form of path interventions. Multiple targets of interest in mediation analysis are defined as edge interventions and for this reason, the extended g-formula has also been extended to the *edge g-formula* [222]. While node interventions are associated with the FFRCISTG model and require the extended g-formula for identification, edge interventions correspond to the NPSEM-ie model and require the edge g-formula for identification as shown in Figure 3.8.

3.6 Conclusion and Future Work

This section has synthesized existing research on causality by situating different research areas within the framework of Pearl's causal hierarchy. The concepts and associated assumptions required to address queries at different levels of the hierarchy have been highlighted. These foundational causal concepts form the basis for the analyses conducted in the remainder of the thesis. Future research should further explore causal inference in systems with feedback, as most existing work assumes acyclicity in structural causal models, despite real-world phenomena, such as climate systems, often exhibiting cyclical causal relations [55, 36].